**Prospective multicentre head-to-head validation of host blood transcriptomic biomarkers for pulmonary tuberculosis by real-time PCR: Public Dataset**

[1. Overview 1](#_Toc70446290)

[2. Abstract 2](#_Toc70446291)

[3. Privacy of Information Disclosed 3](#_Toc70446292)

[4. Cross-Sectional TB Cohort Study Subject-Level Dataset Data Dictionary with Example Values 4](#_Toc70446293)

[5. CORTIS-01 Sub-Study Subject-Level Dataset Data Dictionary with Example Values 5](#_Toc70446294)

[6. CORTIS-HR Sub-Study Subject-Level Dataset Data Dictionary with Example Values 7](#_Toc70446295)

[7. Respiratory Pathogens Sub-Study Subject-Level Dataset Data Dictionary with Example Values 9](#_Toc70446296)

[8. Subject-Level PCR Dataset Data Dictionary with Example Values 12](#_Toc70446297)

# Overview

The Correlate of Risk Targeted Intervention Study(**CORTIS-01**; ClinicalTrials.gov: [NCT02735590](http://clinicaltrials.gov/show/NCT02735590); [doi.org/10.1016/s1473-3099(20)30914-2](http://doi.org/10.1016/s1473-3099(20)30914-2)), and companion study Validation of Correlates of Risk of TB Disease in High Risk Populations (**CORTIS-HR**; [doi.org/10.1016/S2214-109X(21)00045-0](http://doi.org/10.1016/S2214-109X(21)00045-0)), were conducted to prospectively test the diagnostic and prognostic performance of the RISK11 biomarker for tuberculosis (TB) disease in people without HIV and people living with HIV, respectively, in an ambulant community setting. In addition, CORTIS-01 tested the efficacy of biomarker-guided TB preventive therapy.

The Cross-Sectional TB Cohort (**CTBC**) study was previously used to validate the diagnostic performance of the RISK11 and RISK6 biomarkers in a case-control study design ([doi.org/10.1038/s41598-020-65043-8](http://doi.org/10.1038/s41598-020-65043-8)).

The **Respiratory Pathogens** sub-study evaluated the effect of respiratory viral and bacterial pathogens detected by RT-qPCR on nasopharyngeal and oropharyngeal swabs on RISK11 biomarker scores in people without HIV consecutively recruited in the CORTIS-01 study.

Here we provide the public, subject-level datasets containing key variables necessary to reconstruct the study findings for the Cross-Sectional TB Cohort (CTBC) study, and the CORTIS-01, CORTIS-HR, and Respiratory Pathogens sub-studies evaluating the diagnostic and prognostic performance of multiple parsimonious transcriptomic signatures. Data dictionaries are provided below.

The public datasets for each study are divided into (1) subject-level signature scores and metadata with key variables, and (2) subject-level TaqMan RT-qPCR primer-probe assay raw cycle threshold (CT) gene expression data from the Fluidigm microfluidic gene expression integrated fluidic circuits. Summary of included files:

|  |  |
| --- | --- |
| **Public, subject-level datasets with signature scores and metadata with key variables** | |
| 1. CTBC\_SmallSigScores\_pubdata.csv | Cross-Sectional TB Cohort study |
| 2. CORTIS-01\_SmallSigScores\_pubdata.csv | CORTIS-01 sub-study |
| 3. CORTIS-HR\_SmallSigScores\_pubdata.csv | CORTIS-HR sub-study |
| 4. RespPathogens\_SmallSigScores\_pubdata.csv | Respiratory Pathogens sub-study |
| **Public, subject-level PCR datasets with TaqMan gene expression assay CT values** | |
| 5. CTBC\_SmallSigScores\_pubPCRdata.csv | Cross-Sectional TB Cohort study |
| 6. CORTIS-01\_SmallSigScores\_pubPCRdata.csv | CORTIS-01 sub-study |
| 7. CORTIS-HR\_SmallSigScores\_pubPCRdata.csv | CORTIS-HR sub-study |
| 8. RespPathogens\_SmallSigScores\_pubPCRdata.csv | Respiratory Pathogens sub-study |

# Abstract

We tested performance of host-blood transcriptomic tuberculosis (TB) signatures for active case-finding. Among 20,207 HIV-uninfected and 963 HIV-infected adults screened; 2,923 and 861 were enrolled from five South African communities. Eight signatures were measured by microfluidic RT-qPCR and participants were microbiologically-investigated for pulmonary TB at baseline, and actively surveilled for incident disease through 15 months. Diagnostic AUCs for 61 HIV-uninfected (weighted-prevalence 1.1%) and 10 HIV-infected (prevalence 1.2%) prevalent TB cases for the 8 signatures were 0.63–0.79 and 0.65–0.88, respectively. Thereafter, 24 HIV-uninfected and 9 HIV-infected participants progressed to incident TB (1.1 and 1.0 per 100 person-years, respectively). Prognostic AUCs through 15-months follow-up were 0.49–0.66 and 0.54–0.81, respectively. Prognostic performance for incident TB occurring within 6-12 months in HIV-negative participants was higher for all signatures. None of the signatures met WHO Target Product Profile criteria for a triage test to diagnose subclinical TB; most signatures met the criteria for symptomatic TB. Prognostic accuracy of most signatures for incident TB within six months of testing met the criteria for an incipient TB test.

**Funding**

Bill & Melinda Gates Foundation, South African Medical Research Council

**Sponsor**

University of Cape Town

**Corresponding author:** Thomas J Scriba

**Email:** Thomas.Scriba@uct.ac.za **Telephone:** +27 (0) 21 406 6080

**Postal address:** South African Tuberculosis Vaccine Initiative, Institute of Infectious Disease and Molecular Medicine, Division of Immunology, Department of Pathology, University of Cape Town, Anzio Road, Observatory, 7935, Cape Town, South Africa

# Privacy of Information Disclosed

While the Fred Hutchinson Cancer Research Center and by extension the Statistical Center for HIV/AIDS Research and Prevention (SCHARP) is not a covered entity regarding HIPAA requirements, SCHARP uses the 'limited dataset' model as described in HIPAA when publishing public use datasets. This limited Dataset is used for public health and research purposes only. No personal identifying information is used in the Datasets, such as names or social security numbers. The Dataset uses its own unique identifiers that cannot identify the research participants (i.e., the Dataset identifiers are not participant IDs used in the research). SCHARP and South African Tuberculosis Vaccine Initiative (SATVI) at the University of Cape Town (UCT) staff are trained on and sign an agreement to follow SCHARP's and the UCT’s privacy and confidentiality policies, respectively.

These procedures comply with relevant models required in FDA regulations such as Title 21 CFR Parts 20, 21, 803, Title 45 CFR Part 164.514(e)(3)(i), and as recommended in Good Clinical Data Management Practices (GCDMP) defined by the Society of Data Quality Management.

# Cross-Sectional TB Cohort Study Subject-Level Dataset Data Dictionary with Example Values

| **Variable Name** | **Variable Description** | **Format** | **Example Values** |
| --- | --- | --- | --- |
| 1) pubid | Unique identification number for each subject in study | char | 1. 100001, 2. 356282 … |
| 2) TB\_status | TB status at enrollment in case-control study | char | 1. “Control”, 2. “TB Case” |
| 3) HIV\_status | HIV status at enrollment in case-control study | char | 1. “Negative”, 2. “Positive” |
| 4) Darboe11, Francisco2, Maertzdorf4, Penn-Nicholson6, Roe1, Roe3, Suliman4, Sweeney3, and Thompson5 | Signature scores at enrollment | num | 1. 19.895, 2. 60.234, … |

# CORTIS-01 Sub-Study Subject-Level Dataset Data Dictionary with Example Values

| **Variable Name** | **Variable Description** | **Format** | **Example Values** |
| --- | --- | --- | --- |
| 1) pubid | Unique identification number for each subject in study | char | 1. 100001, 2. 356282 … |
| 2) Darboe11, Francisco2, Maertzdorf4, Penn-Nicholson6, Roe1, Roe3, Suliman4, Sweeney3, and Thompson5 | Signature scores at enrollment | num | 1. 19.895, 2. 60.234, … |
| 3) CORTIS-01 Arm | CORTIS-01 clinical trial randomization arms:  Group A: RISK11+ 3HP+  Group B: RISK11+ 3HP-  Group C: RISK11- 3HP- | char | 1. “Group A” 2. “Group B” 3. “Group C” |
| 4) weighting A+B vs C | Respective weighting of study arms:  Group A+B: 1.263388938  Group C: 7.919843049 | num | 1. 1.263388938, 2. 7.919843049 |
| 5) weighting A vs B vs C | Respective weighting of study arms:  Group A: 3.837333333  Group B: 1.883507853  Group C: 7.919843049 | num | 1. 1.883507853, 2. 3.837333333, 3. 7.919843049 |
| 6) endpoint\_osts | Binary indicator for secondary endpoint (≥ 1 sample positive) | int | 1. 0, 2. 1 |
| 7) prevalentTB\_osts | Binary indicator for secondary endpoint prevalent TB (≥ 1 sample positive) | int | 1. 0, 2. 1 |
| 8) prognosticCohort\_osts | Binary indicator for whether participant included in prognostic analysis for secondary endpoint incident TB (≥ 1 sample positive) | int | 1. 0, 2. 1 |
| 9) incidentTB\_osts | Binary indicator for secondary endpoint incident TB (≥ 1 sample positive) | int | 1. 0, 2. 1 |
| 10) FUtime\_osts | Follow-up time for secondary endpoint analysis (months) | num | 1. 0, 2. 1.94, 3. 15, … |
| 11) endpoint\_ts | Binary indicator for primary endpoint (≥ 2 sample positive) | int | 1. 0, 2. 1 |
| 12) prevalentTB\_ts | Binary indicator for primary endpoint prevalent TB (≥ 2 sample positive) | int | 1. 0, 2. 1 |
| 13) prognosticCohort\_ts | Binary indicator for whether participant included in prognostic analysis for primary endpoint incident TB (≥ 2 sample positive) | int | 1. 0, 2. 1 |
| 14) incidentTB\_ts | Binary indicator for primary endpoint incident TB (≥ 2 sample positive) | int | 1. 0, 2. 1 |
| 15) FUtime\_ts | Follow-up time for primary endpoint analysis (months) | num | 1. 0, 2. 1.94, 3. 15, … |
| 16) sex | Participant sex | char | 1. “MALE”,  2. “FEMALE” |
| 17) age | Participant age | num | 1. 18,  2. 30.26,  3. 55 |
| 18) ethnicity | Participant ethnicity | char | 1. “ASIAN”,  2. “BLACK”,  3. “CAUCASIAN”,  4. “MIXED RACE”,  5. “OTHER” |
| 19) bmi | Body-mass index at enrolment | num | 1. 14.2,  2. 21.01,  3. 24.76 |
| 20) weight | Weight at enrolment (kilogram) | num | 1. 92.3,  2. 68.1,  3. 44.8 |
| 21) smoking\_history | History of smoking | char | 1. “YES”,  2. “NO” |
| 22) tb\_history | History of prior TB disease | char | 1. “YES”,  2. “NO” |
| 23) tb\_hhc | Tuberculosis household contact | char | 1. “YES”,  2. “NO” |
| 24) tb\_screen | Any TB symptoms at baseline (at enrollment visit) | char | 1. “POSITIVE”, 2. “NEGATIVE” |

# CORTIS-HR Sub-Study Subject-Level Dataset Data Dictionary with Example Values

| **Variable Name** | **Variable Description** | **Format** | **Example Values** |
| --- | --- | --- | --- |
| 1) pubid | Unique identification number for each subject in study | char | 1. 100001, 2. 356282 … |
| 2) Darboe11, Francisco2, Maertzdorf4, Penn-Nicholson6, Roe1, Roe3, Suliman4, Sweeney3, and Thompson5 | Signature scores at enrollment | num | 1. 19.895, 2. 60.234, … |
| 3) endpoint\_osts | Binary indicator for secondary endpoint (≥ 1 sample positive) | int | 1. 0, 2. 1 |
| 4) prevalentTB\_osts | Binary indicator for secondary endpoint prevalent TB (≥ 1 sample positive) | int | 1. 0, 2. 1 |
| 5) prognosticCohort\_osts | Binary indicator for whether participant included in prognostic analysis for secondary endpoint incident TB (≥ 1 sample positive) | int | 1. 0, 2. 1 |
| 6) incidentTB\_osts | Binary indicator for secondary endpoint incident TB (≥ 1 sample positive) | int | 1. 0, 2. 1 |
| 7) FUtime\_osts | Follow-up time for secondary endpoint analysis (months) | num | 1. 0, 2. 1.94, 3. 15, … |
| 8) endpoint\_ts | Binary indicator for primary endpoint (≥ 2 sample positive) | int | 1. 0, 2. 1 |
| 9) prevalentTB\_ts | Binary indicator for primary endpoint prevalent TB (≥ 2 sample positive) | int | 1. 0, 2. 1 |
| 10) prognosticCohort\_ts | Binary indicator for whether participant included in prognostic analysis for primary endpoint incident TB (≥ 2 sample positive) | int | 1. 0, 2. 1 |
| 11) incidentTB\_ts | Binary indicator for primary endpoint incident TB (≥ 2 sample positive) | int | 1. 0, 2. 1 |
| 12) FUtime\_ts | Follow-up time for primary endpoint analysis (months) | num | 1. 0, 2. 1.94, 3. 15, … |
| 13) sex | Participant sex | char | 1. “MALE”,  2. “FEMALE” |
| 14) age | Participant age | num | 1. 18,  2. 30.26,  3. 55 |
| 15) ethnicity | Participant ethnicity | char | 1. “ASIAN”,  2. “BLACK”,  3. “CAUCASIAN”,  4. “MIXED RACE”,  5. “OTHER” |
| 16) bmi | Body-mass index at enrolment | num | 1. 14.2,  2. 21.01,  3. 24.76 |
| 17) weight | Weight at enrolment (kilogram) | num | 1. 92.3,  2. 68.1,  3. 44.8 |
| 18) smoking\_history | History of smoking | char | 1. “YES”,  2. “NO” |
| 19) tb\_history | History of prior TB disease | char | 1. “YES”,  2. “NO” |
| 20) tb\_hhc | Tuberculosis household contact | char | 1. “YES”,  2. “NO” |
| 21) IPT | Isoniazid preventive therapy (IPT) duration at enrolment. “No IPT recorded” implies that the participant did not take IPT on study. "Started after enrolment" implies participant started IPT during the conduct of the study. | char | 1. "<6 months",  2. ">12 months",  3. "6-12 months",  4. "No IPT recorded",  5. "Started after enrollment" |
| 22) ART | Antiretroviral therapy (ART) duration at enrolment. “No ART recorded” implies that the participant did not take ART on study. "Started after enrolment" implies participant started ART during the conduct of the study. | char | 1. "<6 months",  2. ">12 months",  3. "6-12 months",  4. "No ART recorded",  5. "Started after enrollment" |
| 23) CD4 | CD4-positive cell count at enrolment | num | 1. 45,  2. 834,  3. 345, … |
| 24) viral\_load | HIV plasma viral load at enrolment | char | 1. “<100”,  2. “100-999”,  3. “>=1000” |
| 25) tb\_screen | Any TB symptoms at baseline (at enrollment visit) | char | 1. “Positive”, 2. “Negative” |

# Respiratory Pathogens Sub-Study Subject-Level Dataset Data Dictionary with Example Values

| **Variable Name** | **Variable Description** | **Format** | **Example Values** |
| --- | --- | --- | --- |
| 1) pubid | Unique identification number for each subject in study | char | 1. 100001, 2. 356282 … |
| 2) Darboe11, Francisco2, Herberg2, Maertzdorf4, Penn-Nicholson6, Roe1, Roe3, Suliman4, Sweeney3, and Thompson5 | Signature scores at enrollment | num | 1. 19.895, 2. 60.234, … |
| 3) CORTIS-01 Arm | CORTIS-01 clinical trial randomization arms:  Group A: RISK11+ 3HP+  Group B: RISK11+ 3HP-  Group C: RISK11- 3HP-  Not randomized: participants not randomized and excluded | char | 1. “Group A” 2. “Group B” 3. “Group C” 4. “Not randomized” |
| 4) prevalentTB\_osts | Binary indicator for secondary endpoint prevalent TB (≥ 1 sample positive) amongst randomized participants investigated for TB | int | 1. 0, 2. 1 |
| 5) prevalentTB\_ts | Binary indicator for primary endpoint prevalent TB (≥ 2 sample positive) amongst randomized participants investigated for TB | int | 1. 0, 2. 1 |
| 6) sex | Participant sex | char | 1. “MALE”,  2. “FEMALE” |
| 7) age | Participant age | num | 1. 18,  2. 30.26,  3. 55 |
| 8) ethnicity | Participant ethnicity | char | 1. “ASIAN”,  2. “BLACK”,  3. “CAUCASIAN”,  4. “MIXED RACE”,  5. “OTHER” |
| 9) bmi | Body-mass index at enrolment | num | 1. 14.2,  2. 21.01,  3. 24.76 |
| 10) weight | Weight at enrolment (kilogram) | num | 1. 92.3,  2. 68.1,  3. 44.8 |
| 11) smoking\_history | History of smoking | char | 1. “YES”,  2. “NO” |
| 12) tb\_history | History of prior TB disease | char | 1. “YES”,  2. “NO” |
| 13) tb\_hhc | Tuberculosis household contact | char | 1. “YES”,  2. “NO” |
| 14) tb\_screen | Any TB symptoms at baseline (at enrollment visit) | char | 1. “POSITIVE”, 2. “NEGATIVE” |
| 15) Virus | Binary indicator for presence of any viral pathogen on nasopharyngeal or oropharyngeal swab | char | 1. 0, 2. 1 |
| 16) Bacteria | Binary indicator for presence of any bacterial pathogen on nasopharyngeal or oropharyngeal swab | char | 1. 0, 2. 1 |
| 17)  av = Human adenovirus  mpneu = *Mycoplasma pneumoniae*  flua = Influenza A virus  flub = Influenza B virus  fluc = Influenza C virus  h1n1 = Influenza A (H1N1) virus (swine-lineage)  para1 = Human parainfluenza virus 1  para2 = Human parainfluenza virus 2  para3 = Human parainfluenza virus 3  para4 = Human parainfluenza virus 4  cor43 = Human coronavirus OC43  cor63 = Human coronavirus NL63  cor229 = Human coronavirus 229E  corhku = Human coronavirus HKU1  rsvab = Human respiratory syncytial viruses A/B  ev = Enteroviruses  hbov = Human bocavirus  rhino = Human rhinovirus  cmv = Cytomegalovirus  hmpvab = Human metapneumoviruses A/B  spneu = *Streptococcus pneumoniae*  salm = *Salmonella spp.*  haeinf = *Haemophilus influenzae*  saur = *Staphylococcus aureus*  pcp = *Pneumocystis jirovecii*  morax = *Moraxella catarrhalis*  hib = *Haemophilus influenzae B*  kpneu = *Klebsiella pneumoniae*  bord = *Bordetella spp.* (except *Bordetella parapertussis*)  cpneu = *Chlamydophila pneumoniae*  legio = *Legionella pneumophila* and *Legionella longbeachae*  PV = Human parechovirus | Binary indicator for presence of pathogen on nasopharyngeal or oropharyngeal swab | char | 1. 0, 2. 1 |

# Subject-Level PCR Dataset Data Dictionary with Example Values

The subject-level PCR datasets provide the TaqMan gene expression assay raw CT (cycle threshold) data from the Fluidigm microfluidic Gene Expression Integrated Fluidic Circuits (chips) with sample quality control results.

| **Variable Name** | **Variable Description** | **Format** | **Example Values** |
| --- | --- | --- | --- |
| 1) pubid | Unique identification number for each subject in study; matches the subject pubid in the study metadata | char | 1. 100001, 2. 356282 … |
| 2) CHIP\_BARCODE  “.x” or “.y” or “.z” indicates that the TaqMan primer-probe assays were run on the same Fluidigm microfluidic GE IFC | Unique identification number for each Fluidigm microfluidic Gene Expression (GE) Integrated Fluidic Circuit (IFC) run | num | 1. 1362241199, 2. 1691127095… |
| 3) SAMPLE\_QC  “.x” or “.y” or “.z” indicates that the TaqMan primer-probe assays were run on the same Fluidigm microfluidic GE IFC | Sample quality control result | char | 1. “PASS”, 2. “FAIL” |
| 4) ThermoFisher TaqMan Gene Expression Assays  e.g. GBP2.Hs00894846\_g1.y  GBP2 = Gene symbol  Hs00894846\_g1 = Assay ID  “.x” or “.y” or “.z” indicates that the TaqMan primer-probe assays were run on the same (or different) Fluidigm microfluidic GE IFC | Raw cycle threshold (CT) value for each TaqMan assay  “NA” implies a failed reaction or reaction which did not pass assay or sample quality control checks. | num | 1. 13.11255323151, 2. 17.65212010227, 3. 10.972573523… |