ISSN: 2754-4761

Journal of Pulmonology Research & Reports

SCIENTIFIC Research and Community

Research Article

Open Access

"TimBre" Phase-II Pilot Study Conducted Using Multi-Country Longitudinal Training Data for Screening of Pulmonary Tuberculosis Using Cough (Acoustic Sounds), Clinical & Demographic Inputs

Rahul Pathri*, Sasidhar R Sannadi, Samarth Tandon, Aparna Sykam

docturnal Inc

ABSTRACT

TimBre from docturnal offers screening for multiple lung diseases – Pulmonary Tuberculosis, Pneumonia & Lung Cancer. Detailed studies of TimBre in the past used a third-party Microphone Array with an XY arrangement that recorded high fidelity cough sounds with an average length of >5 seconds and basic demographic data such as Gender, Age, Height, Weight, BMI. In the current study, cough sounds were collected from 7 different countries (India, Vietnam, Philippines, Uganda, Tanzania, Madagascar, and South Africa) using Mobile Phones from different manufacturers & recorded solicited & longitudinal cough sounds for a duration of 0.5 seconds as a part of the phase-I study. We used longitudinally obtained cough sounds numbering 724,694 .WAV files with a resolution of 44.1 kHz & 16 bits. The duration was of 0.5 seconds, with the subject clinical and demographic variables added to the model, which anticipated an improved accuracy over the phase-1 study that used 10,000 plus cough files that obtained a sensitivity & specificity of 75.41% and 68.30% respectively with an AUC of 0.78. The current study (Phase-2) resulted in an overall sensitivity of 68.8%, a specificity of 70.4%, and an AUC of 70% when compared with the microbiological reference standard MRS (sputum culture or Xpert, 25% TB prevalence), and sensitivity of 73.8%, specificity of 69.5%, and an AUC of 71% compared to a sputum Xpert Ultra reference standard (22% prevalence). The phase-1 study used true labels of GeneXpert results as some of the culture results were labelled indeterminate. A comparison between the phases is unwarranted given the fact that the MRS definition is GeneXpert OR Culture while, in Phase-1 it is was strictly GeneXpert. The study and its earlier phase demonstrate the usage of mobile phone-based screening for low resource settings or home-based TB screening or countries that are addressing TB elimination or eradication goals.

***Corresponding author**

Rahul Pathri, docturnal Inc.

Received: April 15, 2025; **Accepted:** April 23, 2025, **Published:** April 30, 2025

Keywords: Cough Sounds, Pulmonary Tuberculosis, Sensitivity, Specificity, XGB, RUS, Ensemble, Spectral Analysis, FFT. STFT, XAI

Abbreviations

UCSF R2D2	University of San Francisco Rapid Research Development & Diagnostics
COPD	Chronic Obstructive Pulmonary Disorder
AUC / ROC	Area Under Curve / Receiver Operating Characteristic Curve
.WAV	Wave Audio File Format
TB	Tuberculosis
EMR	Electronic Medical Record
XY	An arrangement of Microphone Arrays on a Microphone
BMI	Body Mass Index
HIV	Human Immunodeficiency Viruses
PDP	Partial Dependence Plot
MRMR	Maximum Relevance Minimum Redundancy
RUS	Random Under Sampling
FFT	Fast Fourier Transformation
STFT	Short Time Fourier Transformation
ML / DL	Machine Learning / Deep Learning
CNN	Convolutional Neural Network
MFCC	Mel Frequency Cepstral Coefficient
Hz	Hertz
AWS	Amazon Web Services
XAI	Explainable Artificial Intelligence
AI	Artificial Intelligence
HR	Heart Rate
CI	Confidence Interval
XGB	Extreme Gradient Boosting
MRS	Microbiological Reference Standard
SXRS	Sputum Xpert Reference Standard (GeneXpert)

Introduction

Tuberculosis (TB), a communicable disease caused by Mycobacterium tuberculosis, is a major cause of ill health and one of the leading causes of death worldwide. To reduce the burden of disease, there is a critical need to improve TB screening and diagnosis by developing novel point-of-care tools that could be implemented in community settings. To address these challenges

and gaps, docturnal Inc developed TimBre, a low-cost, non-invasive triage tool based on cough sounds using an external third-party microphone array with an XY arrangement that recorded high fidelity cough sounds with an average length of >5 seconds and basic demographic data such as gender, age, height, weight and BMI. The clinical trials were confined to 1200 subjects that yielded a sensitivity of 80% and specificity of 92%, were not large enough. Most importantly, availability of Microphone Array at may be a challenge for at home or primary healthcare & for at home mobile phone screening, this study establishes accuracy metrics for public health stakeholders and decision makers [1].

The current Phase-2 study addresses the limitations of the pilot/phase-1 study where cough sounds were collected from 7 different countries (India, Vietnam, Philippines, Uganda, Tanzania, Madagascar, and South Africa) using mobile phones from different manufacturers & recorded solicited & longitudinal cough sounds for a duration of 0.5 seconds to the tune of 700,00+ .WAV files. A summary of the phase-1 study is depicted below [2].

Summary of Phase-1 Study: To assess the effectiveness of the TimBre machine learning (ML) model in screening pulmonary tuberculosis (TB) using cough acoustic features alongside clinical and demographic variables.

Study Design
- **Countries Involved:** India, Vietnam, Philippines, Uganda, Tanzania, Madagascar, South Africa
- **Participants:** Adults aged 18+ with a persistent cough (>2 weeks)
- **Data Source:** Hyfe Research app (recorded 0.5-second coughs on various mobile phones)
- **Diagnosis Confirmation:** Sputum-based molecular (Xpert MTB/RIF Ultra) and culture tests (MGIT or Lowenstein-Jensen)

Methodology
Two distinct modeling strategies were evaluated:

Per-Participant Model (Concatenated Coughs)
- Cough WAV files were concatenated per subject.
- Spectral and clinical features were extracted.
- **Clinical Variables Used:** Gender, Age, Height, Weight, BMI, Prior TB, Hemoptysis
- **Performance Metrics:**
  - Sensitivity: 68.6%
  - Specificity: 71.7%
  - AUC: 0.75

Per-Cough Model (Individual Snippets)
- Each 0.5s WAV file treated as a separate data point.
- Same clinical/demographic data repeated per snippet.
- **Additional Variables:** HIV status, Fever, Night Sweats, Heart Rate, Temperature
- **Performance Metrics at 95% CI:**
  - Sensitivity: 75.41%
  - Specificity: 68.30%
  - AUC: 0.78

Interpretability & Explainability
- Utilized MATLAB's Classification Learner with RUSBoost ensembles.
- **Key Predictive Features (from PDPs):**
  - BMI >22 decreased likelihood of TB.
  - Heart Rate >110 increased TB likelihood.
  - Presence of Night Sweats and Spectral Centroid variability provided additional separation between classes.

Discussion
- Including clinical variables significantly improved model performance.
- MFCC and FFT-based features offered strong baseline performance without requiring CNNs.
- Models built to maintain explainability and interpretability for clinical deployment.

Limitations
- Only solicited coughs were used.
- Mono-channel audio only; device variance unknown.
- Longitudinal data (~700K coughs) not yet included.

Materials and Methods

Training Data Characteristics
A total of 724694 .WAV files were downloaded from the synapse platform hosted by Sagebionetworks.org and mapped to the meta data for true labels (golden truth) & 280987 were POSITIVE TB & 443706 as confirmed by the Microbiological Reference Standard as defined in the data dictionary. While this seems slightly imbalanced from a class perspective, it did not mandate usage of RUS or ROS Ensemble models & hence XGB was preferred which also supports imbalanced class sets. No strong correlation was found between Prior TB, HIV AND TB positivity while BMI and HR were slightly correlated as observed during Phase-1 [2].

Validation Metrics
Here we leverage data collected from adults 18 years and older who presented to clinics across 7 countries with new or worsening cough for at least 2 weeks. Longitudinal cough sounds were recorded using the Hyfe Research app. Individuals were then comprehensively evaluated for TB with sputum-based molecular (Xpert MTB/RIF Ultra) and culture (MGIT or Lowenstein-Jensen) with microbiological reference standard (MRS) being (TB = positive culture or Xpert) and xpert-only reference standard (SXRS). The total number of training cough files were 724,694. The validation accuracy (Area Under Curve) was 0.96 for the best performing model, XGB (extreme Gradient Boosting) as compared with other models that were used as a reference point for comparison (Random Forest etc.). In addition to the spectral features, the following clinical and demographic variables were used in the model

Clinical & Demographic Variables
Gender
Age
Height
Weight
BMI
HIV Status
Prior TB
Hemoptysis
Fever
Night Sweats
Heart Rate

XGB Hyperparameters
'objective':'binary:logistic'
'max_depth': 3
'alpha': 15
'learning_rate': 0.01
'n_estimators':169
'random_state' :42

These are the baseline hyperparameters and further finetuning shall result in a better AUC for future submissions on the unseen TEST set. docturnal's production instance that uses Microphone Array to record cough sounds are tuned at different values for a RUS boosted model that may be updated to an XGB model in future

For Phase-1 of the study, RUS boosted model was used to predict unseen TEST data that handled imbalanced class sets efficiently & for Phase-2, XGB was used to predict unseen TEST data (figure 1).

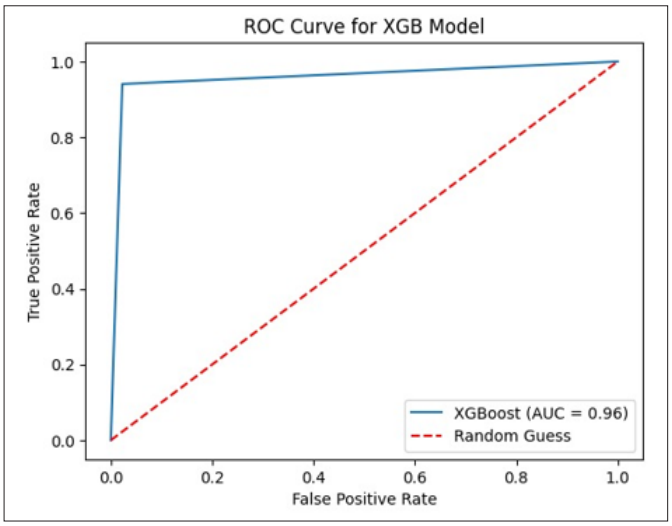

Figure 1: ROC for the Validation Data

Results

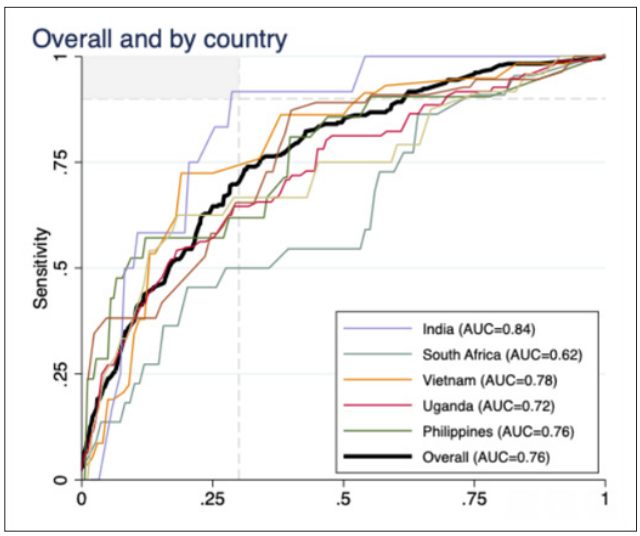

Figure 2: AUC for the 5 Countries and Overall AUC Representing a Cutoff Value of ≥ 0.3128 for India

Overall Accuracy Metrics

Table 1: Overall Accuracy Metrics (MRS Overall at 25% Prevalence)

Metric	Value	95% CI (Lower)	95% CI (Upper)
Prevalence (Pr(A))	25%	22%	27.4%
Sensitivity (Pr(+|A))	69.8%	64.1%	75%
Specificity (Pr(-|N))	70.4%	67.3%	73.4%
ROC Area ((Sens. + Spec.)/2)	0.701	0.67	0.732
Likelihood Ratio (+)	2.36	2.08	2.68
Likelihood Ratio (-)	0.429	0.358	0.514
Odds Ratio (LR(+)/LR(-))	5.5	4.11	7.35
Positive Predictive Value (Pr(A|+))	43.8%	39.2%	48.5%
Negative Predictive Value (Pr(N|-))	87.6%	84.9%	89.9%

Table 2: Overall Accuracy Metrics (SXRS Overall at 22% Prevalence)

Metric	Value	95% CI (Lower)	95% CI (Upper)
Prevalence (Pr(A))	22%	19%	24.3%
Sensitivity (Pr(+|A))	73.8%	68%	79.1%
Specificity (Pr(-|N))	69.5%	66.4%	72.5%
ROC Area ((Sens. + Spec.)/2)	0.717	0.686	0.748
Likelihood Ratio (+)	2.42	2.15	2.74
Likelihood Ratio (-)	0.376	0.305	0.464
Odds Ratio (LR(+)/LR(-))	6.44	4.72	8.79
Positive Predictive Value (Pr(A|+))	40.3%	35.8%	44.9%
Negative Predictive Value (Pr(N|-))	90.5%	88.1%	92.6%

Table 3: India Specific Prevalence

Metric	Value	95% CI (Lower)	95% CI (Upper)
Prevalence (Pr(A))	9%	4.7%	15.1%
Sensitivity (Pr(+|A))	91.7%	61.5%	99.8%
Specificity (Pr(-|N))	71.3%	62.4%	79.1%
ROC Area ((Sens. + Spec.)/2)	0.815	0.724	0.906
Likelihood Ratio (+)	3.2	2.3	4.43
Likelihood Ratio (-)	0.117	0.0178	0.766
Odds Ratio (LR(+)/LR(-))	27.3	4.31	-
Positive Predictive Value (Pr(A|+))	23.9%	12.6%	38.8%
Negative Predictive Value (Pr(N|-))	98.9%	93.8%	100%

For India specific prevalence of 9%, a cutoff value of >= 0.3128 was used to yield a sensitivity of 91% and specificity of 71% that is in line with the WHO target product profile for a screening modality that is point of care or near point of care.

Table 4: Comparison between Phase-1 & 2

	AUC %	Sensitivity %	Specificity %	Geographies	Model Ref Std	TEST Ref Std	Prevalence
Phase-1	78	75.41	68.30	7 countries	SXRS	MRS	-
Phase-2 (Overall)	70.1	69.80	70.4	5 Countries	MRS	MRS	25%
Phase-2 (Overall)	71.7	73.8	69.5	5 Countries	MRS	SXRS	22%
Phase-2 (India)	81.5	91.7	71.3	India	MRS	MRS	9%

Phase 1 included 2 additional countries – Madagascar & Tanzania. Prevalence details for the same are not represented in the table

Reference Standards

Current study demonstrates results verified against the gold standards – Sputum Culture & GeneXpert defined as: MRS, TB = positive culture or Xpert and Xpert-only reference standard (SXRS). With MRS prevalence of 25% and SXRS prevalence of 22%

Testing Metrics (Unseen Data)

A total of 10,008 cough sounds were evaluated against the model and the binary results along with the probability scores were reported to the UCSF team as done for Phase-1 study. In an independent validation, we achieved the following results:

ROC of the TEST Results Across Each Country

Out of the 7 countries, Madagascar & Tanzania were not reported and the best AUC was for India at 0.81+% followed by Vietnam and Philippines at 0.78 and 0.76 respectively with an overall AUC at 0.70+% with an overall sensitivity and specificity of 69.8% and 70.4% respectively. The results are slightly different from the phase-1 study that resulted in an AUC of 0.78% & an overall sensitivity & specificity of 75.41% and 68.30% respectively [3].

Discussion

Machine Learning & Explainability: The nature of the data was such that for each StudyID, redundancy was introduced wherein the cough recording snippet of 0.5 seconds was replicated with the clinical & demographic information for each StudyID aka per cough approach. This accuracy dropped significantly when these snippets were concatenated during phase-1 aka per StudyID approach & the redundancy was removed by having only a single record of clinical & demographic information for each StudyID. The per cough approach however did not result in ambiguity of results but were consistent & unique across each StudyID except for a single StudyID that was dropped from the results however. Segmentation of the .WAV files into 1 second did benefit a COPD study & was done in the current study as well (0.5 seconds). It did improve accuracy and explainability when added with the clinical & demographic information despite the redundancy [4]. As far as explainability is concerned, the clinical & demographic variables such as BMI, HR still played an important role as observed across both the phases while Spectral Centroid & a few others spectral variables (energy, skewness, variance etc.) extracted using STFT across multiple frequency bands ranging between 0-5000 Hz from a single channel with a sampling rate of 44.1 kHz 0.5 second .WAV file continued to provide clear demarcation for the binary class.

Algorithmic Approach: Across both the phases, Ensemble techniques/models performed better than expected and XGB specially outperforms deep learning models which however have a limitation in the current context in that they will not be able to leverage the clinical and demographic information that becomes an important part of explainability and data transparency for regulatory purposes. While Phase-1 used a 10-Fold cross validation using a RUS model, in the current study, cross validation was intentionally avoided to avoid result leakage across the folds for a given StudyID. However, multiple partitioned validation sets were tested to ascertain XGB accuracy over RUS & others that gave an AUC of 96% (figure 1 of results section). Subsequent submissions for the independent benchmarking hosted by Sagebionetworks on the Synapse platform shall fine tune the hyperparameters now that we have a champion model in place that shall implement a 10-Fold cross validation on a need basis for future evaluations.

Longitudinal Data: The addition of additional 724,694 cough files to the training set did not improve the AUC of the unseen

data. It slightly reduced the sensitivity but improved specificity for both MRS & SXRS. The possible explanation as observed could be that of the day-to-day variability. A controlled recording to avoid variance may circumvent this aspect such as a standard device, sitting position during cough recording, distance from the device etc [5]. The longitudinal cough collection entailed subjects carrying the mobile phone home and wearing around the neck [6]. The solicited cough however is more controlled and conducted in a clinical setting.

Most Importantly, the Phase-1 study used true labels of GeneXpert results as some of the culture results were labelled indeterminate. A comparison between the phases is unwarranted.

Cutoff Values: For phase-1, country specific data was not evaluated but for phase-2 we obtained 5- country specific results & the AUC for several countries performed better than overall AUC of Phase-1. The MRS & SXRS prevalence was at 25% (sensitivity/specificity of 69.8/70.4) & 22% (sensitivity/specificity of 73.8/69.5) respectively as represented in table 1 and table 2 respectively. Obtaining prevalence, cutoff values and sensitivity/specificity for remaining 4 countries shall be of great value and the current study obtained these values for India while reporting overall metrics for remaining 4 countries (Vietnam, Philippines, Uganda & South Africa) with South Africa being the lowest as seen from the figure 2 above & table 5 below. Based on a prevalence rate of 9% in India (table 3), a cutoff of ≥ 0.3128 was chosen by the UCSF PI team to arrive at the sensitivity & specificity of 91.7% and 71.3% respectively. This is in line with the WHO target product profile for non-sputum-based point of care or near point of care screening [7-11]. We also plan to use this cutoff for additional studies being conducted in India using the external XY Microphone Array.

Table 5

Country	AUC
India	0.84
Vietnam	0.78
Philippines	0.76
Uganda	0.72
South Africa	0.62

Table Values are Mapped to Figure 2 of Results Section

While day-to-day variance of the cough patterns is usually observed in a longitudinal data sets, the feature extraction approach and the true labels preserved the consistency of the results where in the consistency of the binary results was retained in that the result is either POSITIVE or NEGATIVE across multiple cough recordings for a single StudyID. Most importantly, in situations where there is a paucity of data (cough recordings), the 0.5 to 1 second cough snippet approach provide readily deployable ML models in the cough acoustics domain.

Next Steps

1. Evaluate against additional unseen TEST data as a part of Phase-3 of the study that shall be evaluated by Sagebionetworks.org (Montreal sites etc.) as a part of the Synapse independent benchmarking submission
2. Evaluate against a deep learning model converting the .WAV files into Spectrograms which however shall bereft the explainability aspect of clinical and demographic variables
3. Obtain country specific prevalences, cutoff & sensitivity/specificity values

Data Acknowledgement

1. "The datasets used for the analyses described were contributed by Dr. Adithya Cattamanchi at UCSF and Dr. Simon Grandjean Lapierre at University of Montreal and were generated in collaboration with researchers at Stellenbosch University (PI Grant Theron), Walimu (PIs William Worodria and Alfred Andama); De La Salle Medical and Health Sciences Institute (PI Charles Yu), Vietnam National Tuberculosis Program (PI Nguyen Viet Nhung), Christian Medical College (PI DJ Christopher), Centre Infectiologie Charles Merieux Madagascar (PIs Mihaja Raberahona & Rivonirina Rakotoarivelo), and Ifakara Health Institute (PIs Issa Lyimo & Omar Lweno) with funding from the U.S. National Institutes of Health (U01 AI152087), The Patrick J. McGovern Foundation and Global Health Labs."
2. Solicited Training and Scoring data .WAV files used Hyfe.ai app - https://www.hyfe.ai/ [7].
3. We sincerely thank Dr. Devan Jaganath and Rebecca Crowder (University of California San Francisco) for analysis of the TEST set results & support for Phase-1 and Phase-2 publication reviews (including prevalence & cutoff values) across multiple iterations & countries based on posterior probability scores provided by TimBre algorithm
4. We sincerely thank Dr. Sophie Huddart (University of California San Francisco) for providing the calculated TEST results (including prevalence & cutoff values) across multiple iterations & countries based on posterior probability scores provided by TimBre algorithm
5. We sincerely thanks Dr. Seda Yerlikaya for providing us with the TRAINING and TEST (unseen) data
6. All data clinical, demographic data was collected in accordance with the good clinical practices (GCP) 21 CFR 812.28 by the UCSF Rapid Research in Diagnostic Development TB Network (R2D2 TB Network) [3].

Funding & Support

1. "Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health through award U01AI152087".
2. For Montreal Partner Sites: "Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health through award U01AI152087, The Patrick J. McGovern Foundation and the Global Health Labs".
3. For docturnal, funding & support was provided by "FICCI / Millennium Alliance Covid19 Challenge (MA/R6/SA/2020/0042)" for Phase-1.
4. For Phase-2, funding support provided by USISTEF (United States-India Science & Technology Endowment Fund USISTEF/AI/145/2023.

References

1. Rahul Pathri (2022) Acoustic Epidemiology of Pulmonary Tuberculosis (TB) & Covid19 Leveraging Explainable AI/ML. Journal of Pulmonology Research & Reports 4: 1-6.
2. Rahul Pathri, Shekhar Jha, Ram Samudrala, Aparna Sykam (2023) "TimBre" Pilot Study Conducted Using Multi-Country Training and Validation Data for Screening of Pulmonary Tuberculosis Using Cough (Acoustic Sounds), Clinical & Demographic Inputs. Journal of Pulmonology Research & Reports 5: 1-7.
3. GCP 21 CFR 812.28-https://www.medrxiv.org/content/10.1101/2024.03.27.24304980v1.
4. https://www.pressetext.com/news/first-ai-based-continuous-

ambulatory-cough-monitoring-system-approved-as-medical-device.html?utm_source=chatgpt.com.

5. Windmon A, Minakshi M, Chellappan S, R Athilingam, P Johansson, et al. (2018) On Detecting Chronic Obstructive Pulmonary Disease (COPD) Cough using Audio Signals Recorded from Smart-Phones. In Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018) 156.
6. Training and Scoring data of 0.5 second snippet .WAV files used-https://www.hyfe.ai/.
7. Lucy Asege, Hai Dang, Brigitta Deredinger, Marissa Golla, Sophie Huddart, et al. (2022) Feasibility and analysis considerations of continuous community-based cough monitoring in low- and middle-income settings. MedRxiv https://www.medrxiv.org/content/10.1101/2022.07.06.22277224v1.
8. World Health Organization (2024) Global Tuberculosis Report 2024. WHO https://www.who.int/publications/i/item/9789240101531.
9. Pahar M, Klopper M, Reeve B, Warren R, Theron G, et al. (2021) Automatic cough classification for tuberculosis screening in a real-world environment. Physiol Meas 42: 10.
10. Botha GHR, Theron G, Warren RM, Klopper M, Dheda K, et al. (2018) Detection of tuberculosis by automatic cough sound analysis. Physiol Meas 39: 045005.
11. Coda TB dream challenge - CODA TB DREAM Challenge - syn31472953 - Tables (synapse.org; https://sagebionetworks.org).