Open Forum Infectious Diseases

MAJOR ARTICLE

Accelerating Cough-Based Algorithms for Pulmonary Tuberculosis Screening: Results From the CODA TB DREAM Challenge

Devan Jaganath,[1,2,a] Solveig K. Sieberts,[3,a] Mihaja Raberahona,[4,5,a] Sophie Huddart,[2,6] Larsson Omberg,[3] Rivo Rakotoarivelo,[7,8] Issa Lyimo,[9] Omar Lweno,[9] Devasahayam J. Christopher,[10] Nguyen Viet Nhung,[11,12] William Worodria,[13] Charles Yu,[14] Jhih-Yu Chen,[15] Sz-Hau Chen,[16,17] Tsai-Min Chen,[18,19] Chih-Han Huang,[20] Kuei-Lin Huang,[21] Filip Mulier,[22] Daniel Rafter,[22] Edward S. C. Shih,[23] Yu Tsao,[18,19] Hsuan-Kai Wang,[24] Chih-Hsun Wu,[25] Christine Bachman,[26] Stephen Burkot,[26] Puneet Dewan,[26] Sourabh Kulhare,[26] Peter M. Small,[27,28] Vijay Yadav,[3] Simon Grandjean Lapierre,[29,30,b] Grant Theron,[31,32,b] and Adithya Cattamanchi[2,33,b]; on behalf of the Cough Diagnostic Algorithm for Tuberculosis (CODA TB) DREAM Challenge Consortium[c]

[1]Division of Pediatric Infectious Diseases, University of California, San Francisco, California, San Francisco, USA, [2]Center for Tuberculosis, University of California, San Francisco, California, USA, [3]Sage Bionetworks, Seattle, Washington, USA, [4]CHU Joseph Rasera Befelatanana, Antananarivo, Madagascar, [5]Centre D'Infectiologie Charles Mérieux, Université D'Antananarivo, Antananarivo, Madagascar, [6]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California, USA, [7]CHU Tambohobe Fianarantsoa, Haute-Matsiatra, Madagascar, [8]Université de Fianarantsoa, Fianarantsoa, Madagascar, [9]Environmental and Ecological Sciences & Interventions and Clinical Trials Departments, Ifakara Health Institute, Dar es Salaam, Tanzania, [10]Department of Pulmonary Medicine, Christian Medical College, Vellore (Ranipet Campus), Tamil Nadu, India, [11]National Tuberculosis Programme, Hanoi, Vietnam, [12]VNU University of Medicine and Pharmacy, Hanoi, Vietnam, [13]Walimu, Kampala, Uganda, [14]De La Salle Medical and Health Sciences Institute, Dasmarinas Cavite, Philippines, [15]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan, [16]Industrial Information Department, Development Center for Biotechnology, Taipei, Taiwan, [17]Investment & Wealth Management, FCC Partners, Taipei, Taiwan, [18]Graduate Program of Data Science, National Taiwan University and Academia Sinica, Taipei, Taiwan, [19]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan, [20]Department of Data Science, ANIWARE, Taipei, Taiwan, [21]School of Medicine, China Medical University, Taichung, Taiwan, [22]Flywheel.io, Minneapolis, Minnesota, USA, [23]Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan, [24]Independent Researcher, Taipei, Taiwan, [25]Artificial Intelligence and E-Learning Center, National Chengchi University, Taipei, Taiwan, [26]Global Health Labs, Bellevue, Washington, USA, [27]Department of Global Health, University of Washington, Seattle, Washington, USA, [28]Hyfe, Seattle, Washington, USA, [29]Centre de Recherche du Centre Hospitalier de L'Université de Montréal, Immunopathology Axis, Montreal, Quebec, Canada, [30]Department of Microbiology, Infectious Diseases and Immunology, Université de Montréal, Montreal, Quebec, Canada, [31]DSI-NRF Centre of Excellence for Biomedical Tuberculosis Research, South African Medical Research Council Centre for Tuberculosis Research, Cape Town, South Africa, [32]Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, South Africa, and [33]School of Medicine, University of California Irvine, Orange, California, USA

Background. Open-access data challenges can accelerate innovation in artificial intelligence–based tools. In the Cough Diagnostic Algorithm for Tuberculosis (CODA TB) DREAM Challenge, we developed and independently validated cough sound–based artificial intelligence algorithms for tuberculosis screening.

Methods. We included data from 2143 adults with ≥2 weeks of cough from outpatient clinics in India, Madagascar, the Philippines, South Africa, Tanzania, Uganda, and Vietnam. A standard tuberculosis evaluation was completed, and ≥3 solicited coughs were recorded using a smartphone. We invited teams to develop models using training data to classify microbiologically confirmed tuberculosis disease using (1) cough sound features only and/or (2) cough sound features with routinely available clinical data. After 4 months, they submitted the algorithms for independent test set validation. Models were ranked by area under the receiver operating characteristic curve (AUROC) and partial AUROC (pAUROC) to achieve at least 80% sensitivity and 60% specificity.

Results. Eleven cough models and 6 cough-plus-clinical models were submitted. AUROCs for cough models ranged from 0.69 to 0.74, and the highest performing model achieved 55.5% specificity (95% confidence interval, 47.7%–64.2%) at 80% sensitivity. The addition of clinical data improved AUROCs (range, 0.78–0.83); 5 of the 6 models reached the target pAUROC, and the highest performing model had 73.8% specificity (95% confidence interval, 60.8%–80.0%) at 80% sensitivity. The AUROC varied by country and was higher among male and human immunodeficiency virus–negative individuals.

Conclusions. In a short period, an open-access data challenge facilitated the development of new cough-based tuberculosis algorithms and demonstrated potential as a tuberculosis screening tool.

Keywords. artificial intelligence; cough; data challenge; diagnostics; tuberculosis.

Received 02 June 2025; accepted 04 September 2025; published online 16 September 2025

[a]D. J., S. K. S., and M. R. contributed equally to this work.

[b]S. G. L., G. T., and A. C. contributed equally to this work.

[c]CODA TB DREAM Challenge Consortium coauthors are listed in the Acknowledgments.

Correspondence: Adithya Cattamanchi, MD, MAS, Department of Medicine, University of California, Irvine, 1001 Health Sciences Rd, Irvine, CA 92697-3950 (acattama@hs.uci.edu); Devan Jaganath, MD, MPH, Department of Pediatrics, University of California, San Francisco, 550 16th St, Fourth Floor, San Francisco, CA 94158 (Devan.jaganath@ucsf.edu).

Open Forum Infectious Diseases®

https://doi.org/10.1093/ofid/ofaf572

As ongoing challenges in global infectious diseases intersect with rapid advancements in technology and artificial intelligence (AI), digital health tools have the potential to enhance disease surveillance, diagnosis, and management [1, 2]. The widespread availability of smartphones and wearable sensors create opportunities for low-cost, noninvasive applications to increase healthcare access and quality [3]. However, a major challenge to equitable implementation of these tools is a lack of available datasets from diverse geographic settings, and limited focus on conditions that disproportionally affect low- and middle-income countries [2]. Moreover, available datasets may be proprietary, preventing open-access sharing and transparent algorithm development. The consequence is a dearth of AI tools validated in low- and middle-income countries that address the public health challenges they face.

Tuberculosis is the leading cause of death from infectious disease worldwide [4]. The high mortality rate is driven by a large case detection gap, in which tuberculosis disease has not been diagnosed or reported to public health programs in 3.1 million of the estimated 10 million individuals in whom it occurs each year [4]. AI has already supported tuberculosis diagnosis through automated reading of chest radiographs [5], but the required infrastructure limits implementation at primary health facilities. Cough is a common symptom of tuberculosis, and initial studies suggest that unique acoustic features can distinguish pulmonary tuberculosis from other respiratory conditions [6, 7]. Furthermore, cough detection applications have already been developed for mobile phones and smart watches [8], providing an opportunity to integrate cough-based AI algorithms for point-of-care tuberculosis assessment by providers and patients. In other diseases, including coronavirus disease 2019 (COVID-19) [9], open-access, crowdsourced data challenge initiatives have been used to accelerate the development of novel algorithms [10].

To expedite AI diagnostic development for tuberculosis, we established a cough sound repository from individuals prospectively enrolled with presumptive tuberculosis across 7 high tuberculosis-burden countries and implemented the Cough Diagnostic Algorithm for Tuberculosis (CODA TB) DREAM Challenge [11].

METHODS

CODA TB DREAM Challenge

The CODA TB DREAM Challenge launched on 26 October 2022. Participants were asked to develop a model to classify tuberculosis disease in 2 subchallenges: (1) using cough sounds alone and (2) using cough sounds and basic demographic and clinical variables. Participants were required to submit an outline of their methods and source code. The timeline of the challenge is shown in Supplementary Figure 1.

The challenge was hosted by Sage Bionetworks, which has developed an open-science, collaborative competition framework for evaluating and comparing computational algorithms, using the DREAM Challenges framework. DREAM focuses exclusively on biomedicine with an explicit mandate for transparency, openness, and collaboration. The challenge was set up on Synapse (www.synapse.org/tbcough), which provided all instructions, a secure platform for data sharing, a forum for communication with challenge participants, and supported submission of models for independent validation.

Any individual or team could participate in the challenge after certifying that they understood the Synapse data use policy, verifying their identify, and agreeing to the challenge guidelines to not attempt to identify or contact any study participants, to not share the data with others, and to comply with the intended use of the data. The challenge was broadly advertised, including on social media, to multiple academic institution listservs and departments of global health, bioinformatics and computer science, companies interested in cough-based or tuberculosis diagnosis, and previous DREAM Challenge participants.

Study Dataset

Data for the CODA TB DREAM Challenge were obtained from 2 multicountry tuberculosis diagnostic evaluation studies [11]. The Rapid Research in Diagnostic Development TB Network (R2D2 TB Network) enrolled participants at outpatient health centers in Uganda, South Africa, Vietnam, the Philippines and India [12], and the Digital Cough Monitoring Project enrolled participants in Tanzania and Madagascar [13].

Participants ≥18 years old presenting with new or worsening cough for ≥2 weeks were enrolled and evaluated with a clinical questionnaire and examination, sputum-based molecular testing (Xpert MTB/RIF Ultra; Cepheid) and liquid or solid medium culture testing. Participants were asked to produce ≥3 solicited cough sounds during the baseline visit before any tuberculosis treatment initiation. The coughs were collected on an Android-based smartphone using the Hyfe Research app [14], which uses a convolutional neural network model to automatically detect the cough and saves the 0.5-second peak sound [11]. Solicited cough sounds were collected, though any additional naturally triggered passive coughs were also recorded in order to emulate real-world collection and analysis by a cough sound application. Although the repository includes cough sounds that were collected longitudinally [11], for the challenge we focused only on the coughs collected at enrollment. Tuberculosis disease status was based on a microbiological reference standard, defined by a positive molecular or culture result. Further details on the study procedures and dataset, including a description of participant demographics and country distribution, have been published elsewhere [15]. A summary of clinical and demographic characteristics are shown in Supplementary Table 1.

A training set (n = 1105) for algorithm development was created by taking a 50% sample of the dataset randomized at the

individual level. Of the remaining data, 24% (n = 248) were randomly selected at the individual level for a "leaderboard" test set, from which challenge participants could receive periodic feedback on their model performance, and the remainder (n = 790) was reserved as the final test set for algorithm evaluation. Challenge teams were given direct access only to the training set, which included the raw peak cough sound recordings in WAV format, as well as associated age, sex, height, weight, smoking status, self-reported duration of cough, prior history of tuberculosis, common tuberculosis symptoms (hemoptysis, fever, night sweats, weight loss), heart rate, and temperature. These variables were chosen as data that would be readily available in routine primary care settings. Human immunodeficiency virus status was not included as the testing and/or results may not be available or known at the time of cough assessment.

Patient Consent Statement

All participants provided written informed consent for study participation, cough recording and anonymized data sharing. Ethical approvals for the studies were obtained from institutional review boards in the US (R2D2 TB Network, University of California, San Francisco) and Canada (Digital Cough Monitoring Project, University of Montreal), as well as institutional review boards in each participating country.

Algorithm Development and Evaluation

Participating teams could train an algorithm using any preprocessing approach and model, and with any programming language (eg, R, Python, etc) or framework (eg, Keras or Pytorch). For evaluation, models were required to be saved in Open Neural Network Exchange format and submitted in a Docker container, with any code needed for preprocessing the data.

Challenge teams had 5 interim opportunities to evaluate their algorithms on the "leaderboard" test set before the final algorithms were due for test set evaluation (Supplementary Figure 1). The output of each model was continuous tuberculosis prediction scores used to calculate the area under the receiver operating characteristic curve (AUROC). A limitation of the AUROC is that it considers the accuracy across all thresholds, including those that are not clinically relevant [16]. The partial AUROC [17] (pAUROC) addresses this by measuring the AUROC within predefined sensitivity and specificity targets, and a higher pAUROC indicates that a greater area falls within these targets.

We calculated a 2-way pAUROC within the thresholds of 80% sensitivity and 60% specificity, in order to identify promising algorithms that had an accuracy that was at least within 10% of the minimum World Health Organization (WHO) target product profile (TPP) accuracy for a tuberculosis screening test (≥90% sensitivity and ≥70% specificity) [18]. If no model receiver operating characteristic curve fell within the pAUROC targets, the algorithms were evaluated by the total AUROC. Variability in the AUROCs and pAUROCs was assessed via bootstrap resampling (n = 1000). Challenge teams submitted their preprocessing code (if applicable) and models 4 months after the launch of the challenge. We then independently applied each model to the test set to calculate the AUROC and pAUROC with 95% confidence intervals (CIs). We also calculated the maximum specificity at 80% or 90% sensitivity for each subchallenge, with 95% CIs.

Clinical Data Only Model

As a sensitivity analysis to assess the degree that the clinical variables alone contributed to the models in subchallenge 2, we developed a random forest model [19] using the clinical and demographic variables provided to challenge participants. In addition to the variables provided, body mass index was computed from height and weight variables, and the duration of cough symptoms was log-transformed before model fitting. The model was trained using 1000 trees.

Subgroup Analyses

After the challenge was complete, first- and second-ranked teams in both subchallenges (n = 5 due to ties) were invited to participate in additional model evaluation. We assessed the accuracy of the models by country, sex, and human immunodeficiency virus (HIV) status. We also compared the probability of tuberculosis classification for each model by Xpert MTB/RIF Ultra polymerase chain reaction (PCR) semiquantitative category.

The original model evaluation was performed with Python software (version 3.8.8). All subsequent analyses were performed with R software, version 4.2.2 (2022-10-31). Evaluations of model statistics were done using the pROC R package. Implementation of the pAUROC was provided by Chaibub Neto et al [20, 21].

RESULTS

Dataset Summary

As shown in Supplementary Table 1, randomization without stratification was sufficient to balance clinical and demographic variables across the training (n = 1105) data set and test data set (n = 1038; 248 for leadership board and 790 for final test set). The median ages in the training and test sets were 40 (interquartile range, 28–53) and 40 (29–53) years, respectively; 46.8% in the training and 44.3% in the test set were female. The prevalence of HIV (14.7% and 14.9%, respectively) and prior history of tuberculosis (81.7% and 80.4%) were also balanced between training and test sets, as was the proportion with microbiologically confirmed tuberculosis (26.9% and 24.7%, respectively). The country distribution was largely balanced,

although there was a lower proportion from Madagascar in the test set than in the training set (7.2% vs 14.4%). There were a total of 18 834 coughs (9772 in training and 9062 in test).

Challenge Implementation

In total, 147 individuals and 18 teams registered for the CODA TB DREAM Challenge. In each leaderboard round, 2–8 teams submitted models. Thirteen teams submitted final models for subchallenge 1, and 8 for subchallenge 2. Of those that submitted final models, 11 teams for subchallenge 1 and 6 for subchallenge 2 submitted a summary of methods and model code. The winning models for each subchallenge are described in the Supplementary Data. The reports for the full set of submissions are available through the challenge website [22].

Subchallenge 1

As shown in Table 1, Figure 1*A*, and Supplementary Figure 2, AUROCs ranged from 0.689 to 0.743. The top model achieved a specificity of 55.5% at 80% sensitivity; as further shown in Supplementary Table 2, it did not reach the pAUROC thresholds and did not achieve the WHO TPP-based accuracy goal for a screening test at 90% sensitivity and 70% specificity. Of the 11 groups, 4 (36%) used convolutional neural networks, 4 (36%) used artificial neural networks, and 3 (27%) used gradient-boosting decision tree methods.

Subchallenge 2

All groups used the same algorithm approach they used in subchallenge 1. As shown in Figure 1*B*, Supplementary Figure 3, and Table 2, overall performance improved compared with the use of cough sounds alone, and the top performing model achieved an AUROC of 0.832 (95% CI, .795–.863) and a pAUROC of 0.003 (6.1×10^{-6} to .012). Five of the 6 submissions (83%) achieved at least 80% sensitivity and 60% specificity, with the top model reaching 73.8% (95% CI, 60.8%–80.0%) at 80% sensitivity. Considering the WHO TPP as a tuberculosis screening test, the top-performing model achieved 54% specificity (95% CI, 38%–63%) at 90% sensitivity (Supplementary Table 2). In sensitivity analysis, the clinical data–only model achieved an AUROC of 0.817 (95% CI, .778–.850) and a pAUROC of 0.004 (5.5×10^{-4} to .010). At 80% sensitivity, the specificity was 68.1% (95% CI, 63.4%–75.5%; Supplementary Table 2).

Subgroup Assessment

By country, the AUROC for subchallenge 2 ranged from 0.73 to 0.83, with a higher median AUROC in the Philippines, Uganda, Tanzania and Vietnam (Figure 2*A*). The median AUROC for the cough and clinical data models was slightly higher for male than for female participants (0.82 vs 0.78; $P < .01$) (Figure 2*B*). Model performance was also slightly higher among people without HIV than among those with HIV (median AUROC, 0.83 vs 0.78; $P < .01$) (Figure 2*C*). Subgroup results for subchallenge 1 (cough sounds only) are shown in Supplementary Figures 4–7. Findings were similar, although we found slightly lower accuracy in male than in female participants (median AUROC, 0.69 vs 0.71; $P = .02$), in contrast to subchallenge 2.

For all submitted cough and clinical data models, the median predicted probability of being tuberculosis positive increased with Xpert MTB/RIF Ultra semiquantitative category, from trace positive results to high bacillary load results (Figure 3 and Supplementary Figure 7).

DISCUSSION

The CODA TB DREAM Challenge addressed a critical need to accelerate the development of AI-based tools for tuberculosis screening through an inclusive, open, and transparent approach. The challenge brought together students, researchers, and industry partners from a diverse geographic spectrum with a common goal of developing novel tuberculosis diagnostic algorithms using cough sounds. In a short period, challenge participants created, tested, and improved cough sound–based algorithms that approached the WHO TPP accuracy targets for a high-sensitivity tuberculosis screening test. Open-access research and citizen science represent a potential paradigm shift in how digital health solutions can be developed for global health and infectious diseases.

The cough sound–only models had similar accuracies, with AUROCs ranging from 0.65 to 0.74. This performance is within the wide range of previously developed cough-based COVID-19 models (AUROC, 0.62–0.98) [23–26] A few published cough sound tuberculosis models have shown higher performance (AUROC, 0.79–0.94) [6, 7], but these were small studies and need further validation. A limitation of previous cough models for other conditions was the use of crowd-sourced data [9, 27, 28]. While this approach rapidly generates large real-world datasets, there are multiple challenges, including selection bias, subjective clinical assessment, and heterogenous reference standard definitions. In the CODA TB DREAM Challenge, we used a multicountry cohort of consecutively enrolled symptomatic individuals, standardized clinical data and cough collection protocols, objective tuberculosis testing, and uniform case definitions. Our approach increases the confidence that algorithms are identifying features specific to the disease condition, reduces AI-related biases, and better reflects how the algorithms will perform in the intended settings and populations.

Performance improved when routine demographic and clinical variables were added to models. Five of 6 algorithms approached the WHO-established target accuracy thresholds for a tuberculosis screening test by meeting the pAUROC targets, although the absolute values were small, and external validation is needed. As a postchallenge sensitivity analysis, we developed a clinical data–only model that performed well

Table 1. Model Performance for Cough Sound–Only Model (Subchallenge 1)

Rank[a]	Team	AUROC (95% CI)	Model Type	Sound Features Used[b]
1	Blue Team	0.743 (.703–.780)	CNN	Spectrogram
2	AI-Campus High School Team	0.731 (.691–.771)	Gradient-boosting decision tree	MFCCs, chromagram
2	Raghava_India_TB	0.730 (.690–.773)	CNN	Mel spectrogram
4	Yuanfang Guan Lab Team	0.727 (.685–.768)	Light gradient–boosting machine	MFCC, first- and second-order time derivatives of MFCCs, magnitude-of-pitch tracking, total no. of coughs recorded
5	Metformin-121	0.704 (.660–.746)	MetforNet[c]	*z*-Score normalization of cough recordings
6	Clare	0.699 (.655, .746)	ANN	Top 300 features extracted via OpenSMILE and identified using PCA
7	Sakb	0.695 (.654–.739)	ANN	Top 1024 features extracted via OpenSMILE and identified using PCA
7	chsxashoka	0.693 (.651–.736)	ANN	MFCC, mel spectrogram
9	LCL	0.689 (.644, .733)	CNN, light gradient–boosting machine	Zero-crossing rate, MFCC, chromagram, mel spectrogram, root mean square
9	sasgarian	0.689 (.647–.732)	ANN	Top 1024 features extracted via OpenSMILE identified using PCA
11	yhwei	0.645 (.601–.687)	CNN	Spectrogram

Abbreviations: ANN, artificial neural network; AUROC, area under the receiver operating characteristic curve; CI, confidence interval; CNN, convolutional neural network; MFCC, mel-frequency cepstral coefficient; PCA, principal component analysis.

[a]Models with the same ranking (2, 7 and 9) were statistically indistinguishable.

[b]Key definitions: A spectrogram is a visualization of the audio frequency and amplitude over time, with a chromagram visualization focused on pitch categories; MFCCs are acoustic features derived from the mel scale that approximate how humans perceive sound; and the zero-crossing rate is the number of times an audio signal will change from positive to negative within the time period.

[c]A combined architecture of 5 CNN blocks, followed by a bidirectional gated recurrent unit, an attention layer, and a fully connected layer.

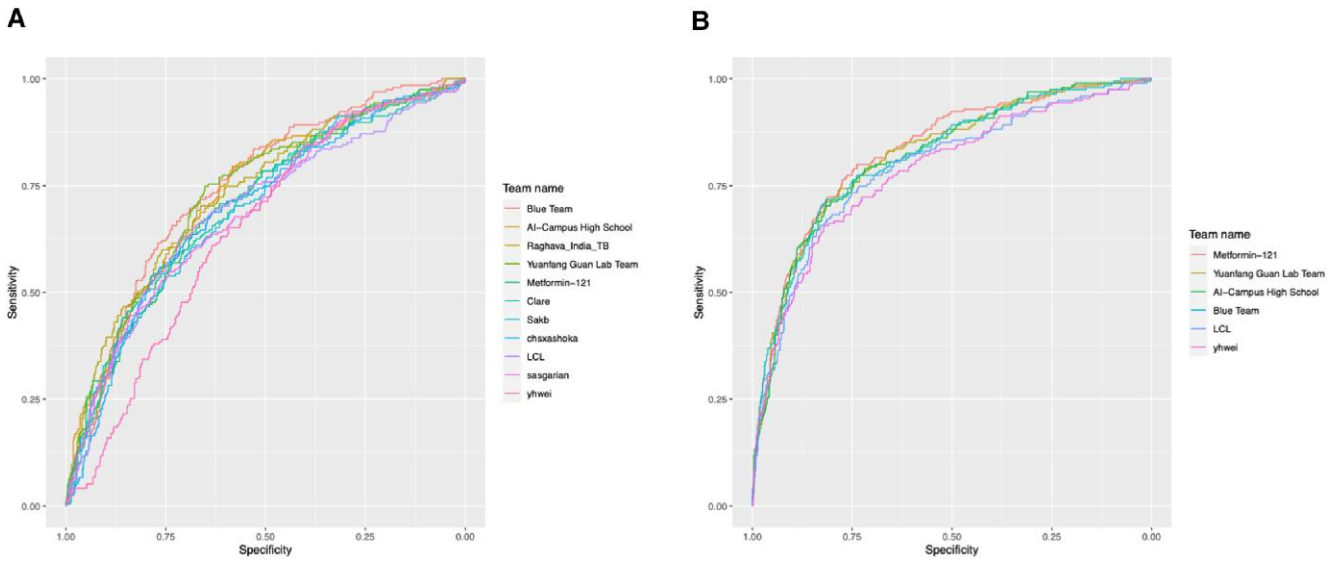

Figure 1. Area under the receiver operating characteristic curves for Cough Diagnostic Algorithm for Tuberculosis (CODA TB) DREAM Challenge final models. *A.* Subchallenge 1—cough sounds only. *B.* Subchallenge 2—cough sounds and routine demographic and clinical data.

(AUROC, 0.817); the top model that combined clinical and cough data had higher AUROC and specificity estimates, although with overlapping CIs with the clinical data–only model. This highlights that clinical care pathways are valuable for tuberculosis assessment, and cough sound models should be integrated and optimized within them to further augment accuracy.

The best-performing models used deep learning algorithms; while interpretability can be limited with such models, subgroup findings increase confidence in a tuberculosis-specific signal. For example, lower performance could be expected among people with HIV as they often have paucibacillary disease, suggesting that different thresholds may be needed depending on the setting or target group [29–31]. Moreover, the probability of tuberculosis classification correlated with bacterial burden as measured by semiquantitative PCR results in both subchallenges, which was also seen in a study in Kenya [6].

It is important to recognize that the final submitted models were developed rapidly over a short time frame, and there is potential for further optimization. This includes exploring more complex deep learning architectures and/or ensembles, increasing the size of the training set and developing country-specific models. We did not include HIV test results in the challenge as it may not be available at assessment, but given the variation in accuracy by HIV status and given that it is an important tuberculosis risk factor, inclusion of HIV results or creation of HIV-stratified models may further bolster performance. The cough sound repository could also be expanded and improved in future collection, including more coughs per participant and using the whole duration of cough sounds rather than the peak period. Passive cough sounds were collected only if the individual naturally coughed after the forced cough, but there are differences between solicited and passive cough sounds that could be more systematically compared for model development [6].

We did not include longitudinal cough sounds for this challenge as the goal was to assess its role for tuberculosis screening, but these data may provide additional tuberculosis-specific features to improve performance. At the same time, the overarching goal of the challenge was to accelerate innovation and gain key insights into cough-based AI models for tuberculosis. In 4 months, the challenge (1) supported multiple new and

Table 2. Model Performance for Cough Sound and Clinical Data Model (Subchallenge 2)

Rank	Team	pAUROC (95% CI)	AUROC (95% CI)
1	Metformin-121	0.003 (6.11 × 10^{-6} to .012)	0.832 (.795–.863)
2	Yuanfang Guan Lab Team	0.003 (0–.009)	0.821 (.784–.853)
3	AI-Campus High School Team	0.001 (0–.008)	0.817 (.778–.850)
4	Blue Team	0.001 (0–.007)	0.818 (.779–.853)
5	LCL	0.001 (0–.006)	0.792 (.750–.829)
6	yhwei	0 (0–.003)	0.784 (.741–.822)

Abbreviations: AUROC, area under the receiver operating characteristic curve; CI, confidence interval; pAUROC, partial AUROC.

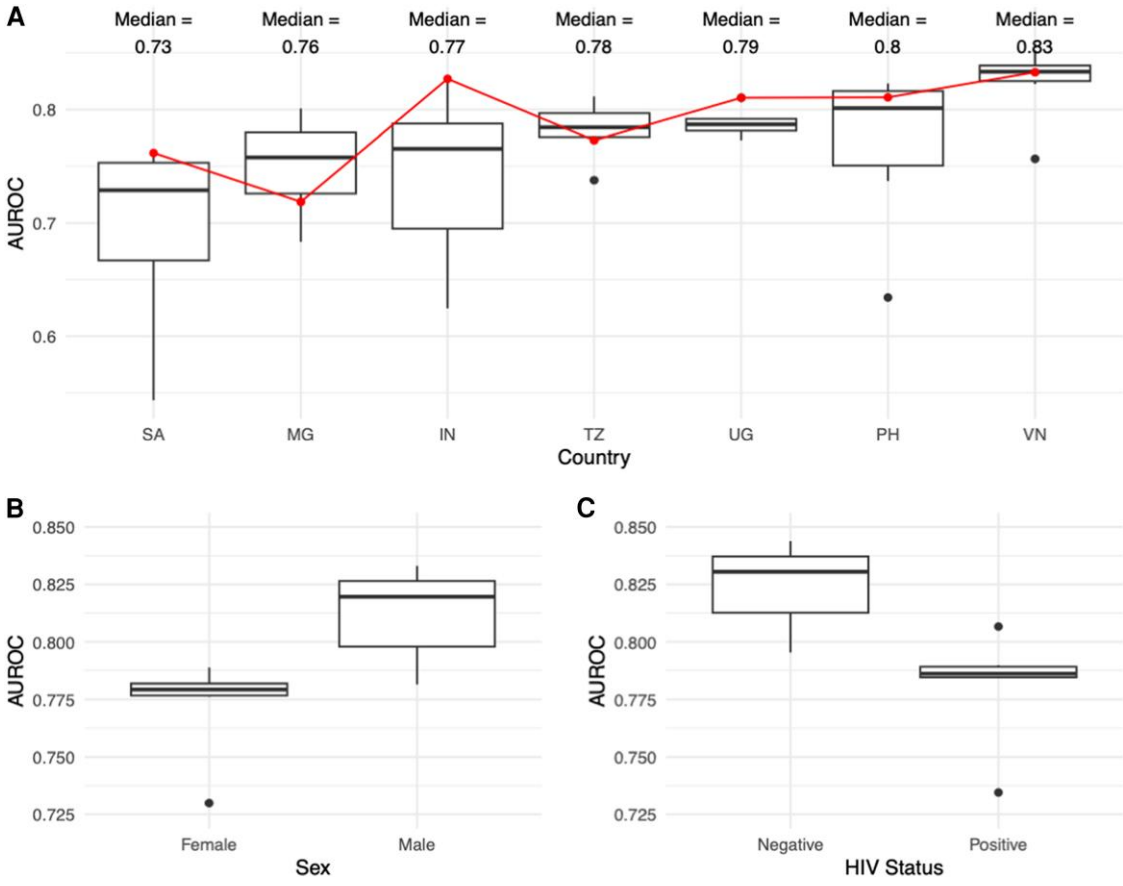

Figure 2. Comparison of area under the receiver operating characteristic curve (AUROC) by country and subgroup in subchallenge 2, shown as box plots of median AUROC with interquartile range, based on all submissions. *A.* Stratified by country: South Africa (SA), n = 98; Madagascar (MG), n = 51; India (IN), n = 86; Tanzania (TZ), n = 87; Uganda (UG), n = 187; Philippines (PH), n = 150; and Vietnam (VN), n = 131 (median AUROC shown at the top; winning model AUROC shown as a connected line). *B.* Stratified by sex: female, n = 362, male n = 428. *C.* Stratified by human immunodeficiency virus (HIV) status: HIV negative, n = 615; HIV positive, n = 115.

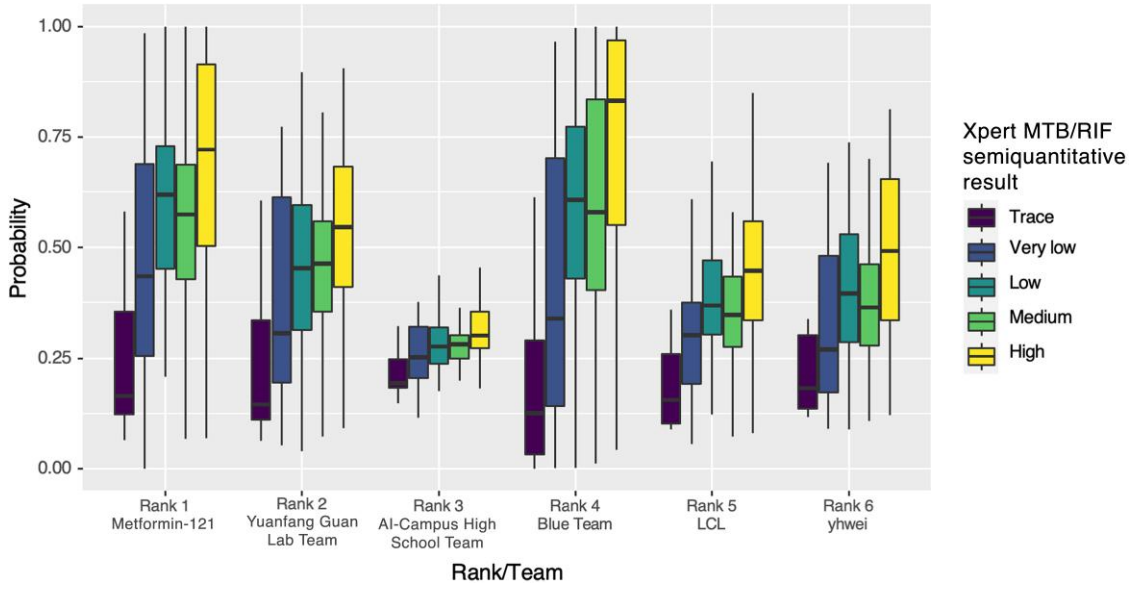

Figure 3. The probability of each cough and clinical model to classify tuberculosis, stratified by Xpert semiquantitative category for subchallenge 2. Box plot shows median probability with interquartile range. Rank indicates the final challenge ranking by team. Higher probability scores indicate higher likelihood that the model would classify the individual as having tuberculosis.

independently validated cough sound algorithms that could discriminate tuberculosis disease, (2) demonstrated that clinical data could augment performance, and (3) transparently shared the best-performing algorithms and processing methods.

To further facilitate ongoing model development, the dataset remains open source and can be downloaded at the challenge website [22]. Moreover, the website supports continuous benchmarking so that developers can submit their algorithms to receive independent feedback on model performance. Through this iterative process, the goal is to support the development of ≥1 cough-based algorithm that could be integrated into a simple mobile device and provide a point-of-care tuberculosis screening tool that could be deployed in community-based settings. Once developed, the continuous benchmarking mechanism and held-out data could potentially support its review by a regulatory body.

The dataset and challenge had some limitations. The cough sounds collected were restricted to 0.5-second recordings around the peak; the use of whole cough sounds may further improve performance [32]. As all participants were symptomatic, there are limitations in extending these models for community-wide screening, and additional data collection from screening cohorts is needed. The participants also all had cough; while solicited cough sounds may have value for those without cough, this needs to be further evaluated. Variation by country may reflect differences in comorbid conditions and disease presentation, but also may be due to differences in phone model used and environmental noise. However, 0.5-second recordings limited background noise, and algorithms should be developed to be compatible with multiple phone models and environments.

The goal of the challenge was to classify microbiologically confirmed tuberculosis; if these algorithms are used as part of 2-step screening to guide further testing, other outcomes could be considered, such as radiographic evidence of lung disease. Future external validation of these models is needed and should include comparative assessment with other screening tools and the added yield when performed in parallel. The greater probability of tuberculosis classification in individuals with higher bacillary loads may be a useful marker of infectiousness and needs further study. By establishing the platform and approach, additional challenges can be created that update the datasets and goals to support new algorithms.

In conclusion, the CODA TB DREAM Challenge accelerated the development of cough sound models that can be integrated into mobile devices for a simple, point-of-care screening tool for tuberculosis. It also highlighted how open science and collaborative efforts can support rapid, inclusive, and effective health innovations.

Supplementary Data

Supplementary materials are available at *Open Forum Infectious Diseases* online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

Acknowledgments

We thank the individuals who participated in the Rapid Research in Diagnostic Development TB Network (R2D2 TB Network) and Digital Cough Monitoring Project studies and the study personnel.

Consortium coauthors. Coauthors in the Cough Diagnostic Algorithm for Tuberculosis (CODA TB) DREAM Challenge Consortium include the following: Gautam Ahuja, BSc, Shalini Balodi, BSc, Diya Khurdiya, BSc, Rintu Kutum, PhD, Aakash M. Rao, PgD, and Ashwin Salampuria, PgD (Department of Computer Science, Ashoka University, Haryana, India); Sina Akbarian, MAsc, and Sepehr Asgarian, MSc (Klick Applied Sciences, Klick, Toronto, Ontario, Canada); Akanksha Arora, MSc (Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India); Shubham Choudhury, MTech, and Gajendra P. S. Raghava, PhD (Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India); Sherry Dong (AI campus, Cedar Sinai, Los Angeles, California); Yuanfang Guan, PhD, and Yiyang Nan, MSc, Hanrui Zhang, PhD (Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor); Aniket Gupta, Tenglong Li, PhD, and Rohan Singh (Arkansas AI Campus, Arkansas); Jouhyun Jeon, PhD, and Qayam Jetha, MPP (Hyfe, Seattle, Washington); Zhixiang Lu, MSc (Xi'an Jiaotong-Liverpool University, Wisdom Lake Academy of Pharmacy, Suzhou, China); Sumeet Patiyal, PhD (Indraprastha Institute of Information Technology, New Delhi, and Cancer Data Science Laboratory, National Cancer Institute, National Institutes of Health, Bethesda, Maryland); and Chandra Suda (Arkansas AI Campus and Bentonville High School, Bentonville, Arkansas).

Author contributions. Conception: S. K. S., L. O., C. B., P. D., P. M. S., S. G. L., and A. C. Data acquisition: M. R., R. R., I. L., O. L., D. J. C., N. V. N., W. W., C. Y., and G. T. Data analysis: S. K. S., S. H., J. Y. C., S. H. C., T. M. C., C. H. H., K. L. H., F. M., D. R., E. S. C. S., Y. T., H. K. W., C. H. W., S. B., S. K., V. Y., and Consortium members. Data interpretation: D. J., S. K. S., S. H., S. B., S. K., V. Y., S. G. L., and A. C. Drafting the manuscript: D. J., S. K. S., and S. H.. Reviewing the manuscript critically for important intellectual content and final approval of the version to be published: All authors. All authors had access to the data in the study and accept responsibility to submit for publication.

Data availability. The challenge deidentified training data, data dictionary, and links to the code and write-ups for the model submissions are available at www.synapse.org/TBcough. To access the data, register for a free Synapse account and then review and agree to the Synapse and CODA TB DREAM Challenge data use policies. Users can also submit models for evaluation against the validation data in an ongoing manner.

Financial support. This work was supported by the Bill & Melinda Gates Foundation (funding for the CODA TB DREAM Challenge and postchallenge evaluation); the US National Institutes of Health (grant U01 AI152087 to the R2D2 TB Network; grant K23HL153581 to D. J.; and grants D43TW010350, U01AI152087, U54EB027049, and R01AI136894 to G. T.); the Patrick J. McGovern Foundation (funding for the Digital Cough Monitoring study); the Fonds de recherche du Québec – Santé (Junior 1 Salary Award to S. G. L.); and the EDCTP2 program, supported by the European Union (grants RIA2018D-2509 [PreFIT]; RIA2018D-2493 [SeroSelectTB], and RIA2020I-3305 [CAGE-TB] to G. T.)

Potential conflicts of interest. P. M. S. is employed by Hyfe AI. All other authors report no potential conflicts.

References

1. World Health Organization (WHO). Global strategy on digital health 2020–2025. Geneva, Switzerland: World Health Organization, **2021**:1–60.
2. Schwalbe N, Wahl B. Artificial intelligence and the future of global health. Lancet **2020**; 395:1579v.
3. Hosny A, Aerts HJWL. Artificial intelligence for global health: socially responsible technologies promise to help address health care inequalities. Science **2019**; 366: 955–6.
4. World Health Organization. Global tuberculosis report. Geneva, Switzerland: World Health Organization, **2023**.
5. World Health Organization. WHO consolidated guidelines on tuberculosis: module 2: screening: systematic screening for tuberculosis disease. Geneva, Switzerland: World Health Organization. **2021**.
6. Sharma M, Nduba V, Njagi LN, et al. TBscreen: a passive cough classifier for tuberculosis screening with a controlled dataset. Sci Adv **2024**; 10:eadi0282.
7. Pahar M, Klopper M, Reeve B, Warren R, Theron G, Niesler T. Automatic cough classification for tuberculosis screening in a real-world environment. Physiol Meas **2021**; 42:105014.
8. Zimmer AJ, Ugarte-Gil C, Pathri R, et al. Making cough count in tuberculosis care. Commun Med (Lond) **2022**; 2:83.
9. Muguli A, Pinto L, Nirmala R, et al. DiCOVA challenge: dataset, task, and baseline system for COVID-19 diagnosis using acoustics. Proc Annual Conf Int Speech Commun Assoc INTERSPEECH **2021**; 6:4241–5.
10. Ellrott K, Buchanan A, Creason A, et al. Reproducible biomedical benchmarking in the cloud: lessons from crowd-sourced data challenges. Genome Biol **2019**; 20:195.
11. Huddart S, Yadav V, Sieberts SK, et al. A dataset of solicited cough sound for tuberculosis triage testing. Sci Data **2024**; 11:1149.
12. Crowder R, Thangakunam B, Andama A, et al. Diagnostic accuracy of TB screening tests in a prospective multinational cohort: chest-X-ray with computer-aided detection, Xpert TB host response, and C-reactive protein. Clin Infect Dis doi:10.1093/cid/ciae549. Published 7 November 2024.
13. Raberahona M, Zimmer A, Rakotoarivelo RA, et al. Continuous digital cough monitoring during 6-month pulmonary tuberculosis treatment. ERJ Open Res **2025**; 11:00655-2024.
14. The Hyfe Team. Smart cough monitoring: an innovation milestone for global respiratory health. Hyfe Res Ser **2021**; 1:1–6
15. Huddart S, Yadav V, Sieberts SK, et al. Solicited cough sound analysis for tuberculosis triage testing: the CODA TB DREAM Challenge dataset. Sci Data **2024**. 28 March 2024. Available from: https://doi.org/10.1101/2024.03.27.24304980
16. Ma H, Bandos AI, Rockette HE, Gur D. On use of partial area under the ROC curve for evaluation of diagnostic performance. Stat Med **2013**; 32:3449–58.
17. Chaibub Neto E, Yadav V, Sieberts SK, Omberg L. A novel estimator for the two-way partial AUC. BMC Med Inform Decis Mak **2024**; 24:57.
18. World Health Organization. High-priority target product profiles for new tuberculosis diagnostics: report of a consensus meeting. **2014**. Available at: https://www.who.int/publications/i/item/WHO-HTM-TB-2014.18. Accessed 23 September 2025.
19. Breiman L. Random forests. Mach Learn **2001**; 45:5–32.
20. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics **2011**; 12:77.
21. Chaibub Neto E. pROC_based_tpAUC. GitHub. **2022**. Available at: https://github.com/echaibub/pROC_based_tpAUC. Accessed 23 September 2025.
22. Sage Bionetworks. CODA TB DREAM Challenge. Available at: https://www.synapse.org/TBcough. Accessed 1 April 2024.
23. Chang Y, Jing X, Ren Z, Schuller BW. CovNet: a transfer learning framework for automatic COVID-19 detection from crowd-sourced cough sounds. Front Digit Health **2022**; 3:799067.
24. Pahar M, Klopper M, Warren R, Niesler T. COVID-19 cough classification using machine learning and global smartphone recordings. Comput Biol Med **2021**; 135:104572.
25. Ghrabli S, Elgendi M, Menon C. Identifying unique spectral fingerprints in cough sounds for diagnosing respiratory ailments. Sci Rep **2024**; 14:593.
26. Pentakota P, Rudraraju G, Sripada NR, et al. Screening COVID-19 by Swaasa AI platform using cough sounds: a cross-sectional study. Sci Rep **2023**; 13:18284.
27. Bhattacharya D, Sharma NK, Dutta D, et al. Coswara: a respiratory sounds and symptoms dataset for remote screening of SARS-CoV-2 infection. Sci Data **2023**; 10:397.
28. Orlandic L, Teijeiro T, Atienza D. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. Sci Data **2021**; 8:156.
29. Swaminathan S, Padmapriyadarsini C, Narendran G. HIV-associated tuberculosis: clinical update. Clin Infect Dis **2010**; 50:1377–86.
30. Tavaziva G, Harris M, Abidi SK, et al. Chest X-ray analysis with deep learning-based software as a triage test for pulmonary tuberculosis: an individual patient data meta-analysis of diagnostic accuracy. Clin Infect Dis **2022**; 74:1390–400.
31. Geric C, Qin ZZ, Denkinger CM, et al. The rise of artificial intelligence reading of chest X-rays for enhanced TB diagnosis and elimination. Int J Tuberc Lung Dis **2023**; 27:367–72. https://www.synapse.org/TBcough
32. Yellapu GD, Rudraraju G, Sripada NR, et al. Development and clinical validation of Swaasa AI platform for screening and prioritization of pulmonary TB. Sci Rep **2023**; 13:4740.