

1 Accelerating cough-based algorithms for pulmonary tuberculosis screening:
2 Results from the CODA TB DREAM Challenge

3
4 Devan Jaganath, MD^{1,2*}, Solveig K Sieberts, PhD^{3*}, Mihaja Raberahona, MD^{4,5*}, Sophie
5 Huddart, PhD^{2,6}, Larsson Omberg, PhD³, Rivo Rakotoarivelo, MD^{7,8}, Issa Lyimo, BVM⁹, Omar
6 Lweno, MD⁹, Devasahayam J. Christopher, MBBS¹⁰, Nguyen Viet Nhung, PhD^{11,12}, William
7 Worodria, MMed¹³, Charles Yu, MD¹⁴, Jhih-Yu Chen, MS¹⁵, Sz-Hau Chen, PhD^{16,17}, Tsai-Min
8 Chen, PhD^{18,19}, Chih-Han Huang, MS²⁰, Kuei-Lin Huang, MD²¹, Filip Mulier, PhD²², Daniel
9 Rafter, MD²², Edward S.C. Shih, PhD²³, Yu Tsao, PhD^{18,24}, Hsuan-Kai Wang, PhD²⁵, Chih-Hsun
10 Wu, PhD²⁶, Christine Bachman, MPH²⁷, Stephen Burkot, MS²⁷, Puneet Dewan, MD²⁷, Sourabh
11 Kulhare, MS²⁷, Peter M. Small, MD^{28,29}, Vijay Yadav, MS³, Simon Grandjean Lapierre, MD^{30,31+},
12 Grant Theron, PhD³²⁺, Adithya Cattamanchi, MD^{2,33+} for the CODA TB DREAM Challenge
13 Consortium

14
15 ¹ Division of Pediatric Infectious Diseases, University of California, San Francisco, San
16 Francisco, USA

17 ² Center for Tuberculosis, University of California, San Francisco, USA

18 ³Sage Bionetworks, Seattle, USA

19 ⁴CHU Joseph Rasera Befelatanana, Antananarivo, 101, Analamanga, Madagascar

20 ⁵Centre d'Infectiologie Charles Mérieux, Université d'Antananarivo, Antananarivo, 101,
21 Analamanga, Madagascar

22 ⁶Department of Epidemiology and Biostatistics, University of California, San Francisco, San
23 Francisco, CA USA

24 ⁷CHU Tambohobe Fianarantsoa, 301, Haute-Matsiatra, Madagascar

25 ⁸Université de Fianarantsoa, Fianarantsoa, 301, Haute-Matsiatra, Madagascar

26 ⁹Ifakara Health Institute, Environmental and Ecological Sciences & Interventions and Clinical
27 Trials Departments, Kiko Avenue, Plot 463, Mikocheni, Dar es Salaam, Tanzania

28 ¹⁰ Christian Medical College, Vellore (Ranipet campus), Tamil Nadu, India

- 29 ¹¹National Tuberculosis Programme, 463 Hoang Hoa Tham, Ba Dinh District, Hanoi, Vietnam
- 30 ¹²VNU, University of Medicine and Pharmacy
- 31 ¹³Walimu, Plot 341 White Close Najjera – Wakiso district, Kampala, Uganda
- 32 ¹⁴De La Salle Medical and Health Sciences Institute, Governor D. Mangubat Avenue,
- 33 Dasmariñas Cavite, Philippines 4114
- 34 ¹⁵ Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University,
- 35 Taipei, Taiwan
- 36 ¹⁶ Industrial Information Department, Development Center for Biotechnology, Taipei, Taiwan
- 37 ¹⁷ Investment & Wealth Management, FCC Partners Inc., Taipei, Taiwan
- 38 ¹⁸ Graduate Program of Data Science, National Taiwan University and Academia Sinica, Taipei,
- 39 Taiwan
- 40 ¹⁹ Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
- 41 ²⁰ Department of Data Science, ANIWARE, Taipei, Taiwan
- 42 ²¹ School of Medicine, China Medical University, Taichung, Taiwan
- 43 ²² Flywheel.io, Minneapolis, Minnesota USA
- 44 ²³ Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan
- 45 ²⁴ Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
- 46 ²⁵ Independent Researcher, Taipei, Taiwan
- 47 ²⁶ Artificial Intelligence and E-learning Center, National Chengchi University, Taipei, Taiwan
- 48 ²⁷ Global Health Labs, 14360 SE Eastgate Way, Bellevue, USA
- 49 ²⁸ Department of Global Health, University of Washington, Seattle USA
- 50 ²⁹ Hyfe, Inc, Seattle USA
- 51 ³⁰ Centre de Recherche du Centre Hospitalier de l'Université de Montréal, Immunopathology
- 52 Axis, 900 St-Denis, Montréal, Québec, H2X 0A9 Canada
- 53 ³¹ Université de Montréal, Department of Microbiology, Infectious Diseases and Immunology,
- 54 2900 Edouard-Montpetit, Montréal, Québec, H3T 1J4 Canada
- 55 ³² DSI-NRF Centre of Excellence for Biomedical Tuberculosis Research, South African Medical
- 56 Research Council Centre for Tuberculosis Research, Division of Molecular Biology and Human
- 57 Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, South Africa
- 58 ³³ University of California Irvine, School of Medicine, Orange, USA

59

60 *First authors contributed equally

61 †Senior authors contributed equally

62

63 Corresponding author
64 Adithya Cattamanchi
65 1001 Health Sciences Road
66 Irvine CA 92697-3950
67 (714) 456-2959
68 acattama@hs.uci.edu
69 Word Count: 2,993

70 Key Points

71 **Question:** Can an open-access data challenge support the rapid development of cough-based
72 artificial intelligence (AI) algorithms to screen for tuberculosis (TB)?

73

74 **Findings:** In this diagnostic study, teams were provided well-characterized cough sound data
75 from seven countries, and developed and submitted AI models for independent validation.
76 Multiple models that combined clinical and cough data achieved the target accuracy of at least
77 80% sensitivity and 60% specificity to classify microbiologically-confirmed TB.

78

79 **Meaning:** Cough-based AI models have promise to support point-of-care TB screening, and
80 open-access data challenges can accelerate the development of AI-based tools for global
81 health.

82 Abstract

83 **Importance.** Open-access data challenges have the potential to accelerate innovation in
84 artificial-intelligence (AI)-based tools for global health. A specimen-free rapid triage method for
85 TB is a global health priority.

86

87 **Objective.** To develop and validate cough sound-based AI algorithms for tuberculosis (TB)
88 through the Cough Diagnostic Algorithm for Tuberculosis (CODA TB) DREAM challenge.

89

90 **Design.** In this diagnostic study, participating teams were provided cough-sound and clinical
91 and demographic data. They were asked to develop AI models over a four-month period, and
92 then submit the algorithms for independent validation.

93

94 **Setting.** Data was collected using smartphones from outpatient clinics in India, Madagascar, the
95 Philippines, South Africa, Tanzania, Uganda, and Vietnam.

96

97 **Participants.** We included data from 2,143 adults who were consecutively enrolled with at least
98 two weeks of cough. Data were randomly split evenly into training and test partitions.

99

100 **Exposures.** Standard TB evaluation was completed, including Xpert MTB/RIF Ultra and
101 culture. At least three solicited coughs were recorded using the Hyfe Research app.

102

103 **Main Outcomes and Measures.** We invited teams to develop models using 1) cough sound
104 features only and/or 2) cough sound features with routinely available clinical data to classify
105 microbiologically confirmed TB disease. Models were ranked by area under the receiver

106 operating characteristic curve (AUROC) and partial AUROC (pAUROC) to achieve at least 80%
107 sensitivity and 60% specificity.

108

109 **Results.** Eleven cough models were submitted, as well as six cough-plus-clinical models.
110 AUROCs for cough models ranged from 0.69-0.74, and the highest performing model achieved
111 55.5% specificity (95% CI 47.7-64.2) at 80% sensitivity. The addition of clinical data improved
112 AUROCs (range 0.78-0.83), five of the six submitted models reached the target pAUROC, and
113 highest performing model had 73.8% (95% CI 60.8-80.0) specificity at 80% sensitivity. In post-
114 challenge subgroup analyses, AUROCs varied by country, and was higher among males and
115 HIV-negative individuals. The probability of TB classification correlated with Xpert Ultra semi-
116 quantitative levels.

117

118 **Conclusions and Relevance.** In a short period, new and independently validated cough-based
119 TB algorithms were developed through an open-source and transparent process. Open-access
120 data challenges can rapidly advance and improve AI-based tools for global health.

121 Introduction

122 As global health challenges intersect with rapid advancements in technology and artificial
123 intelligence (AI), digital health tools have the potential to enhance disease surveillance,
124 diagnosis, and management.¹⁻³ In particular, the widespread availability of smartphones and
125 wearable sensors create opportunities for low-cost, non-invasive applications to increase
126 healthcare access and quality.⁴ However, development and deployment of AI solutions have
127 primarily focused on commercial applications in high-income markets. A major challenge to
128 equitable implementation of these tools is a lack of available datasets from diverse geographic
129 settings, and limited focus on conditions that disproportionately affect low- and middle-income
130 countries (LMICs).^{2,5} Moreover, available datasets may be proprietary, preventing open-access
131 sharing and transparent algorithm development. The consequence is a dearth of AI tools
132 validated in LMICs and that address the public health challenges they face.

133

134 Tuberculosis (TB) is the leading cause of death from an infectious disease worldwide.⁶ The high
135 mortality is driven by a large case detection gap, in which 3.1 million of the estimated 10 million
136 individuals who develop TB disease each year have not been diagnosed or reported to public
137 health programs.⁶ AI has already supported TB diagnosis through automated chest X-ray
138 reading, and computer-assisted detection algorithms (CAD) have been endorsed by the World
139 Health Organization (WHO) as a triage tool.⁷ However, CAD systems require infrastructure and
140 expertise to obtain chest X-rays which limit their impact at primary health facility levels. Cough is
141 a common symptom of TB, and initial studies suggest that there are unique acoustic features
142 that can distinguish pulmonary TB from other respiratory conditions.^{8,9} Furthermore, cough
143 detection applications have already been developed for mobile phones and smart watches,¹⁰
144 providing an opportunity to integrate cough-based AI algorithms for point-of-care TB
145 assessment by providers and patients.

146

147 In other diseases, including COVID-19,¹¹ open-access, crowd-sourced data challenge initiatives
148 have been used to accelerate the development of novel algorithms.¹² These initiatives provide a
149 transparent platform to share methods and findings, and support independent validation. To
150 expedite AI diagnostic development for TB, we established a cough sound repository from
151 individuals prospectively enrolled with presumptive TB across seven high TB-burden countries,
152 and launched the Cough Diagnostic Algorithm for TB (CODA TB) Dream Challenge.¹³ We
153 present the results of the challenge and highlight the role of this approach to rapidly advance AI
154 tools for global diseases that impact LMICs.

155

156 Methods

157 CODA TB DREAM Challenge

158 The CODA TB DREAM Challenge launched on October 26, 2022. Participants were asked to
159 develop a model to classify TB disease in two sub-challenges: (1) using cough sounds alone
160 and (2) using cough sounds and basic demographic and clinical variables. Challenge teams
161 were allowed to submit results for one or both sub-challenges. To be considered in the official
162 ranking, teams needed to submit a final report that outlined their methods and conclusions, and
163 a link to their source code. The timeline of the challenge is shown in **Supplemental Figure 1**.

164

165 The challenge was hosted by Sage Bionetworks, which has developed an open-science,
166 collaborative competition framework for evaluating and comparing computational algorithms,
167 using the DREAM Challenges framework. DREAM focuses exclusively on biomedicine with an
168 explicit mandate for transparency, openness, and collaboration. The challenge was set up on
169 Synapse (www.synapse.org/tbcough), which provided all instructions, a secure platform for data

170 sharing, a forum for communication with challenge participants, and supported submission of
171 models for independent validation.

172
173 Any individual or team could participate in the challenge. After registering for a free Synapse
174 account, they certified that they understood the Synapse data use policy, verified their identify,
175 and agreed to the challenge guidelines to not attempt to identify or contact any study
176 participants, to not share the data with others, and that they must comply with the intended use
177 of the data. If they agreed to these conditions, they were given access to the de-identified
178 training data as described below. The challenge was advertised as broadly as possible,
179 including on social media, to multiple academic institution listservs and departments of global
180 health, bioinformatics and computer science, companies interested in cough-based or TB
181 diagnosis, and previous DREAM Challenge participants.

182

183 Study dataset

184 Data for the CODA TB DREAM Challenge were obtained from two multi-country TB diagnostic
185 evaluation studies.¹³ The Rapid Research in Diagnostic Development TB Network (R2D2 TB
186 Network) enrolled participants at outpatient health centers in Uganda, South Africa, Vietnam,
187 the Philippines and India. The Digital Cough Monitoring Project enrolled participants in Tanzania
188 and Madagascar. Ethical approvals for the studies were obtained from institutional review
189 boards (IRB) in the US (R2D2 TB Network, University of California, San Francisco) and Canada
190 (Digital Cough Monitoring Project, University of Montreal), as well as IRBs in each country in
191 which participants were enrolled. All participants provided written informed consent for study
192 participation, cough recording and anonymized data sharing.

193

194 In both studies, eligible participants were 18 years or older and had a new or worsening cough
195 for at least two weeks. Participants completed a standard evaluation for TB including a clinical

196 questionnaire and examination, sputum-based molecular testing (Xpert MTB/RIF Ultra,
197 Cepheid, Sunnyvale) and liquid or solid medium culture testing. Participants were asked to
198 produce at least three solicited cough sounds during the baseline visit prior to any TB treatment
199 initiation. The coughs were collected on an Android-based smartphone using the Hyfe Research
200 app,¹⁴ which uses a convolutional neural network (CNN) model to automatically detect the
201 cough and saves the 0.5 second peak sound.¹³ Solicited cough sounds were collected, though
202 any triggered passive coughs were also recorded. TB disease status was based on a
203 microbiological reference standard, defined by a positive molecular or culture result. Further
204 details on the study procedures and dataset including a summary of participant demographics
205 and country distribution have been published previously.¹³

206

207 A training set (n=1,105) for algorithm development was created by taking a 50% sample of the
208 dataset randomized at the individual level. Of the remaining data, 24% (n=248) was randomly
209 selected at the individual level for the “leaderboard” test set, from which challenge participants
210 could receive periodic feedback on their model performance, and the remainder (n=790) was
211 reserved as the final test set for algorithm evaluation. Challenge teams were given direct access
212 only to the training set, which included the raw peak cough sound recordings in WAV format, as
213 well as associated age, sex, height, weight, smoking status, self-reported duration of cough,
214 history of prior TB, common TB symptoms (hemoptysis, fever, night sweats, weight loss), heart
215 rate and temperature. These variables were chosen as data that would be readily available in
216 routine primary care settings. We did not include HIV status as the testing and/or results may
217 not be available or known at the time of cough assessment.

218

219 Algorithm development and evaluation

220 Participating teams could train an algorithm using any pre-processing approach and model, and
221 with any programming language (i.e. R, Python, etc.) or framework (such as Keras or Pytorch).

222 For evaluation, models were required to be saved in Open Neural Network Exchange (ONNX)
223 format and submitted in a Docker container, with any code needed for pre-processing the data.
224
225 Challenge teams had five interim opportunities to evaluate their algorithms on the “leaderboard”
226 test set before the final algorithms were due for test set evaluation (**Supplemental Figure 1**).
227 The teams submitted their preliminary models and we independently applied those models to
228 the leaderboard test set. The output of each model was continuous TB prediction scores used to
229 generate calculate area under the receiver operating characteristic curve (AUROC) and two-
230 way partial AUROC¹⁵ (pAUROC) with 80% sensitivity and 60% specificity. We set this threshold
231 to identify promising algorithms that had an accuracy that was at least within 10% of the
232 minimum WHO target product profile (TPP) accuracy for a TB triage test ($\geq 90\%$ sensitivity,
233 $\geq 70\%$ specificity).¹⁶ A higher pAUROC indicates that a greater area meets the minimum target
234 sensitivity and specificity. Original model evaluation was performed in Python (version 3.8.8). All
235 subsequent analyses were performed in R Software version 4.2.2 (2022-10-31). Evaluations of
236 model statistics were done using the pROC R package. Implementation of the pAUROC was
237 provided by Chaibub Neto, et al.^{17,18}
238
239 The final submission deadline was on February 13, 2023, four months from the launch of the
240 challenge. Similar to the leaderboard rounds, challenge teams submitted their pre-processing
241 code (if applicable) and trained models, and we independently applied the model to the test set.
242 Final model performance was evaluated by the pAUROC. If no model could meet the accuracy
243 threshold, the algorithms were evaluated by the total AUROC. Variability in the AUROCs and
244 pAUROCs were assessed via bootstrap resampling (n = 1,000).
245
246 Clinical Data Only Model

247 As a sensitivity analysis to assess the degree that the clinical variables alone contributed to the
248 models in sub-challenge 2, we developed a Random Forest model¹⁹ using the clinical and
249 demographic variables provided to challenge participants. In addition to the variables provided,
250 body mass index (BMI) was computed from height and weight variables, and duration of cough
251 symptoms was log-transformed prior to model fitting. The model was trained using 1,000 trees.

252

253 Subgroup analyses

254 After the challenge was complete, first- and second-ranked teams (n=5 due to ties) in both sub-
255 challenges were invited to participate in additional model evaluation. We assessed the accuracy
256 of the models by country, sex, and HIV status. We also compared the probability of TB
257 classification for each model by Xpert MTB/RIF Ultra semi-quantitative PCR result.

258

259 Results

260 *Challenge Implementation*

261 147 individuals and 18 teams registered for the CODA TB DREAM Challenge. In each
262 leaderboard round, two to eight teams submitted models. Thirteen teams submitted final models
263 for sub-challenge 1, and eight for sub-challenge 2. Of those that submitted final models, eleven
264 (sub-challenge 1) and six (sub-challenge 2) teams submitted a summary of methods and model
265 code. The winning models for each sub-challenge are described in the Supplemental Methods.
266 The reports for the full set of submissions are available through the challenge website.²⁰

267

268 *Sub-challenge 1*

269 As shown in **Table 1**, **Figure 1A** and **Supplemental Figure 2**, AUROCs ranged from 0.689 to
270 0.743. The top model achieved a specificity of 55.5% at 80% sensitivity; as further shown in
271 **Supplemental Table 1**, it did not achieve the WHO TPP-based accuracy thresholds for a triage

272 test at 90% sensitivity and 70% specificity. Of the 11 groups, 4 (36%) used CNNs, 4 (36%) used
273 artificial neural networks, and 3 (27%) used gradient boosting decision tree methods.

274

275 *Sub-challenge 2*

276 All groups used the same algorithm approach they utilized in sub-challenge 1. As shown in
277 **Figure 1B, Supplemental Figure 3** and **Table 2**, overall performance improved compared to
278 the use of cough sounds alone, and the top performing model achieved an AUROC of 0.832
279 (95% CI 0.795-0.863) and a pAUROC of 0.003 (95% CI 6.1e-06-0.012). Five of the six (83%)
280 submission achieved at least 80% sensitivity and 60% specificity, with the top model reaching
281 73.8% (95% CI 60.8-80.0) at 80% sensitivity. For the WHO TPP for a TB triage test, the top
282 performing model achieved 54% specificity (95% CI = (38%, 63%)) at 90% sensitivity
283 (**Supplemental Table 1**). In sensitivity analysis, the clinical data only model achieved an
284 AUROC of 0.817 (95% CI 0.778-0.850) and pAUROC of 0.004 (95% CI 5.5e-4-0.010). This was
285 higher than the cough only model, but the top combined cough and clinical data model
286 outperformed both.

287

288 *Subgroup Assessment*

289 Model performance for the combined cough sound and clinical data models (sub-challenge 2)
290 was variable across country of data collection (**Figure 2A**). In general, the models performed
291 better on data from the Philippines, Uganda, Tanzania and Vietnam. The median AUROC of the
292 cough and clinical data models was slightly higher for males compared to females (median 0.82
293 vs. 0.78, $p < 0.01$, **Figure 2B**). Model performance was also slightly higher among people not
294 living with HIV compared to people living with HIV (median AUROC 0.83 vs. 0.78, $p < 0.01$,
295 **Figure 2C**). Subgroup results for sub-challenge 1 (cough sounds only) are shown in
296 **Supplemental Figures 4-7**. Findings were similar, although we found slightly lower accuracy in
297 males vs. females (median AUROC 0.69 vs 0.71, $p = 0.02$) in contrast to sub-challenge 2.

298

299 For all submitted cough and clinical data models, the median predicted probability of being TB-
300 positive increased with Xpert MTB/RIF Ultra semi-quantitative level, from trace positive results
301 to high bacillary load results (**Figure 3, Supplemental Figure 7**).

302

303 Discussion

304 The CODA TB Dream Challenge addressed a critical need to accelerate the development of AI-
305 based tools for global health through an inclusive, open and transparent approach. The
306 challenge brought together students, researchers, and industry partners from a diverse
307 geographical spectrum with a common goal of developing novel TB diagnostic algorithms using
308 cough sounds. In a short period, challenge participants created, tested and improved algorithms
309 using cough sounds and routine clinical and demographic data that approached the WHO TPP
310 accuracy targets for a TB triage test. Open-access research and citizen science represent a
311 potential paradigm shift in how digital health solutions can be developed for global health by
312 harnessing the collective expertise of an international community to address a common
313 scientific and humanitarian goal.

314 The cough-sound only models had similar accuracies with AUROCs that ranged from
315 0.65-0.74. This performance is within the wide range of cough-based models that have been
316 developed to detect COVID-19 (AUROCs 0.62 to 0.98).²¹⁻²⁴ In India, for example, a cough-
317 based CNN model achieved an AUROC of 0.75 to detect COVID-19.²⁴ There are a few
318 published cough-sound models in TB that have shown higher performance (AUROC 0.79-
319 0.94),^{8,9} but these have been small studies that were not validated on independent datasets,
320 and may overestimate accuracy. A major limitation of previous cough models for other
321 conditions was the use of crowd-sourced data.^{11,25,26} While this approach rapidly generates
322 large real-world datasets, there are multiple challenges, including selection bias, subjective

323 clinical assessment and heterogenous reference standard definitions. In CODA, we utilized a
324 multi-country cohort of consecutively enrolled symptomatic individuals with indications for TB
325 evaluation, standardized clinical data and cough collection protocols, objective TB testing and
326 uniform case definitions. This increases the confidence that algorithms are identifying features
327 specific to the disease condition, reduces AI-related biases, and better reflects how the
328 algorithms will perform in the intended settings and populations.

329 Performance improved when routine demographic and clinical variables were added to
330 models, and five of six algorithms approached the WHO-established target accuracy thresholds
331 for a TB triage test. We chose demographic and clinical variables that are associated with TB
332 and could be collected in primary care settings or self-reported on a mobile application. As a
333 post-challenge sensitivity analysis, we developed a clinical data only model that performed well
334 (AUROC 0.817), but the addition of cough sound data improved accuracy and supports the role
335 of integrating both data types. The best performing challenge models utilized deep learning
336 algorithms; while interpretability can be limited with such models, subgroup findings increase
337 confidence in a TB-specific signal. First, the probability of TB classification correlated with
338 bacterial burden as measured by semi-quantitative PCR results in both sub-challenges, which
339 was also seen in a recent study in Kenya.⁸ Moreover, worse performance among people living
340 with HIV who often have paucibacillary disease may be expected.²⁷ These differences in
341 accuracy have also been seen in CAD algorithms for chest x-ray interpretation,²⁸ and different
342 thresholds may be needed depending on the setting or target group.²⁹

343 It is important to recognize that the final submitted models were developed rapidly over a
344 short timeframe, and there is potential for further optimization. This includes exploring more
345 complex CNN architectures and/or ensembles, increasing the size of the training set and
346 developing country-specific models. At the same time, the overarching goal of the challenge
347 was to accelerate innovation and gain key insights into cough-based AI models for TB. In four
348 months, the challenge 1) supported multiple new and independently validated cough sound

349 algorithms that could discriminate TB disease; 2) demonstrated that clinical data could augment
350 performance; and 3) transparently shared the best performing algorithms and processing
351 methods.

352 To further facilitate ongoing model development, the dataset remains open-source and
353 can be downloaded at the challenge website.³⁰ Moreover, the website supports continuous
354 benchmarking so that developers can submit their algorithms to receive independent feedback
355 on model performance. Through this iterative process, the goal is to support the development of
356 at least one cough-based algorithm that could be integrated into a simple mobile device and
357 provide a point-of-care TB triage tool which could be deployed in community-based settings.
358 Once developed, the continuous benchmarking mechanism and held-out data could potentially
359 support its review by a regulatory body.

360 The dataset and challenge had some limitations. The cough sounds collected were
361 restricted to 0.5 second recordings around the peak; the use of whole cough sounds may further
362 improve performance.³¹ As all participants were symptomatic, there are limitations in extending
363 these models for community-wide screening, and additional data collection from screening
364 cohorts is needed. The participants also all had cough; while solicited cough sounds may have
365 value for those without cough, this needs to be further evaluated. Variation by country may
366 reflect differences in co-morbidities and disease presentation, but also may be due to
367 differences in phone model used and environmental noise. However, 0.5 second recordings
368 limited background noise and algorithms should be developed to be compatible with multiple
369 phone models and environments. The goal of the challenge was to classify microbiologically-
370 confirmed TB; if these algorithms are used as part of two-step screening to guide further testing,
371 other outcomes could be considered such as a radiographic evidence of lung disease. The
372 greater probability of TB classification in individuals with higher bacillary loads may be a useful
373 marker of infectiousness and needs further study. By establishing the platform and approach,
374 additional challenges can be created that update datasets and goals to support new algorithms.

375 In conclusion, the CODA TB Dream Challenge accelerated the development of cough-
376 sound models that can be integrated into mobile devices for a simple, point-of-care triage tool
377 for TB. It also highlighted how open science and collaborative efforts can support rapid,
378 inclusive, and impactful health innovations. Through such initiatives, we move closer to realizing
379 the expansive potential of digital tools for TB and global health.

380

381 Data Sharing

382 The challenge training data and links to the code and write-ups for the model submissions are
383 available at www.synapse.org/TBcough. Additionally, users can register to submit models for
384 evaluation against the validation data in an ongoing manner.

385

386 Acknowledgements

387 The CODA TB DREAM Challenge and post-challenge evaluation was funded in part by the Bill
388 & Melinda Gates Foundation. R2D2 was funded by the U.S. National Institutes of Health (U01
389 AI152087), and the Digital Cough Monitoring study was funded by the Patrick J. McGovern
390 Foundation. SGL is supported by a Junior 1 Salary Award from the Fonds de Recherche Santé
391 Québec. DJ is supported by funding by the National Institutes of Health. GT acknowledges
392 funding from the EDCTP2 programme supported by the European Union (RIA2018D-2509,
393 PreFIT; RIA2018D-2493, SeroSelectTB; RIA2020I-3305, CAGE-TB) and the National Institutes
394 of Health (D43TW010350; U01AI152087; U54EB027049; R01AI136894).

395

396

397 The CODA TB DREAM Challenge Consortium

398 Gautam Ahuja, BSc³⁴, Sina Akbarian, MAsc³⁵, Akanksha Arora, MSc³⁶, Sepehr Asgarian, MSc³⁵,
399 Christine Bachman, MPH²⁷, Shalini Balodi, BSc³⁴, Stephen Burkot, MS²⁷, Adithya Cattamanchi,
400 MD^{2,33}, Jhih-Yu Chen, MS¹⁵, Sz-Hau Chen, PhD^{16,17}, Tsai-Min Chen, PhD^{18,19}, Shubham
401 Choudhury, MTech³⁶, Devasahayam J. Christopher, MBBS¹⁰, Puneet Dewan, MD²⁷, Sherry
402 Dong³⁷, Yuanfang Guan, PhD³⁸, Aniket Gupta³⁹, Chih-Han Huang, MS²⁰, Kuei-Lin Huang, MD²¹,
403 Sophie Huddart, PhD^{2,6}, Devan Jaganath, MD^{1,2}, Jouhyun Jeon, PhD²⁹, Qayam Jetha, MPP²⁹,
404 Diya Khurdiya, BSc³⁴, Sourabh Kulhare, MS²⁷, Rintu Kutum, PhD³⁴, Simon Grandjean Lapierre,
405 MD,^{24,25} Tenglong Li, PhD⁴⁰, Zhixiang Lu, MSc,⁴⁰ Omar Lweno, MD⁹, Issa Lyimo, BVM⁹, Filip
406 Mulier, PhD²², Yiyang Nan, MSc³⁸, Nguyen Viet Nhung, PhD^{11,12}, Larsson Omberg, PhD³,
407 Sumeet Patiyal, PhD^{36,41}, Mihaja Raberahona, MD^{4,5}, Gajendra P.S. Raghava, PhD³⁶, Dan
408 Rafter, MD²², Rivo Rakotoarivelo, MD^{7,8}, Aakash M. Rao, PgD³⁴, Ashwin Salampuria, PgD³⁴,
409 Edward S.C. Shih, PhD²³, Solveig K Sieberts, PhD,³ Rohan Singh³⁹, Peter M. Small, MD^{28,29},
410 Chandra Suda^{39,42}, Grant Theron, PhD³², Yu Tsao, PhD^{18,24}, Hsuan-Kai Wang, PhD²⁵, William
411 Worodria, MMed¹³, Chih-Hsun Wu, PhD²⁶, Vijay Yadav, MS³, Charles Yu, MD¹⁴, Hanrui Zhang,
412 PhD³⁸

413

414 ³⁴ Department of Computer Science, Ashoka University, Haryana, India

415 ³⁵ Klick Applied Sciences, Klick Inc., 175 Bloor Street East, Toronto, Canada

416 ³⁶ Department of Computational Biology, Indraprastha Institute of Information Technology, New
417 Delhi, India

418 ³⁷ AI campus, Cedar Sinai, Los Angeles, CA, USA

419 ³⁸ Department of Computational Medicine & Bioinformatics, University of Michigan - Ann Arbor.

420 Ann Arbor, MI, USA

421 ³⁹ Arkansas AI Campus, AR, USA

422 ⁴⁰ Xi'an Jiaotong-Liverpool University, Wisdom Lake Academy of Pharmacy, Suzhou, China

423 ⁴¹ Cancer Data Science Laboratory, National Cancer Institute, National Institutes of Health,
424 Bethesda, Maryland, USA.

425 ⁴² Bentonville High School, Bentonville, Arkansas, USA

426

427 Author Contributions

428 Author contributions include conception (SKS, LO, CB, PD, PMS, SGL, AC), data acquisition
429 (MR, RR, IL, OL, DJC, NVN, WW, CY, GT), data analysis (SKS, SH, JYC, SHC, TMC, CHH,
430 KLH, FM, DR, ESCS, YT, HKW, CHW, SB, SK, VY, and consortium), data interpretation (DJ,
431 SKS, SH, SB, SK, VY, SGL, AC), drafting the manuscript (DJ, SKS, SH), and reviewing the
432 manuscript critically for important intellectual content and final approval of the version to be
433 published (all co-authors).

434

435 We agree to be accountable for all aspects of the work in ensuring that questions related to the
436 accuracy or integrity of any part of the work are appropriately investigated and resolved (DJ,
437 SKS, MR, SGL, GT, AC).

438

439 Competing Interests

440 PMS is employed by Hyfe AI. The other authors declare no conflicts of interest.

441 References

- 442 1. World Health Organization (WHO). Global strategy on digital health 2020-2025. *Who*.
443 Published online 2021:1-60.
- 444 2. Schwalbe N, Wahl B. Artificial intelligence and the future of global health. *Lancet*.
445 2020;395(10236):1579. doi:10.1016/S0140-6736(20)30226-9
- 446 3. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*.
447 2022;28(1):31-38. doi:10.1038/s41591-021-01614-0
- 448 4. Hosny A, Aerts HJWL. Artificial intelligence for global health: Socially responsible
449 technologies promise to help address health care inequalities. *Science*.
450 2019;366(6468):955. doi:10.1126/SCIENCE.AAY5189
- 451 5. USAID. *Artificial Intelligence in Global Health: Defining a Collective Path Forward.*; 2019.
- 452 6. World Health Organization. *Global Tuberculosis Report.*; 2023. <https://iris.who.int/>.
- 453 7. World Health Organization. *WHO Consolidated Guidelines on Tuberculosis: Module 2:
454 Screening: Systematic Screening for Tuberculosis Disease.*; 2021.
- 455 8. Sharma M, Nduba V, Njagi LN, et al. TBscreen: A passive cough classifier for
456 tuberculosis screening with a controlled dataset. *Sci Adv*. 2024;10(1).
457 doi:10.1126/SCIADV.ADI0282
- 458 9. Pahar M, Klopper M, Reeve B, Warren R, Theron G, Niesler T. Automatic cough
459 classification for tuberculosis screening in a real-world environment. *Physiol Meas*.
460 2021;42(10). doi:10.1088/1361-6579/AC2FB8
- 461 10. Zimmer AJ, Ugarte-Gil C, Pathri R, et al. Making cough count in tuberculosis care.
462 *Communications medicine*. 2022;2(1). doi:10.1038/S43856-022-00149-W
- 463 11. Muguli A, Pinto L, Nirmala R, et al. DiCOVA Challenge: Dataset, task, and baseline
464 system for COVID-19 diagnosis using acoustics. *Proceedings of the Annual Conference*

- 465 of the International Speech Communication Association, *INTERSPEECH*. 2021;6:4241-
466 4245. doi:10.21437/Interspeech.2021-74
- 467 12. Ellrott K, Buchanan A, Creason A, et al. Reproducible biomedical benchmarking in the
468 cloud: Lessons from crowd-sourced data challenges. *Genome Biol.* 2019;20(1):1-9.
469 doi:10.1186/S13059-019-1794-0/FIGURES/3
- 470 13. Sophie Huddart, Vijay Yadav, Solveig K. Sieberts, et al. Solicited Cough Sound Analysis
471 for Tuberculosis Triage Testing: The CODA TB DREAM Challenge Dataset [Preprint].
472 *MedRxiv*. Published online March 28, 2024. doi:10.1101/2024.03.27.24304980
- 473 14. The Hyfe Team. Smart cough monitoring: an innovation milestone for global respiratory
474 health. *Hyfe Research Series*. 2021;(1):1-6. Accessed March 14, 2021.
475 www.hyfeapp.com
- 476 15. Chaibub Neto E, Yadav V, Sieberts SK, Omberg L. A novel estimator for the two-way
477 partial AUC. *BMC Med Inform Decis Mak.* 2024;24(1):57. doi:10.1186/s12911-023-
478 02382-2
- 479 16. World Health Organization. *High-Priority Target Product Profiles for New Tuberculosis*
480 *Diagnostics: Report of a Consensus Meeting.*; 2014. www.who.int
- 481 17. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to
482 analyze and compare ROC curves. *BMC Bioinformatics.* 2011;12(1):77.
483 doi:10.1186/1471-2105-12-77
- 484 18. pROC_based_tpAUC. GitHub. Published 2022.
485 https://github.com/echaibub/pROC_based_tpAUC
- 486 19. Breiman L. Random Forests. *Mach Learn.* 2001;45(1):5-32.
487 doi:10.1023/A:1010933404324
- 488 20. SAGE Bionetworks. CODA TB DREAM Challenge. Accessed April 1, 2024.
489 <https://www.synapse.org/TBcough>

- 490 21. Chang Y, Jing X, Ren Z, Schuller BW. CovNet: A Transfer Learning Framework for
491 Automatic COVID-19 Detection From Crowd-Sourced Cough Sounds. *Front Digit Health*.
492 2022;3:799067. doi:10.3389/FDGTH.2021.799067/BIBTEX
- 493 22. Pahar M, Klopper M, Warren R, Niesler T. COVID-19 cough classification using machine
494 learning and global smartphone recordings. *Comput Biol Med*. 2021;135.
495 doi:10.1016/J.COMPBIOMED.2021.104572
- 496 23. Ghrabli S, Elgendi M, Menon C. Identifying unique spectral fingerprints in cough sounds
497 for diagnosing respiratory ailments. *Sci Rep*. 2024;14(1). doi:10.1038/S41598-023-
498 50371-2
- 499 24. Pentakota P, Rudraraju G, Sripada NR, et al. Screening COVID-19 by Swaasa AI
500 platform using cough sounds: a cross-sectional study. *Sci Rep*. 2023;13(1).
501 doi:10.1038/S41598-023-45104-4
- 502 25. Bhattacharya D, Sharma NK, Dutta D, et al. Coswara: A respiratory sounds and
503 symptoms dataset for remote screening of SARS-CoV-2 infection. *Scientific Data 2023*
504 *10:1*. 2023;10(1):1-11. doi:10.1038/s41597-023-02266-0
- 505 26. Orlandic L, Teijeiro T, Atienza D. The COUGHVID crowdsourcing dataset, a corpus for
506 the study of large-scale cough analysis algorithms. *Sci Data*. 2021;8(1).
507 doi:10.1038/S41597-021-00937-4
- 508 27. Swaminathan S, Padmapriyadarsini C, Narendran G. HIV-Associated Tuberculosis:
509 Clinical Update. *Clinical Infectious Diseases*. 2010;50(10):1377-1386.
510 doi:10.1086/652147
- 511 28. Tavaziva G, Harris M, Abidi SK, et al. Chest X-ray Analysis With Deep Learning-Based
512 Software as a Triage Test for Pulmonary Tuberculosis: An Individual Patient Data Meta-
513 Analysis of Diagnostic Accuracy. *Clinical Infectious Diseases*. 2022;74(8):1390-1400.
514 doi:10.1093/cid/ciab639

- 515 29. Geric C, Qin ZZ, Denkinger CM, et al. The rise of artificial intelligence reading of chest X-
516 rays for enhanced TB diagnosis and elimination. *Int J Tuberc Lung Dis.* 2023;27(5):367-
517 372. doi:10.5588/IJTL.D.22.0687
- 518 30. Sage Bionetworks. CODA TB DREAM Challenge. <http://synapse.org/tbcough>
- 519 31. Yellapu GD, Rudraraju G, Sripada NR, et al. Development and clinical validation of
520 Swaasa AI platform for screening and prioritization of pulmonary TB. *Sci Rep.*
521 2023;13(1). doi:10.1038/S41598-023-31772-9
- 522
- 523

524 **Table 1. Model performance for cough-only model (Sub-challenge 1)**

Rank¹	Team	AUROC (95% CIs)	Model type	Sound features used
1	Blue Team	0.743 (0.703, 0.780)	Convolutional neural network	Spectrogram
2	AI-Campus High School Team	0.731 (0.691, 0.771)	Gradient Boosting Decision Tree	Mel-Frequency Cepstral Coefficients (MFCC), chromagram
2**	Raghava_India_TB	0.730 (0.690, 0.773)	Convolutional neural network	Mel spectrogram
4	Yuanfang Guan Lab Team	0.727 (0.685, 0.768)	Light gradient- boosting machine	Mel Frequency Cepstral Coefficients , first and second order time derivatives of MFCC, magnitude of pitch tracking, total number of coughs recorded
5	Metformin-121	0.704 (0.660, 0.746)	MetforNet ²	Z-score normalization of cough recordings
6	Clare	0.699 (0.655, 0.748)	Artificial Neural	Top 300 features

		0.746)	Network	extracted via OpenSMILE identified via principal component analysis
7	Sakb	0.695 (0.654, 0.739)	Artificial Neural Network	Top 1,024 features extracted via OpenSMILE identified via principal component analysis
7	chsxashoka	0.693 (0.651, 0.736)	Artificial Neural Network	Mel Frequency Cepstral Coefficients, Mel spectrogram
9	LCL	0.689 (0.644, 0.733)	Convolutional neural network, Light gradient- boosting machine	Zero-crossing rate, Mel frequency cepstral coefficients, chromagram, mel spectrogram, root mean square
9	sasgarian	0.689 (0.647, 0.732)	Artificial Neural Network	Top 1,024 features extracted via OpenSMILE

				identified via principal component analysis
11	yhwei	0.645 (0.601, 0.687)	Convolutional Neural Network	Spectrogram

525 1. Models with the same ranking were statistically indistinguishable

526 2. A combined architecture of five convolutional neural network blocks, followed by a

527 bidirectional gated recurrent unit, an attention layer, and a fully connected layer.

528 **Table 2. Model performance for cough sound and clinical data model (Sub-challenge 2)**

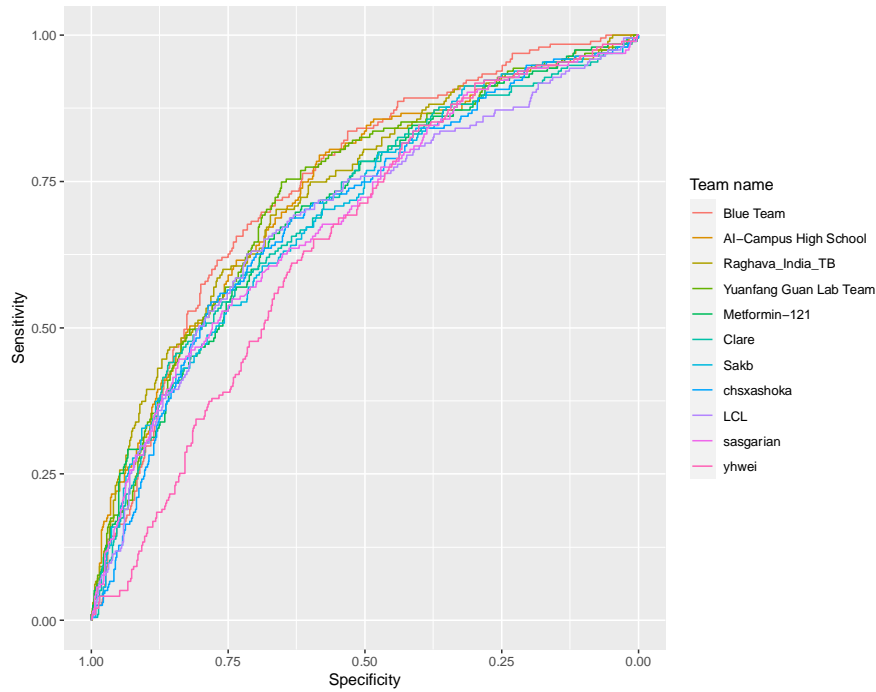
Rank	Team	pAUROC (95% CI)	AUROC (95% CI)
1	Metformin-121	0.003 (6.11e-06, 0.012)	0.832 (0.795, 0.863)
2	Yuanfang Guan Lab Team	0.003 (0, 0.009)	0.821 (0.784, 0.853)
3	AI-Campus High School Team	0.001 (0, 0.008)	0.817 (0.778, 0.850)
4	Blue Team	0.001 (0, 0.007)	0.818 (0.779, 0.853)
5	LCL	0.001 (0, 0.006)	0.792 (0.750, 0.829)
6	yhwei	0 (0, 0.003)	0.784 (0.741, 0.822)

529

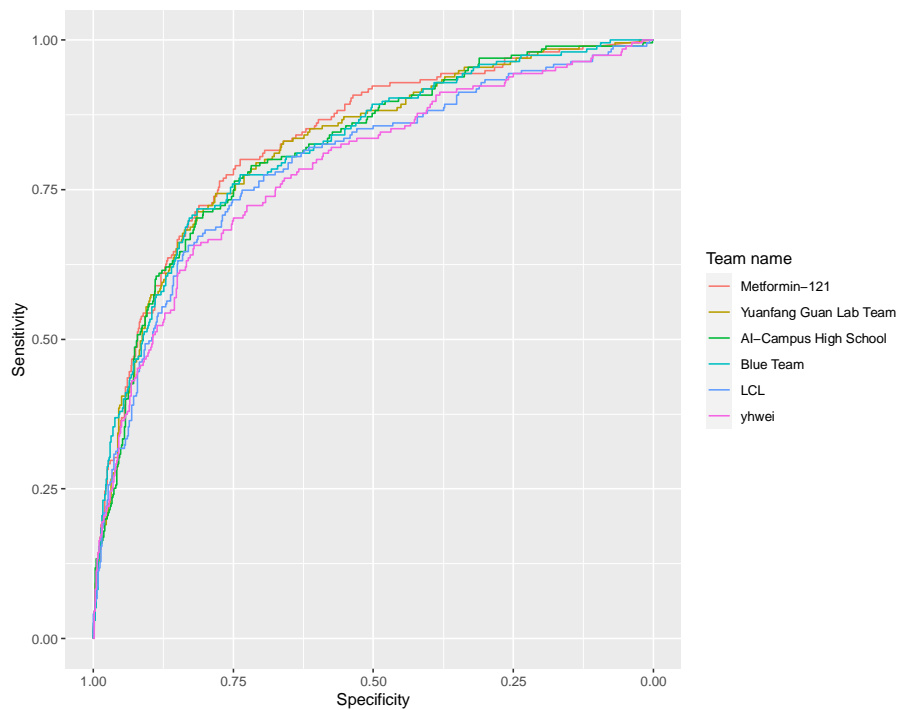
530 **Figure 1. Receiver Operating Characteristic Curves for CODA TB DREAM Challenge Final**

531 **Models**

532 **A. Sub-challenge 1 – Cough Sounds Only**

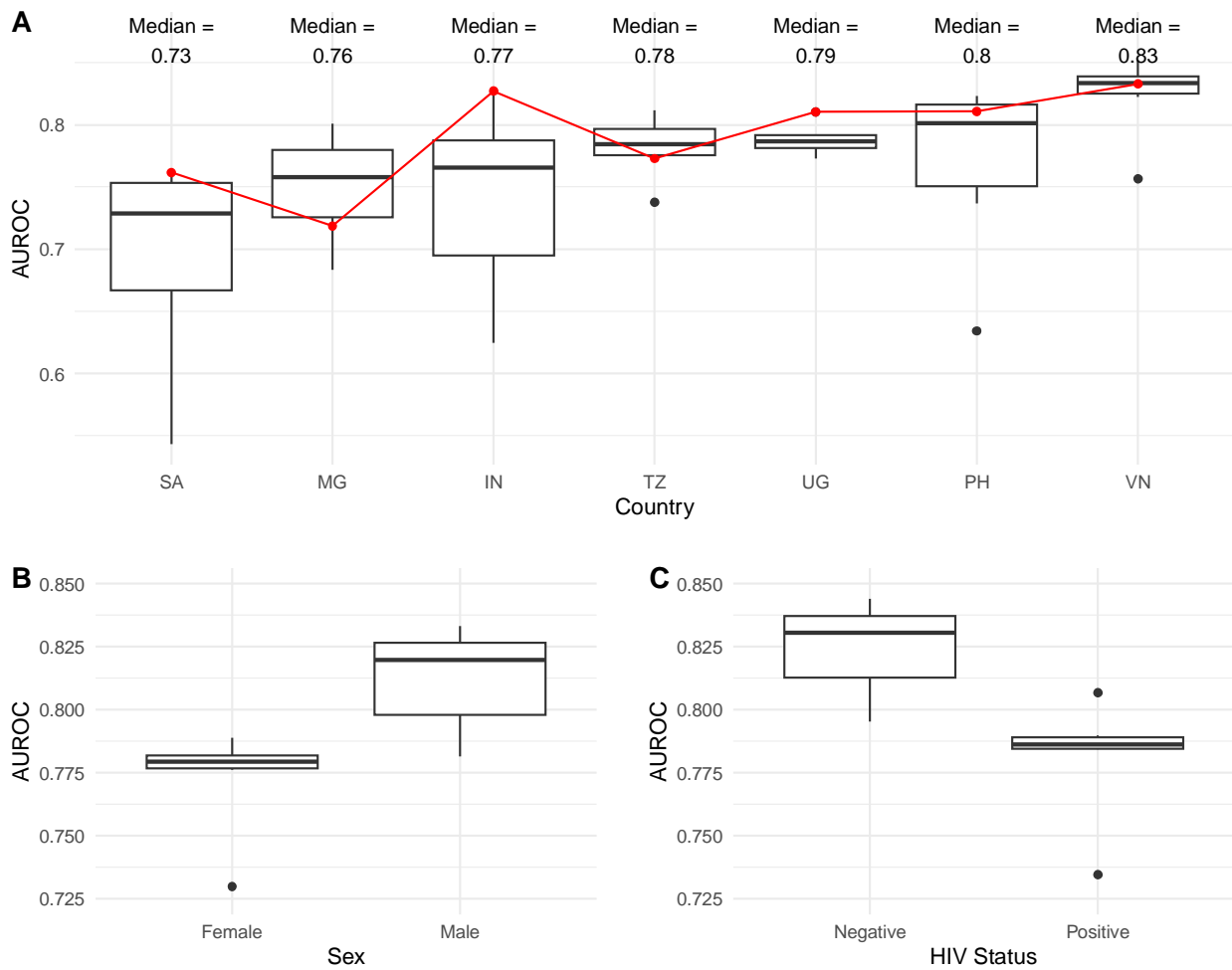


534 **B. Sub-challenge 2 – Cough Sounds and Routine Demographic and Clinical Data**



536 **Figure 2. Comparison of Area under the Receiver Operating Characteristic Curve by**
 537 **Country and Subgroup in Sub-challenge 2.** Box plots of median area under the curve
 538 (AUROC) with interquartile range (IQR) based on all submissions, and stratified by (A) country;
 539 (B) sex and (C) HIV status. For (A), median AUROC indicated at the top, and winning model
 540 AUROC shown in red. SA: South Africa; MG: Madagascar; IN: India; TZ: Tanzania; UG:
 541 Uganda; PH: Philippines; VN: Vietnam.

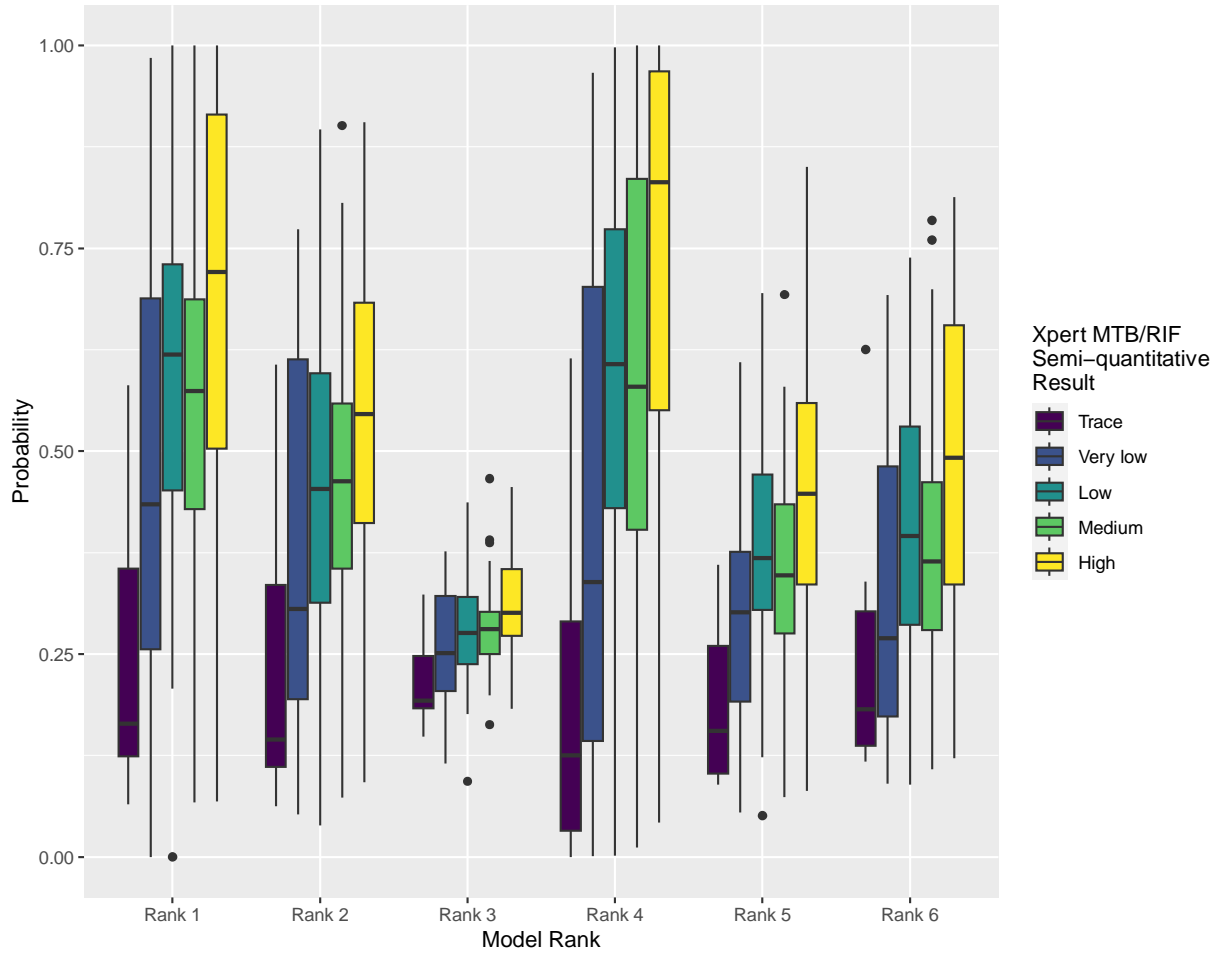
542



543

544

545 **Figure 3. TB probability scores of the cough sound and clinical data models stratified by**
 546 **Xpert semi-quantitative status.** Box plot of the median probability with interquartile range
 547 (IQR). Rank indicates the final challenge ranking. Higher probability scores indicate higher
 548 likelihood that the model would classify the individual as having TB.
 549



550