

1 **TITLE**

2 **Solicited Cough Sound Analysis for Tuberculosis Triage Testing: The CODA TB**
3 **DREAM Challenge Dataset**

4
5 **RUNNING TITLE**

6 Cough sounds for tuberculosis triage

7
8 **AUTHORS**

9 Sophie Huddart^{1*}, Vijay Yadav^{2*}, Solveig K. Sieberts², Larson Omberg^{2,3}, Mihaja
10 Raberahona^{4,5}, Rivo Rakotoarivelo^{6,7}, Issa N. Lyimo⁸, Omar Lweno⁸, Devasahayam J
11 Christopher⁹, Nguyen Viet Nhung¹⁰, Grant Theron¹¹, William Worodria¹², Charles Y.
12 Yu¹³, Christine M Bachman¹⁴, Stephen Burkot¹⁴, Puneet Dewan¹⁴, Sourabh Kulhare¹⁴,
13 Peter M Small¹⁴, Adithya Cattamanchi¹⁵, Devan Jaganath^{15#}, Simon Grandjean
14 Lapierre^{16,17#}

15 *, #Equal contributions

16
17 **AFFILIATIONS**

18 ¹University of California San Francisco, School of Medicine, 533 Parnassus Ave, San
19 Francisco, CA 94143 USA

20 ²Sage Bionetworks, Seattle, WA 98103 USA

21 ³Currently at Koneksa Health, One World Trade Center 285 Fulton St. 77th Floor New
22 York, NY, 10007

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

24 ⁵Centre d'Infectiologie Charles Mérieux, Antananarivo, 101, Analamanga, Madagascar

25 ⁶CHU Tambohobe Fianarantsoa, 301, Haute-Matsiatra, Madagascar

26 ⁷Université de Fianarantsoa, Fianarantsoa, 301, Haute-Matsiatra, Madagascar

27 ⁸Ifakara Health Institute, Environmental and Ecological Sciences & Interventions and
28 Clinical Trials Departments, Kiko Avenue, Plot 463, Mikocheni, Dar es Salaam,
29 Tanzania

30 ⁹Christian Medical College, Ida Scudder Road, Vellore 632004, Tamil Nadu, India

31 ¹⁰National Tuberculosis Programme, 463 Hoang Hoa Tham, Ba Dinh District, Hanoi,
32 Vietnam

33 ¹¹Stellenbosch University, Division of Molecular Biology and Human Genetics,
34 Matieland, 7602 South Africa

35 ¹²Walimu, Plot 5-7, Coral Crescent, Kololo, Kampala, Uganda

36 ¹³De La Salle Medical and Health Sciences Institute, Governor D. Mangubat Avenue,
37 Dasmariñas Cavite, Philippines 4114

38 ¹⁴Global Health Labs, 14360 SE Eastgate Way, Bellevue, WA 98007 USA

39 ¹⁵University of California Irvine, School of Medicine, 333 City Blvd. W Suite 400, Orange
40 CA 92868 USA

41 ¹⁶Centre de Recherche du Centre Hospitalier de l'Université de Montréal,
42 Immunopathology Axis, 900 St-Denis, Montréal, Québec, H2X 0A9 Canada

43 ¹⁷Université de Montréal, Department of Microbiology, Infectious Diseases and
44 Immunology, 2900 Edouard-Montpetit, Montréal, Québec, H3T 1J4 Canada

45

46 **CORRESPONDING AUTHOR**

47 Simon Grandjean Lapierre

48 Simon.grandjean.lapierre@umontreal.ca

49

50 **WORD COUNT**

51 2286

52 **ABSTRACT**

53 Cough is a common and commonly ignored symptom of lung disease. Cough is often
54 perceived as difficult to quantify, frequently self-limiting, and non-specific. However,
55 cough has a central role in the clinical detection of many lung diseases including
56 tuberculosis (TB), which remains the leading infectious disease killer worldwide. TB
57 screening currently relies on self-reported cough which fails to meet the World Health
58 Organization (WHO) accuracy targets for a TB triage test. Artificial intelligence (AI)
59 models based on cough sound have been developed for several respiratory conditions,
60 with limited work being done in TB. To support the development of an accurate, point-
61 of-care cough-based triage tool for TB, we have compiled a large multi-country
62 database of cough sounds from individuals being evaluated for TB. The dataset
63 includes more than 700,000 cough sounds from 2,143 individuals with detailed
64 demographic, clinical and microbiologic diagnostic information. We aim to empower
65 researchers in the development of cough sound analysis models to improve TB
66 diagnosis, where innovative approaches are critically needed to end this long-standing
67 pandemic.

68 BACKGROUND AND SUMMARY

69 Tuberculosis remains the leading infectious disease killer globally, partly due to public
70 health systems' inability to accurately diagnose millions of infected individuals every
71 year.¹ Insufficient access to high-quality TB screening and diagnosis is recognized as
72 one of the most important gaps in the cascade of care.² Here we describe a cough
73 sound database including detailed demographic, clinical and microbiologic information
74 for the development of AI-based sound classification TB triage models. As the WHO's
75 End TB Strategy calls for intensified research and innovation including the discovery of
76 new tools for community-based screening, digital cough monitoring and *Acoustic*
77 *Epidemiology* could represent new tools that can help bend the TB pandemic curve and
78 accelerate the achievement of global TB elimination goals.³⁻⁵

79
80 The “missing millions” of undiagnosed patients living with active TB disease represent
81 an heterogeneous group including those who did not access triage or diagnosis testing
82 or weren't appropriately referred for effective treatment. Improving the accuracy,
83 portability, point-of-care amenability and connectivity of diagnostic tools and algorithms
84 would have significant value. Most health systems build their TB programs on a
85 combination of complementary screening followed by diagnostic tests. The WHO's
86 target product profile (TPP) for a community-based TB triage test suggests that it should
87 be at least 90% sensitive and 70% specific.⁶ According to the 2021 WHO TB screening
88 guidelines, symptom-based screening with questionnaires, including cough, is 42%
89 sensitive.⁷ Besides having poor accuracy these guidelines have operational challenges
90 that impede its sustained and uniform implementation within resource-challenged TB
91 programs. Other tools such as digital chest X-rays combined with computer-aided
92 detection (CAD) algorithms have also been evaluated in the context of TB triage. This
93 approach was shown to be highly sensitive but had variable specificity and remains
94 difficult to deploy due to limited availability of chest X-ray platforms at primary-level
95 health facilities.⁸ Whether in the context of community-based outreach screening or
96 healthcare facility-based evaluation prior to confirmatory testing, cough classification
97 models could complement or replace other triage strategies including symptom-based
98 screening.

99

100 We historically have been unable to objectively monitor cough sounds and consequently
101 reduced this data-rich symptom into subjective and dichotomous information (e.g.,
102 cough versus no cough, chronic versus acute, better versus worse). Advances in
103 acoustics and machine learning (ML) have enabled the identification and recording of
104 human coughs in real-world acoustic environments (cough detection) as well as
105 differentiation of coughs from patients with distinct clinical conditions or at different
106 stages of disease (cough classification). As part of the emerging field of *Acoustic*
107 *Epidemiology*, this has the potential to develop novel screening or diagnostic assays
108 with simple digital recording devices, such as a smartphone, tablet or watch.⁵ Proof-of-
109 concept studies previously showed that cough associated with TB contains a specific
110 acoustic signature which can be recognized by ML models. A study by *Pahar et al.*
111 suggests that a cough-based TB screening model can discriminate TB cough sounds
112 from those associated with other lung conditions with 93% sensitivity and 95%
113 specificity, exceeding the WHO TPPs.⁹ In a study combining cough sound analysis and
114 patients' clinical characteristics, *Yellapu et al.* report that ML can be used to detect TB
115 with 90% sensitivity and 85% specificity.¹⁰ Those pilot studies report on ML models
116 which were designed on small datasets and were not validated in external populations.
117 Given the potential impact on performance of local disease epidemiology and
118 population ethnicity among other confounders, large and diverse cough datasets are
119 needed to replicate those studies.

120
121 We collected and are here releasing a dataset including 733,756 cough sounds from
122 2,143 patients across 7 countries with accurately annotated demographic, clinical and
123 microbiologic diagnostic information. These data were initially used to enable and
124 evaluate the CODA TB DREAM Challenge which invited participants to develop
125 algorithms for prediction of TB diagnosis. The training data are now available for
126 general use, and researchers are invited to leverage acoustic and clinical data to further
127 develop and evaluate sound classification models for TB screening against a held-out
128 test partition.¹¹ We aim to enable the development of models which could achieve the
129 WHO TPP performance targets for the current 'community-based TB triage test' or the
130 forthcoming TPP for a TB screening test.^{6,12} This data set has limitations which include
131 some selection bias since it was collected from a symptomatic presumptive TB

132 population. The developed models which will be developed may hence not perform as
133 well if used for asymptomatic screening at population level. Accordingly, more data
134 should be collected from community screening activities.

135

136 **METHODS**

137 **Participants**

138 A total of 2,143 participants were recruited from two parent studies described below. To
139 be eligible, participants had to be 18 years or older and have a new or worsening cough
140 for at least two weeks. took place at outpatient clinics in India, Madagascar, the
141 Philippines, South Africa, Tanzania, Uganda, and Vietnam. All participants provided
142 informed consent. A summary of participant demographics and country distribution are
143 available in Table 1.

144

145 *Rapid Research in Diagnostic Development TB Network (R2D2 TB Network) study:*

146 The R2D2 TB Network study evaluates novel TB diagnostics in various stages of
147 development among people with presumptive TB in five low- and middle-income
148 countries: Uganda, South Africa, Vietnam, the Philippines and India.¹³ Ethical approval
149 for this study was obtained from institutional review boards (IRB) in the US and in each
150 study site. In the US, approval was obtained from the University of California San
151 Francisco IRB (# 20-32670). In Vietnam, approval was obtained from the Ministry of
152 Health Ethical Committee for National Biological Medical Research (94/CN-HĐĐĐ), the
153 National Lung Hospital Ethical Committee for Biological Medical Research
154 (566/2020/NCKH) and the Hanoi Department of Health, Hanoi Lung Hospital Science
155 and Technology Initiative Committee (22/BVPHN). In India, approval was obtained from
156 Christian Medical College IRB (13256). In South Africa, approval was obtained from
157 Stellenbosch University Health Research Ethics Committee (17047). In Uganda,
158 approval was obtained from Makerere University, College of Health Sciences, School of
159 Medicine, Research Ethics Committee (2020-182). In the Philippines, approval was
160 obtained from De La Salle Health Sciences Institute Independent Ethics Committee
161 (2020-33-02-A).

162

163 *The Digital Cough Monitoring for screening, diagnosis and clinical follow-up of*
164 *tuberculosis and other respiratory diseases project:* This project was designed to embed
165 digital cough monitoring within existing health facility-based TB diagnostic cohorts in
166 Madagascar and Tanzania. Ethical approval for this study was obtained from
167 institutional review boards (IRB) in Canada and in each study site. In Canada, approval
168 was obtained from the Centre de Recherche du Centre Hospitalier de l'Université de
169 Montréal IRB (# 2021-9270, 20.226). In Madagascar, approval was obtained from the
170 Comité d'Éthique à la Recherche Biomédicale (IORG0000851 - N°051-
171 MSANP/SG/AMM/CERBM).). In Tanzania, approval was obtained from the Ifakarah
172 Health Institute IRB (31-2021) and the National Institute for Medical Research
173 (NIMR/HQ/R.8a/Vol IX/3805).

174

175 **Data Collection**

176 *Demographic and clinical data.* At enrollment into the parent studies, participants
177 underwent a baseline questionnaire, clinical examination, and sputum collection for TB
178 testing. Study staff also recorded participants' age, gender, height, weight, smoking
179 status and duration of cough. HIV diagnosis was made either based on participant self-
180 report of a positive HIV diagnosis or a positive test result. A summary of the available
181 variables is shown in Table 2.

182

183 *TB Reference Standard Testing.* Both Xpert MTB/RIF Ultra PCR and mycobacterial
184 culture (Lowenstein-Jensen solid medium or MGIT liquid medium) were performed on
185 sputum collected from all participants. Any participant whose first sputum Xpert
186 MTB/RIF Ultra result was indeterminate or trace-positive, received a second sputum
187 Xpert MTB/RIF Ultra test. Results from those assays were combined to determine TB
188 status according to two reference standards: a microbiologic reference standard and a
189 sputum Xpert reference standard. The sputum Xpert reference standard is restricted to
190 Xpert MTB/RIF Ultra results on sputum samples. The microbiologic reference standard
191 includes culture results, allowing for more individuals to be classified as TB positive.
192 The microbiologic reference standard is considered the primary reference standard. Full
193 details of the reference standards are described in Table 3.

194

195 *Cough Recording.* Cough sounds were collected using smartphones loaded with the
196 Hyfe research app.¹⁴ Specific phone models used in the different participating sites are
197 presented in Supplementary Materials 1. Hyfe research app is designed to listen for
198 explosive sounds and record ~0.5 seconds sound fragments corresponding to putative
199 cough sounds. Hyfe research app uses a server-based convolutional neural network
200 (CNN) model to classify explosive sounds as coughs and recordings of these cough
201 sounds are saved on a protected health information (PHI)-regulated server for analysis.
202 This model has been shown to be 96% sensitive and 96% specific for cough detection
203 using human-labeled sounds as a reference standard.¹⁵ Smartphones were positioned
204 on tripods in rooms within the clinic. Participants were asked to cough five times
205 (solicited cough) while standing 60-90 cm from the tripod; participants who managed to
206 produce at least three coughs were retained in the dataset. Some participants produced
207 more than five coughs due to a triggered coughing fit and those additional coughs were
208 also collected and included in the dataset. Solicited and triggered coughs could not be
209 labeled distinctively and are treated the same in the dataset. After enrollment and
210 onboarding, a subset of participants (n = 565) were also asked to carry a study phone
211 for two weeks and collect longitudinal coughs sounds in an outpatient setting. Those
212 sounds are labeled as longitudinal and made available within the dataset. A tally of
213 solicited and longitudinal cough sounds per data partition are available in Table 4.

214

215 **Data Partitioning**

216 The dataset was split into a training (n=1,105) and validation set (n=1,038). The dataset
217 was randomly partitioned evenly between the training and testing set at the level of the
218 participant (i.e., all of a participant's cough sounds are in either the training or validation
219 set).

220

221 **Data Pre-Processing**

222 *Cough sounds:* The sound recordings available in this dataset have not undergone pre-
223 processing beyond their identification as a cough sound by the Hyfe research app CNN
224 model.

225

226 *Clinical Data:* Data from all participating sites were collected with standardized
227 questionnaires and definitions. Data formatting was harmonized in the open access
228 database.

229

230 **Dataset Description**

231 Sage Bionetworks independently verified the variable balance between the training and
232 validation sets as demonstrated in Table 1. A breakdown of key demographics and
233 microbiologic reference standard results by country are shown in Table 5.

234 **DATA RECORDS**

235 De-identified participant demographic and clinical data, including TB reference standard
236 results, cough sound WAV files, and a datafile linking participant IDs to sound file IDs
237 were exported to a dedicated project in Synapse. Synapse is a general-purpose data
238 and analysis sharing service where members can work collaboratively, analyze data,
239 share insights, and have attributions and provenance of those insights to share with
240 others. Synapse is developed and operated by Sage Bionetworks¹⁶. A total of 1,105
241 participants' data are made available for access and download as a training dataset.
242 The validation set is withheld, but models can be evaluated against the validation set via
243 the instructions provided in the Synapse project.

244
245 All training set files are stored and are accessible via the Synapse platform with
246 associated metadata and documentation and can be accessed at the following URL:
247 www.synapse.org/TBcough - <https://doi.org/10.7303/syn31472953>.

248 249 **TECHNICAL VALIDATION**

250 All cough collection periods were observed by study staff and cough sounds were spot-
251 checked for accurate recording. Patient metadata was reviewed by study staff for
252 accuracy. The data described in this article were collected using the Hyfe Research app
253 which uses a proprietary algorithm to identify cough sounds. We used a prediction score
254 of 0.8 from this algorithm to filter potential non-cough sounds. To validate the precision
255 of the Hyfe algorithm, a standalone computer vision and deep learning model was
256 trained using Log Mel spectrogram images from ESC-50 and Coswara datasets.^{17,18}
257 The "VGG16" CNN based pre-trained model was trained for accurate classification of
258 cough sounds and achieved a model accuracy of approximately 96% on Hyfe cough
259 recordings.¹⁹ Most of the recordings that were classified incorrectly had a Hyfe
260 prediction score less than 0.8.

261

262 **USAGE NOTES**

263 Users can register to evaluate predictive models of TB diagnosis against the held-out
264 test partition via the instructions on the Synapse project. The scoring mechanism can
265 evaluate two different types of models: (1) those that use only cough sounds, or (2)
266 those which also incorporate clinical metadata variables which have been provided in
267 the training dataset (sex, age, height, weight, reported duration of cough, prior TB
268 diagnosis and type, hemoptysis, heart rate, temperature, weight loss, smoking in the
269 last week, fever and night sweats). Models are submitted to the scoring queues as
270 Docker images. Full instructions and example code is available on Synapse project
271 website (www.synapse.org/TBcough).

272

273 **Downloading the Data**

274 Given the number of files represented in the data, users should consider downloading
275 the data via one of the programmatic Synapse clients (available in R or Python). For
276 convenience, Python code for downloading the data is provided in the Synapse project
277 wiki. The training dataset size is (0.43 GB) for the solicited coughs and (31.6 GB) for the
278 longitudinal coughs.

279

280 **Data Use Agreement**

281 To access the data, individuals must become Certified and Validated users of Synapse
282 and maintain an active account on Synapse: <http://www.synapse.org>. They must also
283 submit an Intended Data Use Statement and agree to the Terms of Use of the dataset.
284 Terms of Use are summarized in Supplementary Materials 2.

285

286 **CODE AVAILABILITY**

287 No additional data processing was conducted other than what has been described
288 above.

289

290 **DATA AVAILABILITY**

291 All training set files are stored and are accessible via the Synapse platform with
292 associated metadata and documentation and can be accessed at the following URL:

293 www.synapse.org/TBcough.

294
295 Specific acoustic and clinical metadata as well as dataset information can be found
296 under the following doi references.
297 Clinical data: <https://doi.org/10.7303/syn53710097>
298 Cough metadata: <https://doi.org/10.7303/syn53710098>
299 Solicited Coughs: <https://doi.org/10.7303/syn40358494>
300 Longitudinal Coughs: <https://doi.org/10.7303/syn40358476>
301 Data Dictionary: <https://doi.org/10.7303/syn41743692>
302 Data sharing and model benchmarking website: <https://doi.org/10.7303/syn31472953>

303 REFERENCES

- 304 1 WHO. Global TB report 2023. (2023).
- 305 2 Faust, L. *et al.* Improving measurement of tuberculosis care cascades to
306 enhance people-centred care. *Lancet Infect Dis* **23**, e547-e557,
307 doi:10.1016/S1473-3099(23)00375-4 (2023).
- 308 3 Houben, R. *et al.* Feasibility of achieving the 2025 WHO global tuberculosis
309 targets in South Africa, China, and India: a combined analysis of 11
310 mathematical models. *Lancet Glob Health* **4**, e806-e815, doi:10.1016/S2214-
311 109X(16)30199-1 (2016).
- 312 4 WHO. The end TB strategy. (2015).
- 313 5 Zimmer, A. J. *et al.* Making cough count in tuberculosis care. *Commun Med*
314 (*Lond*) **2**, 83, doi:10.1038/s43856-022-00149-w (2022).
- 315 6 WHO. High-priority target product profiles for new tuberculosis diagnostics: report
316 of a consensus meeting. (2014).
- 317 7 WHO. WHO operational handbook on tuberculosis. Module 2: screening -
318 systematic screening for tuberculosis disease. (2021).
- 319 8 Tavaziva, G. *et al.* Chest X-ray Analysis With Deep Learning-Based Software as
320 a Triage Test for Pulmonary Tuberculosis: An Individual Patient Data Meta-
321 Analysis of Diagnostic Accuracy. *Clin Infect Dis* **74**, 1390-1400,
322 doi:10.1093/cid/ciab639 (2022).
- 323 9 Pahar, M. *et al.* Automatic cough classification for tuberculosis screening in a
324 real-world environment. *Physiol Meas* **42**, doi:10.1088/1361-6579/ac2fb8 (2021).
- 325 10 Yellapu, G. D. *et al.* Development and clinical validation of Swaasa AI platform
326 for screening and prioritization of pulmonary TB. *Sci Rep* **13**, 4740,
327 doi:10.1038/s41598-023-31772-9 (2023).
- 328 11 Bionetworks, S. CODA TB DREAM Challenge, <www.synapse.org/TBcough>
329 (2024).
- 330 12 WHO. *Public consultation for Target Product Profiles for TB screening tests*,
331 <[https://www.who.int/news-room/articles-detail/public-consultation-for-target-](https://www.who.int/news-room/articles-detail/public-consultation-for-target-product-profiles-for-tb-screening-tests)
332 [product-profiles-for-tb-screening-tests](https://www.who.int/news-room/articles-detail/public-consultation-for-target-product-profiles-for-tb-screening-tests)> (2024).
- 333 13 R2D2. *Rapid Research in Diagnostics Development for TB Network*,
334 <<https://www.r2d2tbnetwork.org/>> (2024).
- 335 14 HyfeAI. *Hyfe AI*, <<https://www.hyfe.ai/>> (2024).
- 336 15 Gabaldon-Figueira, J. C. *et al.* Digital acoustic surveillance for early detection of
337 respiratory disease outbreaks in Spain: a protocol for an observational study.
338 *BMJ Open* **11**, e051278, doi:10.1136/bmjopen-2021-051278 (2021).
- 339 16 Omberg, L. *et al.* Enabling transparent and collaborative computational analysis
340 of 12 tumor types within The Cancer Genome Atlas. *Nat Genet* **45**, 1121-1126,
341 doi:10.1038/ng.2761 (2013).
- 342 17 Piczak, K. Dataset for Environmental Sound Classification. *Proceedings of the*
343 *23rd ACM international conference on Multimedia*,
344 doi:doi.org/10.1145/2733373.2806390 (2015).
- 345 18 Bhattacharya, D. *et al.* Coswara: A respiratory sounds and symptoms dataset for
346 remote screening of SARS-CoV-2 infection. *Sci Data* **10**, 397,
347 doi:10.1038/s41597-023-02266-0 (2023).

348 19 Simonyan K., Z. A. Very deep convolutional networks for large-scale image
349 recognition. *conference paper* @CLR2015,
350 doi:<https://doi.org/10.48550/arXiv.1409.1556> (2015).
351

352 **TABLES**

353 **Table 1 – Participant demographics across training and test sets**

	Training set (N=1105)	Testing set (N=1038)	Complete set (N=2143)
Sex			
Female	517 (46.8%)	460 (44.3%)	977 (45.6%)
Male	588 (53.2%)	578 (55.7%)	1166 (54.4%)
Age			
Median [Q1, Q3]	40.0 [28, 53]	40.0 [29, 53]	40.0 [28, 53]
Height (cm)			
Median [Q1, Q3]	162 [155, 168]	162 [156, 168]	162 [156, 168]
Weight (Kg)			
Median [Q1, Q3]	55.0 [49, 65]	56.0 [49, 65]	55.8 [49, 65]
HIV status			
Negative	878 (79.5%)	809 (77.9%)	1687 (78.7%)
Positive	162 (14.7%)	155 (14.9%)	317 (14.8%)
Unknown	65 (5.9%)	74 (7.1%)	139 (6.5%)
Duration of cough (days)			
Median [Q1, Q3]	30.0 [16, 60]	30.0 [16, 50]	30.0 [16, 60]
Prior TB			
No	903 (81.7%)	835 (80.4%)	1738 (81.1%)
Not sure	3 (0.3%)	2 (0.2%)	5 (0.2%)
Yes	199 (18.0%)	201 (19.4%)	400 (18.7%)
Country			
India	119 (10.8%)	122 (11.8%)	241 (11.2%)
Madagascar	159 (14.4%)	75 (7.2%)	234 (10.9%)
The Philippines	198 (17.9%)	190 (18.3%)	388 (18.1%)
South Africa	137 (12.4%)	138 (13.3%)	275 (12.8%)
Tanzania	87 (7.9%)	110 (10.6%)	197 (9.2%)
Uganda	242 (21.9%)	245 (23.6%)	487 (22.7%)

	Training set (N=1105)	Testing set (N=1038)	Complete set (N=2143)
Vietnam	163 (14.8%)	158 (15.2%)	321 (15.0%)
Microbiologic reference standard			
TB Negative	807 (73.0%)	782 (75.3%)	1589 (74.1%)
TB Positive	297 (26.9%)	256 (24.7%)	553 (25.8%)
Sputum Xpert reference standard			
Indeterminate	4 (0.4%)	2 (0.2%)	6 (0.3%)
TB Negative	839 (75.9%)	808 (77.8%)	1647 (76.9%)
TB Positive	262 (23.7%)	226 (21.8%)	488 (22.8%)

355 **Table 2 - Available demographic, clinical and microbiologic variables**

Variable	Values	Format/Definition
Country	Philippines Vietnam South Africa Uganda India Madagascar Tanzania	Country where the participant was enrolled
Sex	Male Female	Sex at birth reported by participant
Age	Numeric	Age calculated as date of collection - date of birth if known. If date of birth is unknown, reported age at time of collection.
Height	Numeric	Height in centimeters
Weight	Numeric	Weight in Kg
HIV status	Positive Negative	HIV status. All participants who do not report being HIV-positive receive HIV testing (using capillary or venous blood) Positive = Positive reported by patient or positive on a HIV test Negative = Negative on a HIV test
Reported duration of cough	Numeric	Self-reported duration of current cough (days). At baseline, we ask: How many days have you had this new cough or cough that has been worse?
Prior TB	Yes No	Self-reported. At baseline, we ask: Have you ever had or been told you had tuberculosis (TB)?
Prior TB type: Pulmonary	Checked Unchecked	Asked at baseline: "With what kind of TB were you diagnosed?" (may select more than one) Participant selected pulmonary TB.
Prior TB type: Extrapulmonary	Checked Unchecked	Asked at baseline: "With what kind of TB were you diagnosed?" (may select more than one) Participant selected extrapulmonary TB.
Prior TB type: Unknown	Checked Unchecked	Asked at baseline: "With what kind of TB were you diagnosed?" (may select more than one) Participant selected Unknown.
Hemoptysis	Yes No	In the past 30 days, have you ever coughed up blood?
Heart rate	Numeric	Participant's heartrate (beats per minute) measured at baseline
Temperature	Numeric	Participant's temperature (Celsius) measured at baseline.
Weight loss	Yes No	Subjective, Self-reported. At baseline, we ask: In the past 30 days, have you experienced any weight loss?

Smoked in last week	Yes No	Asked at baseline "Have you used combustible tobacco and/or vaping products in the last 7 days?"
Fever	Yes No	Self-reported fever at baseline: "In the past 30 days, have you ever felt or experienced fever?"
Night sweats	Yes No	Self-reported night sweats at baseline: "In the past 30 days, have you ever experienced night sweats?"
Microbiologic reference standard*	TB Positive TB Negative Indeterminate	TB diagnosis based on sputum and culture results
Sputum Xpert reference standard*	Positive Negative Indeterminate	TB diagnosis based on sputum results alone
Xpert combined semi-quant	Trace Very Low Low Medium High	Highest semiquantitative result from sputum Xpert Ultra test conducted at baseline visit. A semiquantitative result will only be available if the test was positive.

357 **Table 3 – Reference standard definitions**

Result	Microbiologic Reference Standard	Sputum Xpert Reference Standard
TB positive	<p>One or more of the following:</p> <ul style="list-style-type: none"> • a positive sputum Xpert result • a positive urine Xpert result • a positive liquid culture result • a positive solid culture result • a positive Xpert Ultra results on contaminated liquid culture • two trace positive Xpert Ultra results on any sample type. 	<p>One of the following:</p> <ul style="list-style-type: none"> • a positive sputum Xpert result • two trace-positive sputum Xpert Ultra results
TB negative	<p>No positive results on any Xpert or Xpert Ultra testing and one of the following conditions:</p> <ul style="list-style-type: none"> • two negative liquid cultures • two negative solid cultures • one negative liquid and one negative solid cultures 	A negative sputum Xpert result
Indeterminate	No positive results on any Xpert or Xpert Ultra testing and only one trace result on Xpert Ultra testing of any sample type	An indeterminate or trace-positive sputum Xpert result followed by a non-positive result on repeat sputum Xpert Ultra testing

359 **Table 4 – Number of cough recordings per data partition**

	Training set	Testing set
Solicited coughs (Participants)	9,772 (1,082)	9,062 (1,038)
Longitudinal coughs (participants)	714,922 (565)	0 (0)

360 *Longitudinal coughs are not used for evaluation on the testing set, so are not quantified*
 361 *here.*

362

363 **Table 5 – Key variables summarized by country**

	India n = 241	Madagascar n = 234	Philippines n = 388	South Africa n = 275	Tanzania n = 197	Uganda n = 487	Vietnam n = 321
Sex							
Female	100 (41%)	115 (49%)	208 (54%)	144 (52%)	86 (44%)	203 (42%)	121 (38%)
Male	141 (59%)	119 (51%)	180 (46%)	131 (48%)	111 (56%)	284 (58%)	200 (62%)
Age							
Median [Q1, Q3]	47 [34, 58]	32 [24, 51]	39 [26, 53]	40 [32, 49]	40 [31, 50]	32 [26, 42]	54 [41, 64]
Duration of cough (days)							
Median [Q1, Q3]	45 [30, 90]	30 [17, 60]	21 [15, 30]	21 [14, 28]	30 [14, 30]	30 [16, 60]	30 [30, 90]
Prior TB							
No	204 (85%)	207 (88%)	325 (84%)	170 (62%)	147 (75%)	427 (88%)	258 (80%)
Not sure	0 (0%)	0 (0%)	4 (1.0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.3%)
Yes	37 (15%)	27 (12%)	59 (15%)	105 (38%)	50 (25%)	60 (12%)	62 (19%)
Microbiologic reference standard							
TB Negative	216 (90%)	131 (56%)	348 (90%)	223 (81%)	161 (82%)	313 (64%)	199 (62%)
TB Positive	25 (10%)	103 (44%)	40 (10%)	52 (19%)	36 (18%)	174 (36%)	122 (38%)

364

365 **SUPPLEMENTARY MATERIALS**

366 **Supplementary materials 1 - Phone models used in the different participating**
367 **sites**

368 India

369 - Redmi 9 Prime

370 - Realme Narzo20

371 Madagascar

372 - Motorola G9 play

373 Philippines

374 - Myphone myWX2 Pro

375 - Xiaomi 9C

376 South Africa

377 - Nokia 3.1

378 - Nokia 5.4

379 - Xiaomi Redmi 9A

380 Tanzania

381 - Nokia 3.4 Ta-1288

382 Uganda

383 - Motorola G16

384 - Samsung M11

385 - Nokia model 5.3

386 Vietnam

387 - OPPOA54

388 **Supplementary materials 2 – Data Use Agreement**

389

390 Researchers wishing to access the data must:

391

392 ☐ You must reaffirm your commitment to the Synapse Pledge and must abide by
393 the guiding principles for responsible research use and data handling within the
394 Synapse Commons Platform as described in the Synapse Governance
395 documents.

396 ☐ You will not attempt to establish the identity of, or attempt to contact any of the
397 subjects included in the data.

398 ☐ You confirm that if you inadvertently receive identifiable information or otherwise
399 identify a subject, you will promptly notify the ACT by emailing
400 act@sagebase.org.

401 ☐ You agree to establish appropriate administrative, technical, and physical
402 safeguards to prevent unauthorized use of or access to the Data.

403 ☐ You will report any data misuse or breach of data security to ACT by emailing
404 act@sagebase.org.

405 ☐ You will use the data only as identified in your intended data use statement
406 (IDU), submitted through Synapse. The IDU should be written in English and
407 must describe the objectives of the proposed research and study design and
408 analysis plan (500 word maximum).

409 ☐ Data accessors must acknowledge the following in all publications or
410 presentations as follows:

411 “The datasets used for the analyses described were contributed by Dr. Adithya
412 Cattamanchi at UCSF and Dr. Simon Grandjean Lapierre at University of
413 Montreal and were generated in collaboration with researchers at Stellenbosch
414 University (PI Grant Theron), Walimu (PIs William Worodria and Alfred Andama);
415 De La Salle Medical and Health Sciences Institute (PI Charles Yu), Vietnam
416 National Tuberculosis Program (PI Nguyen Viet Nhung), Christian Medical
417 College (PI DJ Christopher), Centre Infectiologie Charles Mérieux Madagascar
418 (PIs Mihaja Raberahona & Rivonirina Rakotoarivelo), and Ifakara Health Institute
419 (PIs Issa Lyimo & Omar Lweno) with funding from the U.S. National Institutes of

420 Health (U01 AI152087), The Patrick J. McGovern Foundation and Global Health
421 Labs.”

422

423 **ACKNOWLEDGMENT**

424 We thank all patients and families who participated in this study. We thank healthcare
425 providers, research personnel, laboratory technicians involved in patient recruitment
426 and data collection.

427

428 - Funders / operations

429 Data from the CODA TB DREAM Challenge was generated with support from Global
430 Health Labs, the Patrick J. McGovern Foundation, and the National Institute of Allergy
431 and Infectious Diseases of the US National Institutes of Health under award number
432 U01AI152087. The CODA TB DREAM Challenge and post-challenge evaluation was
433 funded in part by the Bill & Melinda Gates Foundation.

434

435 - Direct salary support

436 SGL is supported by a Junior 1 Salary Award from the Fonds de Recherche Santé
437 Québec. DJ is supported by funding by the National Institutes of Health

438

439 **AUTHOR CONTRIBUTION**

440 Conception (SKS, LO, CB, PD, PMS, AC, SGL), data acquisition (MR, RR, IL, OL, DJC,
441 NVN, GT, WW, CY), data analysis (SH, VY, SKS, SK, DJ), data interpretation (N/A),
442 drafting the work (SH, SGL), reviewing the work critically for important intellectual
443 content and final approval of the version to be published (all co-authors).

444

445 Agreement to be accountable for all aspects of the work in ensuring that questions
446 related to the accuracy or integrity of any part of the work are appropriately investigated
447 and resolved (SH, SKS, AC, SGL).

448

449 **COMPETING INTERESTS**

450 The authors declare no competing interest.