

Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: Cain KP, McCarthy KD, Heilig CM, et al. An algorithm for tuberculosis screening and diagnosis in people with HIV. *N Engl J Med* 2010;362:707-16.

SUPPLEMENTARY APPENDIX for An Algorithm for Tuberculosis Screening and Diagnosis in Persons with HIV

EPIDEMIOLOGIC PROFILE OF STUDY SITES

This study enrolled patients in Cambodia, Vietnam, and Thailand, all three of which are on the World Health Organization (WHO) list of 22 countries with a high burden of TB. The estimated TB incidence (for all forms of TB) in 2007 was 495/100,000 population for Cambodia, 142/100,000 for Thailand, and 171/100,000 for Vietnam.¹

METHODS

Laboratory Methods

All specimens (sputum, urine, stool, and lymph node aspirates) were transported to one reference laboratory in each country within 48 hours of collection for acid-fast bacilli (AFB) smear microscopy and mycobacterial culture on solid media [Lowenstein-Jensen (LJ)]. In Vietnam and Thailand, broth-based mycobacterial culture using the fully automated BACTEC MGIT 960 (Becton Dickinson, Cockeysville, Md.) was also performed. The MGIT 960 technology was not available in Cambodia at the time of this study. The automated BACTEC 9050/9120 instrument (Becton Dickinson, Cockeysville, Md.) was used for mycobacterial blood culture in all 3 countries.

Sputum specimen processing

To ensure standardization of sputum processing methods across sites, sputum specimens were digested and decontaminated using BBL MycoPrep Reagent (BD Diagnostic

Systems, Sparks, Md.) containing 2% NaOH-Na-citrate and *N*-acetyl-L-cysteine (NALC). Specimens were first adjusted to 10mL using sterile water, and then treated with an equal volume of the processing reagent (final concentration of 1% NaOH). Specimens were mixed well, and incubated for 15-20 minutes at room temperature. Following incubation, BBL MycoPrep phosphate buffer solution (PBS) [pH 6.8] was added for a total specimen volume of 45mL. Specimens were mixed well, and concentrated using centrifugation at 3,000 x g, 4°C for 15-20 minutes. Specimens were then decanted completely and re-suspended in 2mL of BBL MycoPrep PBS [pH 6.8]. The suspension was mixed thoroughly and used to inoculate solid media [LJ] x 2 (150µL-200µL each) and broth-based media (Mycobacterial Growth Indicator Tube [MGIT] x 1[500µL]) (BD, Franklin Lakes, New Jersey). The use of *N*-acetyl-L-cysteine–sodium citrate–sodium hydroxide (NALC-NaOH) followed by neutralization with PBS [pH 6.8] is the standard recommended procedure for use with MGIT, and is recommended by the US Centers for Disease Control. The use of solid and liquid media for maximum recovery of mycobacteria has also been previously recommended.²⁻⁴

Non-sputum specimen processing

Non-sputum specimens were handled similarly to sputum specimens with the following exceptions. Blood from patients was inoculated directly into Myco/F Lytic bottles (BD, Franklin Lakes, New Jersey) and placed into an automated blood culture instrument (BACTEC 9050/9120 system).³ Lymph node aspirates were collected aseptically and considered sterile. Therefore, aspirates were directly inoculated into MGIT; if sufficient specimen volume remained after inoculation into MGIT, an LJ culture was inoculated

and a smear was prepared.³ Stool was prepared for decontamination by emulsifying 1 gram in 10mL of sterile water and sterile glass beads, then filtering. Urine was first concentrated by centrifugation, decanted, and adjusted to 10mL using sterile water.

Smear Microscopy

After specimen processing and culture inoculation, concentrated smears were prepared, stained with Ziehl-Neelsen stain, and graded according to WHO recommendations. The actual number of AFB observed on smear was documented for all smears having 1-9 AFB per 100 fields. Smears were then documented and reported using the following scale: Negative (no AFB seen per 100 fields), 1+ (10-99 AFB per 100 fields), 2+ (1-10 AFB per field in at least 50 fields) and 3+ (>10 AFB in at least 20 fields). Smears documented as having 1-3 AFB per 100 fields were recoded as negative for the purpose of analysis based on the known low correlation of 1 to 3 AFB and positive cultures.^{5,6} Smears with 4-9 AFB per 100 fields and those reported as 1+, 2+, or 3+ were classified as positive.

Mycobacterial Culture

LJ cultures were evaluated twice within the first week of inoculation and then once per week for 42 days. Cultures with growth were confirmed as AFB positive by visual inspection or smear. MGIT and Myco/F Lytic cultures were incubated for 42 days in the BACTEC MGIT 960 and BACTEC 9050/9120, respectively. Cultures flagged as positive by either instrument were removed, had AFB smears performed, and were sub-cultured on blood agar plate (BAP). AFB-positive cultures with no contamination were sub-

cultured onto 2 LJ; AFB-positive but contaminated cultures were re-decontaminated and sub-cultured onto 2 LJ; and cultures less than 42 days old with no organisms present on smear and BAP were returned to the instrument. AFB-negative and contaminated MGIT cultures were documented and discarded; AFB-negative Myco/F Lytic cultures that had growth on BAP underwent bacterial or fungal identification. All MGIT tubes and Myco/F Lytic cultures were removed after 42 days, visually inspected for growth, and then discarded. Blood cultures were sub-cultured onto 2 LJ and incubated for an additional 3 weeks.

MTB Identification

Positive LJ or MGIT cultures were identified as MTB using the niacin production and nitrate reduction tests. In patients for whom no culture was positive for MTB and one or more MGIT cultures grew NTM, the MGIT cultures that grew NTM additionally were tested with a DNA hybridization system (AccuProbe, Gen-Probe, San Diego, California) to confirm the presence or absence of MTB.⁷

Quality Control

Extensive quality control procedures were implemented in each collection site and laboratory to reduce the possibility of false-positive results, including from cross-contamination. Microscopists were not blinded to the results of previous tests, but smears were only judged to be positive if they were confirmed by two different readers. Smears with 1-9 AFB per 100 fields were re-read on-site by independent microbiologists from CDC. We systematically evaluated all patients with at least one culture positive for TB

for possible cross-contamination with the following approach: 1) If a patient had more than one specimen culture positive for TB, then we assumed that it was a true positive; 2) If a patient had only one culture-positive specimen, then we checked to see if another positive culture was processed in the same laboratory on the same day, which would make cross-contamination a possibility. If not, then the investigation ended, and we assumed that the result was a true positive; 3) All single positive cultures were compared to all other positive cultures processed in the same lab on the same day. These cultures were evaluated by genotyping, including spoligotyping and 24-loci mycobacterial interspersed repetitive unit-variable number tandem repeat analysis (MIRU-VNTR). If 2 or more isolates had matching spoligotype and MIRU, cross-contamination was considered possible. A matching isolate makes “cross-contamination” possible but does not confirm it, because community transmission could result in the same finding, especially since some patients enrolled were family members, e.g., husband-wife pairs. In this study, we only found one patient of the 267 TB patients for whom cross-contamination was a possibility. We included this patient in the analysis as a TB patient, since cross-contamination could not be confirmed. Exclusion of the patient from the analysis would not have substantially changed our findings. Genotyping was performed at US CDC supported laboratories in the U.S. using established methods.^{8,9}

Chest Radiography

For the purposes of patient care, chest radiographs were read immediately on site by the clinician caring for the patient. For the purposes of the study and data analysis, all chest radiographs were read by one of a team of three trained readers who were hired to read

all chest radiographs. This person was blinded to the patient's clinical information. One reader read all of the films for patients enrolled in Vietnam, another reader read all films for patients enrolled in Thailand and films of half of the patients enrolled in Cambodia, and one read the films for the other half of the patients enrolled in Cambodia. Both the clinician-reader and the study reader were asked to review standardized elements of the chest radiograph and record the results on a standardized data collection form. For data analysis, we used the reading from the study readers.

Statistical Methods

We defined any patient with at least 1 specimen from any site MTB culture-positive as a TB case, and any patient for whom no culture from any site was MTB culture-positive and at least 1 sputum specimen and 1 non-sputum specimen was MTB culture-negative as not having TB. Patients who did not meet either definition (e.g., had multiple contaminated or missing cultures) were excluded from further analysis.

We compared characteristics of TB patients to those of patients without TB to derive an algorithm for TB screening and diagnosis. Our general approach was to divide the analysis into a TB screening step, which made use of only those variables which could be readily ascertained at any level of the health system (e.g., signs, symptoms, exposure history), followed by a diagnostic step, which included tests for which the availability may be more limited (e.g., smear microscopy, chest radiography, blood tests).

TB screening

The goal of the TB screening step is to divide the population into two groups: 1) Those who do not have TB and can be safely started on ART and IPT without further evaluation for TB; and 2) those who need further evaluation for TB. We calculated the sensitivity, specificity, predictive values, and likelihood ratios for individual predictors and for all potential combinations of 2-5 predictors. In total, 100 different predictors were considered. Some of these predictors considered were variations of a single predictor (i.e., any cough in the past 4 weeks, cough >2 weeks, cough >3 weeks, cough with sputum).

We conducted an exhaustive search over nearly 80 million combinations of 100 signs, symptoms, and medical history indicators to identify combinations with sensitivity $\geq 85\%$ and high specificity for a given sensitivity level. Among combinations with sensitivity $\geq 85\%$, the most specific combinations tended to lie above a line with slope -2 or -2.5 . Therefore, as a guide for comparing the large number of combinations, we computed scores equaling $2 \times \text{sensitivity} + \text{specificity}$ and $2.5 \times \text{sensitivity} + \text{specificity}$ as a basis for comparing different algorithms. Because missing a case of TB in a person with HIV can have serious adverse consequences, we required the screening step to be highly sensitive, and weighted sensitivity greater than specificity in these scores. We considered both $2 \times$ and $2.5 \times$ sensitivity in our calculations because the optimal relative weights are not known. The algorithms presented in the results were those with the highest calculated scores.

Many of the “best” algorithms have similar performance characteristics. The diagnostic odds ratio, which can be calculated by dividing likelihood ratio positive by likelihood ratio negative, is one way of comparing the algorithms. When we did this, we found that the highest diagnostic odds ratio for a 1 of 2 combination meeting our desired criteria was 6.26 (the first listed combination). The diagnostic odds ratios for the 1 of 3 combinations shown range from 6.67-7.63, all exceeding the maximum diagnostic odds ratio achieved by the best 1 of 2 combinations. The 1 of 4 combinations shown have diagnostic odds ratios of 8.22 and 7.04. This range overlaps the range of diagnostic odds ratios for the 1 of 3 combinations. Additionally, the combination with a diagnostic odds ratio of 8.22 has the same sensitivity and predictive value negative as the best 1 of 3 combination, thus adding complexity with minimal gain. This approach to comparing solutions is often applied when using statistical learning methods.¹⁰ The best way to accurately compare the “best” performing algorithms from each group to truly determine which one is superior in clinical practice would be in a prospective replication / validation study, which we propose should be done. Such an evaluation should include both statistical and non-statistical measures, evaluating not only the operating characteristics of each algorithm, but also assessing other factors, including clinician acceptability.

For each candidate rule, we applied the results to subgroups of the data to determine whether performance varied markedly across groups. We used predictive value negative as our primary measure of consistency of performance, since this equates best to the number of false negatives that could be expected. Candidates with the most consistent performance across subgroups were favored, since these are more likely to perform well

in a variety of settings. The subgroups that were planned at the beginning of the study included sex, country / site, CD4, and age. The most useful subgroup for age would be children <18 compared to adults. Since only 4 children under 18 were enrolled, we did not do this analysis, but we did analyze the other subgroups. We tested for heterogeneity in performance across subgroups using the Breslow-Day test.

We also used recursive partitioning to determine how well a rule could be expected to perform and whether tree-based algorithms performed acceptably well at varying levels of complexity.^{11, 12} We were not able to identify any single tree that performed better than the combinations. By allowing us to value sensitivity more than specificity, as was needed for this analysis, the combinatorial approach proved to be superior to certain other methods, including traditional logistic regression, which do not readily lend themselves to doing this. In addition, the end result of a tool based on the combinatorial approach is a tool where a provider can assess a patient for any of several specific features. If a pre-designated number of features are identified, then the patient is classified as screening positive. This is easier to implement than a scoring system or multi-level decision tree. Because of its simplicity, it should be more readily usable at the lowest levels of the health system.

TB diagnosis in patients for whom TB was not excluded in the screening step

We analyzed data for all patients needing further evaluation based on the above screening step. The goal of this analysis was to divide the remaining population into three groups:

1) Those for whom TB is diagnosed without further evaluation needed; 2) those for

whom TB is excluded (e.g., can start ART and IPT safely); and 3) those for whom a clinical judgment is needed, followed by confirmatory TB culture. In this step, the predictors we evaluated included smear microscopy for AFB, chest radiography, and blood tests. We used recursive partitioning analysis to construct a decision tree.¹²

Determining potential validity of results on future data

Random forest analysis is a method that generates a classification rule by combining large number of trees from randomly re-sampled subsets of the data. This process helps to estimate the optimal expected performance (known as the Bayes error rate) of decision rules. Rules which appear to perform better than the random forest analysis would predict can be assumed to result from over-fitting and cannot be expected to generalize well to new data. A random forest analysis of our data showed that screening rules with sensitivity around 90% achieved optimal specificity around 40%. Rules which would appear to perform better than this can be assumed to result from over-fitting and cannot be expected to generalize well to new data.^{11, 13} For each candidate rule, we also evaluated performance across subgroups as described above.

Comparison of algorithms to other approaches

We used data from the study to compare the outcomes of using algorithms developed in this study to the current WHO algorithm for diagnosing smear-negative TB,¹⁴ which is used for TB screening and diagnosis in people with HIV in some settings, and additionally to an approach of doing chest radiography and 2 sputum smears for all patients. We determined the number and characteristics of patients misclassified. For the

purposes of this comparison, we defined as a false negative any patient who had TB but for whom TB was neither diagnosed nor evaluated with TB culture.

ADDITIONAL RESULTS

Chest radiography

Because chest radiography is often not available at peripheral levels of the health system in developing countries, we did not include it in our primary analysis of screening combinations. We did determine, however, that the addition of an abnormal chest radiograph to the above combination of 3 symptom predictors eliminated 11 false negatives (including all 3 false negatives with positive smears), while requiring an additional 56 patients to undergo diagnostic evaluation (**Appendix table 1**). This represents an increase in sensitivity to 97%.

Subgroup analyses

The performance of the 3 predictor combination was heterogeneous in different CD4 subgroups ($p=0.03$). However, while the prevalence of TB was higher in patients with CD4 <200, the 3 predictor combination had higher sensitivity in patients with CD4 <200 than in those with CD4 ≥ 200 (97% and 85%, respectively), making the negative predictive value the same for both groups (97%). Full results of subgroup comparisons are shown in **Appendix table 2**. There was no statistically significant heterogeneity in the performance of the combination across countries ($p=0.36$). The predictive value negative was, not surprisingly, lowest in Vietnam, where prevalence was highest. That the algorithm did well in all settings, i.e. predictive value negative exceeded 90% in all

countries, and sensitivity ranged from 88-98%, supports its potential generalizability. In settings with even higher prevalence, predictive value negative may be lower. At some level of prevalence, routine diagnostic testing may be indicated for all patients. Performance of the combination was not heterogenous with respect to gender ($p=0.22$).

Appendix Table 1. Smear and culture results of patients with TB (N=267), stratified by symptoms and chest radiograph result.

Symptoms*	Category Chest radiograph	Enrolled patients, n	TB diagnosed, n (% of enrolled patients)	Positive acid-fast smear, n (% of TB diagnosed)	Number of positive cultures, n (% of TB diagnosed)	
					1	>1
Absent	Normal	493	7 (1)	0	5 (71)	2 (29)
Present	Normal	865	87 (10)	26 (30)	40 (46)	47 (54)
Absent	Abnormal	56	11 (20)	3 (27)	2 (18)	9 (82)
Present	Abnormal	334	162 (49)	92 (57)	21 (13)	140 (87)

*Any one of: any cough in the past 4 weeks, any fever in the past 4 weeks, or night sweats for ≥ 3 weeks.

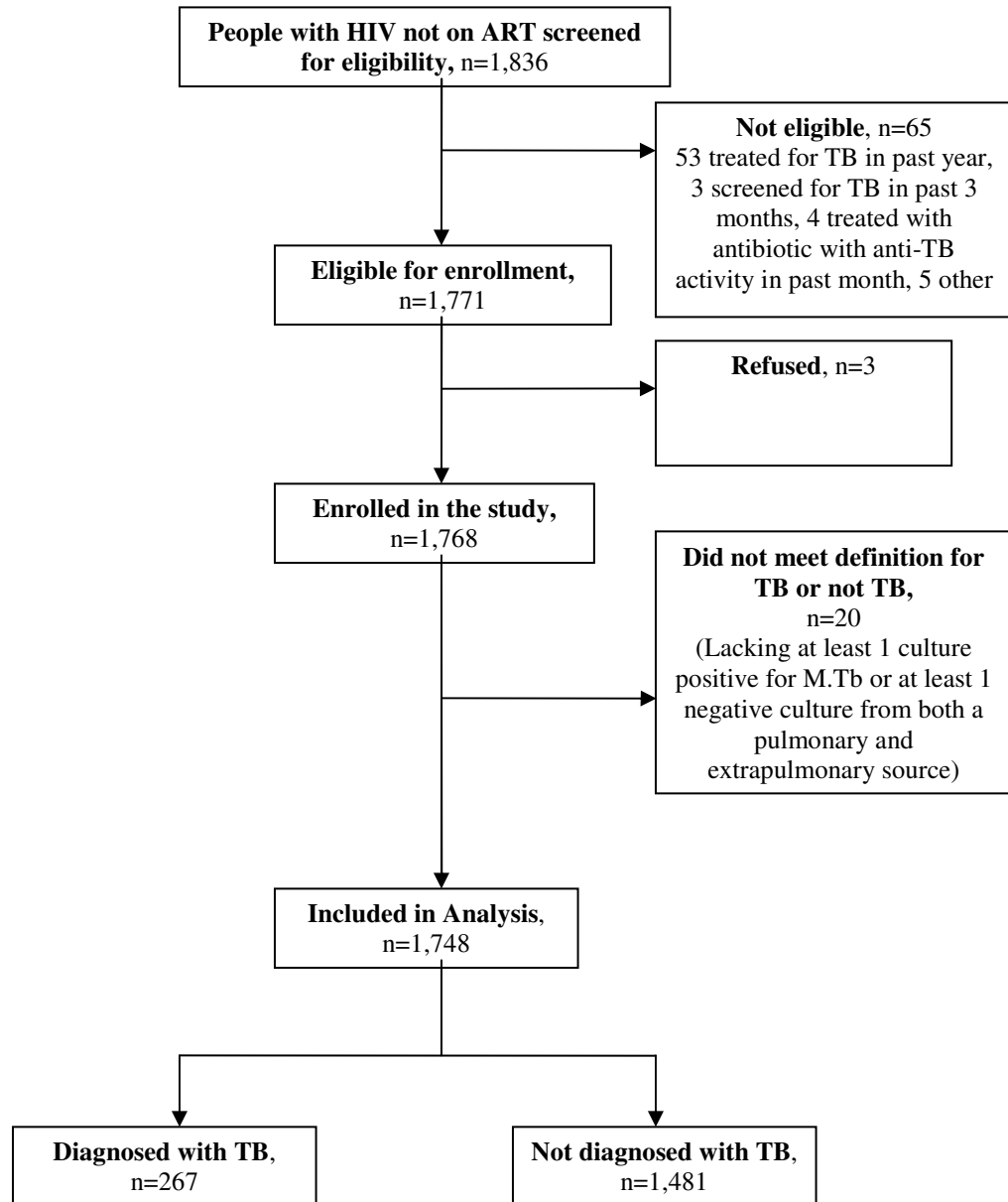
Appendix Table 2. Performance characteristics of proposed symptom screening combination*, stratified by country, CD4, and sex.

	Sens. (%)	Spec. (%)	PVN (%)	PVP (%)	LR [†] negative	LR [†] positive
Country						
Cambodia	98	23	99	18	0.07	1.28
Thailand	90	49	99	10	0.20	1.75
Vietnam	88	47	91	39	0.25	1.67
CD4						
<200	97	26	97	30	0.10	1.32
≥ 200	85	42	97	12	0.37	1.46
Gender						
Male	92	36	95	25	0.23	1.43
Female	96	36	99	16	0.12	1.50

*Any one of: any cough in the past 4 weeks, any fever in the past 4 weeks, or night sweats for ≥ 3 weeks.

† LR negative = (1-sensitivity)/specificity; LR positive = sensitivity / (1-specificity).¹⁵
Small differences in numbers shown may be present due to rounding.

APPENDIX FIGURE: Flow chart of patient enrollment



REFERENCES

1. WHO. Global tuberculosis control - epidemiology, strategy, financing: WHO Report 2009. Geneva, Switzerland: World Health Organization; 2009. Report No.: WHO/HTM/TB/2009.411.
2. Kent PT, Kubica GP, Centers for Disease Control (U.S.). Public health mycobacteriology : a guide for the level III laboratory. Atlanta, Ga.: U.S. Dept. of Health and Human Services, Public Health Service, Centers for Disease Control; 1985.
3. Murray PR, Baron EJ. Manual of clinical microbiology. 9th ed. Washington, D.C.: ASM Press; 2007.
4. Siddiqi SH. MGIT™ Procedure manual for BACTEC™ MGIT 960™ TB system. In. Sparks, MD, USA: Becton Dickinson; 2005.
5. Rieder HL, Van Deun A, Kam KM, et al. Priorities for tuberculosis bacteriology services in low income countries. Paris, France: International Union Against Tuberculosis and Lung Disease; 2007.
6. WHO. Laboratory Services in Tuberculosis Control Part II: Microscopy. Geneva, Switzerland: World Health Organization; 1999.
7. Isenberg HD. Essential procedures for clinical microbiology. Washington, D.C.: ASM Press; 1998.
8. Cowan LS, Diem L, Brake MC, Crawford JT. Transfer of a Mycobacterium tuberculosis genotyping method, Spoligotyping, from a reverse line-blot hybridization, membrane-based assay to the Luminex multianalyte profiling system. J Clin Microbiol 2004;42:474-7.

9. Cowan LS, Diem L, Monson T, et al. Evaluation of a two-step approach for large-scale, prospective genotyping of *Mycobacterium tuberculosis* isolates in the United States. *J Clin Microbiol* 2005;43:688-95.
10. Hastie T, Tibshirani R, Friedman JH. *The elements of statistical learning : data mining, inference, and prediction : with 200 full-color illustrations*. New York: Springer; 2001.
11. R Development Core Team. *R: A language and environment for statistical computing*. In. Vienna, Austria: R Foundation for Statistical Computing; 2008.
12. Zhang H, Singer B. *Recursive Partitioning in the Health Sciences*. New York: Springer-Verlag; 1999.
13. Breiman L. Random Forests. *Machine Learning* 2001;45:5-32.
14. WHO. *Improving the diagnosis and treatment of smear-negative pulmonary and extrapulmonary tuberculosis among adults and adolescents: Recommendations for HIV-prevalent and resource-constrained settings*. Geneva, Switzerland: World Health Organization; 2007. Report No.: WHO/HTM/TB/2007.379.
15. Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios. *Lancet* 2005;365:1500-5.