METHOD ARTICLE

Predicting TB treatment outcomes using baseline risk and treatment response markers: developing the PredictTB early treatment completion criteria [version 1; peer review: 1 approved, 2 not approved]

Ray Y. Chen^{1,2}, Jing Wang³, Lili Liang⁴, Yingda L. Xie^{1,5}, Stephanus T. Malherbe⁶, Jill Winter⁷, Laura E. Via^{1,2}, Xiang Yu¹, Joel Vincent¹, Derek Armstrong⁸, Gerhard Walzl⁶, David Alland⁵, Clifton E. Barry 3rd^{1,2}, Lori E. Dodd⁹

¹Tuberculosis Research Section, Laboratory of Clinical Immunology and Microbiology, Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, 20892, USA

²Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa

³Clinical Monitoring Research Program Directorate, Frederick National Laboratory for Cancer Research, Frederick, MD, USA

⁴Henan Provincial Chest Hospital, Zhengzhou, Henan, China

⁵Department of Medicine and the Public Health Research Institute, Rutgers, New Jersey Medical School, ICPH Building, Room 2232, 225 Warren Street, Newark, NJ, 07103, USA

⁶DST-NRF Centre of Excellence for Biomedical Tuberculosis Research, South African Medical Research Council Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Department of Biomedical Sciences, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa

⁷Catalysis Foundation for Health, 2010 Crow Canyon Pl. STE 100, San Ramon, CA, 94583, USA

⁸Johns Hopkins University School of Medicine, Baltimore, MD, 21231, USA

⁹Biostatistics Research Branch, Division of Clinical Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA

V1 First published: 14 Oct 2020, 4:157
<https://doi.org/10.12688/gatesopenres.13179.1>
 Latest published: 14 Oct 2020, 4:157
<https://doi.org/10.12688/gatesopenres.13179.1>

Abstract

Standard treatment of drug-sensitive pulmonary tuberculosis requires six months of treatment. Several randomized clinical trials have attempted to shorten treatment to four months using various strategies but thus far all have failed. The PredictTB trial is an ongoing international randomized clinical trial testing a treatment shortening strategy whereby only drug-sensitive pulmonary TB patients who meet the study early treatment completion criteria are randomized to four vs. six months of treatment. The PredictTB early treatment completion criteria were developed based on a cohort of 92 pulmonary tuberculosis patients treated programmatically through the local tuberculosis treatment program in Cape Town, South Africa, with FDG-PET/CT scans also performed at baseline and week 4 of treatment. Patients were followed for one year after the end of

Open Peer Review

Approval Status

	1	2	3
version 1
14 Oct 2020	view	view	view
1. Mariza Vorster , University of Pretoria, Pretoria, South Africa			
2. Vineet K. Chadha , National TB Institute, Bengaluru, India			
3. Richard Menzies , McGill University			

therapy for programmatic treatment outcomes. This methodology paper describes how the PET/CT scans and GeneXpert cycle threshold data of this cohort were analyzed to develop the early treatment completion algorithm currently being used in the PredictTB trial.

Keywords

pulmonary tuberculosis, drug sensitive, predict tb, PET/CT, treatment shortening

International TB Centre, Montreal, Canada

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Ray Y. Chen (ray.chen@nih.gov)

Author roles: **Chen RY:** Conceptualization, Data Curation, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Wang J:** Formal Analysis, Methodology, Software, Validation, Visualization, Writing – Review & Editing; **Liang L:** Data Curation, Writing – Review & Editing; **Xie YL:** Data Curation, Methodology, Writing – Review & Editing; **Malherbe ST:** Data Curation, Investigation, Methodology, Writing – Review & Editing; **Winter J:** Funding Acquisition, Project Administration, Resources, Writing – Review & Editing; **Via LE:** Conceptualization, Funding Acquisition, Methodology, Resources, Writing – Review & Editing; **Yu X:** Data Curation, Formal Analysis, Methodology, Software, Writing – Review & Editing; **Vincent J:** Project Administration, Resources, Writing – Review & Editing; **Armstrong D:** Methodology, Resources, Validation, Writing – Review & Editing; **Walzl G:** Conceptualization, Funding Acquisition, Investigation, Resources, Supervision, Writing – Review & Editing; **Alland D:** Conceptualization, Methodology, Resources, Supervision, Writing – Review & Editing; **Barry 3rd CE:** Conceptualization, Data Curation, Funding Acquisition, Methodology, Resources, Supervision, Writing – Review & Editing; **Dodd LE:** Conceptualization, Formal Analysis, Methodology, Supervision, Validation, Visualization, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This study was supported in part by the Bill and Melinda Gates Foundation [OPP51919, OPP1155128], the National Cancer Institute, National Institutes of Health, under Contract No. 75N91019D00024, Task Order No. 75N91019F00130, and the Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health. The content of this publication does not necessarily reflect the views or policies of the U.S. Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2020 Chen RY *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The author(s) is/are employees of the US Government and therefore domestic copyright protection in USA does not apply to this work. The work may be protected under the copyright laws of other jurisdictions when used in those jurisdictions.

How to cite this article: Chen RY, Wang J, Liang L *et al.* **Predicting TB treatment outcomes using baseline risk and treatment response markers: developing the PredictTB early treatment completion criteria [version 1; peer review: 1 approved, 2 not approved]** Gates Open Research 2020, 4:157 <https://doi.org/10.12688/gatesopenres.13179.1>

First published: 14 Oct 2020, 4:157 <https://doi.org/10.12688/gatesopenres.13179.1>

Background

Multiple clinical trials over the last 40 years attempting to shorten therapy for pulmonary, drug-sensitive tuberculosis from six to four months have failed¹⁻⁴. Despite this, the various four-month treatment arms consistently cured about 80–85% of patients. The premise of the ongoing PredictTB clinical trial is that this sub-population of lower risk patients who are cured at four months can be identified prospectively⁵. Two subsequent secondary analyses of the three fluoroquinolone treatment shortening trials²⁻⁴ support this premise by finding that a subset of lower risk participants, those without cavity on baseline chest-x-ray, acid-fast bacilli (AFB) smear less than 2+ at baseline, or AFB smear negative at month 2 of treatment were associated with successful four-month treatment outcomes^{6,7}. Another trial prospectively attempted a similar risk stratification by shortening treatment for those without cavity on baseline chest x-ray, a measure of baseline risk, and those with negative sputum cultures at two months of treatment, a measure of treatment response⁸. Although this trial also failed, the treatment success rate in the four-month arm achieved 93%, higher than other four-month treatment trials that did not risk stratify participants. The PredictTB trial was developed on the hypothesis that more precise methods of evaluating baseline risk and treatment response will successfully identify lower risk participants cured with four months of treatment.

The value of early chest x-ray changes to predict TB treatment outcomes has been recognized for over 60 years⁹. Cavities on baseline chest x-ray¹⁰⁻¹² and residual cavity at cure¹³ have been associated with poor treatment outcomes. Computed tomography scans are more sensitive than x-ray. Further, changes on 2-deoxy-2-[¹⁸F]fluoro-D-glucose (FDG)-positron emission tomography/computed tomography (PET/CT) scans have been associated with treatment outcomes in nonhuman primates^{14,15}. Results in patients have been limited to small numbers due to the difficulty of collecting PET/CT imaging in this setting. However, a study of patients with pulmonary, multidrug-resistant tuberculosis (MDR-TB) showed a relationship with long-term treatment outcomes¹⁶. In drug sensitive TB (DS-TB), distinct response patterns on PET/CT scan from baseline to month 6 on treatment correlated with treatment outcomes¹⁷.

Month 2 sputum culture conversion has traditionally been considered the best biomarker of treatment outcome available¹⁸ despite a meta-analysis demonstrating poor sensitivity and specificity in predicting outcomes¹⁹. This has been confirmed by additional analyses of the REMoxTB treatment shortening trial², where month 2 culture conversion status poorly predicted final outcomes²⁰. Time-delays associated with culture conversion tests can be as long as 6-8 weeks, resulting in delayed assessments of the patient's bacillary status. Immediately available test results are likely to predict outcomes better than delayed results because the results reflect the patient's current status rather than a historical status from 1-2 months ago and thus poor results can be acted upon immediately. The value of contemporaneous results has been shown with CD4 cell counts in HIV²¹, as well as in TB where a month 6 culture conversion

status predicted final outcomes significantly better than month 2 culture conversion status²². A point-of-care test that measures TB bacterial load has the potential to predict outcomes better than a delayed culture conversion result.

The aim of this analysis is to describe how the early treatment completion criteria used in the ongoing PredictTB clinical trial (NIH IRB #16IN133; [NCT02821832](#)) were developed. In this analysis, we assess the relationship between measures of baseline risk and treatment response with poor treatment outcomes in a cohort of pulmonary DS-TB patients treated programmatically in South Africa. Baseline risk is assessed using quantitative markers from FDG-PET/CT scans at baseline. Early treatment response is assessed by changes in these markers on FDG-PET/CT scan at week 4 of treatment. Adherence is also assessed, as is a quantitative measure of residual bacterial load using sputum Xpert MTB/RIF cycle threshold at week 16.

Methods

Dataset

The dataset we used to develop our algorithm were the PET/CT scans (DICOM format), the Xpert MTB/RIF cycle threshold values, and the final treatment outcomes from a cohort of 92 pulmonary DS-TB patients treated programmatically in Cape Town, South Africa with PET/CT scans performed at baseline, week 4, and at the end-of-treatment (Catalysis cohort; end-of-treatment scans not used in this analysis)¹⁷. Ninety-nine participants were included in the original study but seven participants did not have a complete set of PET/CT scans, Xpert MTB/RIF cycle threshold results, and treatment outcomes available. Data were collected under written informed consent and the study was reviewed by the Stellenbosch University ethics committee (approval number N10/01/013). The Catalysis cohort dataset used for this analysis was de-identified and no additional ethical approval was required. We developed a risk stratification algorithm for the PredictTB trial⁵ that predicted participants who would successfully complete TB treatment early at four months (compared to the standard six months) with the following aims: 1) capture all treatment failure and recurrent TB patients as high risk; and 2) stratify 50% of all patients as low-risk. Although previous four-month treatment trials consistently cured 80-85% of patients, we conservatively lowered this estimate to target 50% of patients as eligible for treatment shortening. Patients included in this analysis from the Catalysis cohort were contacted at ≥ 1 year after treatment completion for final treatment outcomes, including cures, treatment failures, and retreatments. Retreatment outcomes were defined programmatically as patients who restarted TB treatment for any reason and may include true relapses, re-infections, or nontuberculous infections with symptoms that mimic TB. Culture confirmation was not routinely obtained nor was mycobacterial genetic strain-typing performed to differentiate relapse from re-infection.

Criteria development and rationale

We patterned our risk stratification algorithm on the Johnson *et al.* trial that, although was stopped early by its Data and

Safety Monitoring Board as a failure, increased the treatment success rate in the four-month arm to 93%⁸. This study used a measure of baseline disease burden (cavity on baseline chest x-ray) as well as a measure of treatment response (sputum culture conversion at week 8). For baseline disease burden, we used disease severity measured on PET/CT scan. For treatment response, we measured change in disease severity on the week 4 PET/CT scan. As we developed the specific risk stratification algorithm thresholds, it became clear that we would not be able to accomplish both aims simultaneously. Any algorithm sensitive enough to capture all unfavorable outcomes as high risk was poorly specific, with well below 50% remaining as low risk. Any algorithm specific enough to capture 50% as low risk was not sensitive enough to capture all or nearly all unfavorable outcomes as high risk. We realized that we could not capture all treatment failure and retreatment patients as high risk in our risk stratification algorithm because some patients who failed may have failed due to patient-related factors, such as poor treatment adherence, that we could not predict because we only had adherence data based on monthly pill counts, which can be inaccurate^{23,24}. For example, a patient with less severe baseline disease and a good response after one month of treatment and therefore predicted to be low risk may subsequently fail treatment due to poor adherence after the initial month of therapy. Retreatment TB patients were even more complex because, in addition to the possibility of poor adherence, retreatment TB was defined as programmatic restart of TB treatment by the local TB clinic for any reason. In most cases, retreated patients were not confirmed as true TB by culture, which also prevented strain typing to differentiate relapse with the same strain from re-infection with a different strain. Re-infection contributed to about 50% of recurrent TB cases in a previous analysis from Cape Town²⁵ and may not be differentiated from relapse in an algorithm based on radiology. A rigorous analysis of risk criteria would have required a larger prospective dataset of cures, treatment failures, and confirmed relapses with PET/CT scans to facilitate model training, testing and validation. Because of these limitations of the Catalysis dataset for our purposes, particularly our lack of clearly defined poor treatment outcomes, we shifted our aims to emphasize: 1) stratifying about 50% of the cohort as low-risk and eligible for treatment shortening, while 2) capturing as many treatment failures over retreatments as possible because failures were felt to be more reliably determined than retreatments.

Measurement of criteria

Each PET/CT scan was read using MIM software version 6 (MIM Software Inc, Cleveland, Ohio USA; freely available alternatives include ITK-SNAP, 3D Slicer, and MeVisLab), with all diseased areas of the lung included in regions of interest (ROI). Data exported from each scan included cavity air measurements and Hounsfield unit (HU) histograms of volumes of each ROI on CT scans and total lesion glycolysis (TLG) for each ROI on PET scans. Hounsfield units are a measure of density, with air about -1000 HU, normal lung around -700 to -950 HU, water at 0 HU, and bone ranging from +500 to +1000 HU. TB lesion density ranges from near normal lung to about +200 HU, with dense consolidations measuring

around -100 to +200 HU. As there are few other densities in this “hard” region in normal lung (for example, blood vessels), the volumes measured in this range almost completely represent TB lesion density so we focused on this “hard” HU range for this analysis (compared to “softer” TB lesions below 100 HU). For PET scans, we determined total lesion glycolysis (TLG) in diseased lung regions, which is a measure of the amount of FDG uptake and is calculated by the mean standardized uptake value in each lung region multiplied by the volume of that region. However, neither CT hard volume nor PET TLG have previously been validated as markers of treatment outcome in TB, whereas previous studies have found chest x-ray cavity size to correlate with unfavorable outcomes^{10–12}. Therefore, for this analysis, cavity air volume measured at baseline and change in cavity air at week 4 were weighted more heavily than CT hard volume and PET TLG.

In addition to PET/CT scan quantitation, two other variables were incorporated into the risk stratification algorithm. Instead of the month 2 sputum culture traditionally used to predict treatment outcome, the Xpert MTB/RIF cycle threshold assay at week 16 was included in our early treatment completion algorithm. Xpert MTB/RIF correlates well with sputum smear and culture results, with excellent sensitivity but poor specificity²⁶. When the assay cycle threshold is incorporated, the balance between sensitivity and specificity improves correlations with sputum smear^{27,28} or culture²⁹. We applied this test at week 16 as a measure of residual bacterial load at the time of potential treatment completion because this is a point-of-care test with immediate results. The second variable incorporated at week 16 is an adherence dose count requirement of about 90% (minimum 100 out of possible 112 doses [7 doses/week x 16 weeks]) because at least 90% has been correlated with better treatment outcomes^{6,30}.

Statistical analysis

Statistical analyses were conducted in R (version 3.6.1). Primary analyses compared tested imaging markers (cavity air, hard volume, and TLG) measured at baseline and one-month after treatment initiation between cures vs failures (and retreatments) using Wilcoxon rank-sum tests. As an exploratory analysis with limited statistical power, statistical significance was defined by $p < 0.05$, without adjustment for multiplicity. Non-parametric receiver operating characteristic (ROC) curves were generated using R packages, pROC and ROCr^{31,32}. Sensitivity and specificity estimates were computed as binomial proportions, along with 95% confidence intervals using a normal approximation.

Results

Among the 92 patients that we analyzed from the Catalysis cohort, 73 were cured (asymptomatic two years after the end of treatment), eight failed treatment, and 11 programmatically restarted TB treatment during follow-up. For the PET/CT imaging analysis, the baseline and week 4 PET/CT scans were each read by a single reader, with overall summary statistics presented in [Table 1](#). At baseline, cured patients were significantly different from treatment failure patients in CT cavity

Table 1. Summary statistics of PET/CT scan read results at baseline and change at week 4. Wilcoxon rank-sum test was performed to assess the difference in image features by outcome groups.

	All (n=92) Median (IQR)	Cure (n=73) Median (IQR)	Failure (n=8) Median (IQR)	Retreatment (n=11) Median (IQR)	P-value (Cure vs. Failure)	P-value (Cure vs. Retreatment)	P-value (Failure vs. Retreatment)
Baseline	Largest cavity air volume (mL)	7.2 (1.7, 17.4)	32.1 (12.9, 58.2)	5.1 (0.6, 7.4)	0.008	0.212	0.004
	Total cavity air volume (mL)	7.2 (1.8, 20.9)	36.4 (19.8, 73.7)	5.1 (0.7, 7.4)	0.004	0.189	0.003
	CT hard volume (mL)	56.5 (32.7, 111.4)	142.3 (67.5, 157.8)	51.1 (33.1, 67.6)	0.059	0.353	0.012
	PET total lesion glycolysis	522.6 (292.2, 995.4)	1127.2 (732.1, 1469.2)	451.9 (349.0, 568.8)	0.059	0.419	0.051
Percent change at week 4	Largest cavity air volume (mL)	-61.8 (-81.0, -32.2)	-39.6 (-53.9, -24.7)	-63.4 (-90.4, -16.3)	0.065	0.851	0.321
	Total cavity air volume (mL)	-61.2 (-78.4, -33.6)	-40.1 (-51.2, -20.3)	-63.4 (-89.0, -16.3)	0.033	0.825	0.167
	CT hard volume (mL)	-16.7 (-31.1, -4.9)	-8.2 (-22.6, -0.4)	-21.4 (-29.6, -6.4)	0.376	0.724	0.310
	PET total lesion glycolysis	-17.7 (-31.3, -2.6)	-8.6 (-21.6, 16.7)	-8.0 (-22.1, -2.0)	0.125	0.308	0.545

PET, positron emission tomography; CT, computed tomography; IQR, interquartile range.

air volume, with CT hard volume and PET TLG differences being borderline significant ($P=0.059$ for both). At week 4, only the difference in total cavity air remained significantly different. In contrast, the results for patients who were cures and retreatments were not significantly different from each other in any parameter at baseline or at week 4, making it very difficult to differentiate these two cohorts using these parameters. The comparison of treatment failures with retreatments was similar to that of treatment failures with cures. We therefore developed our criteria based primarily on differences between the cured and failure cohorts.

To identify specific thresholds that predicted cure vs treatment failure, ROC curves were drawn for each variable (Figure 1). When using the optimal ROC thresholds of all baseline and week 4 PET/CT criteria into a single algorithm, the combined criteria predicted cure with 100% sensitivity, capturing all eight failures and 11 retreatments as high risk. However, specificity was very poor at 19.2% and only 14/81 (17.3%) of subjects with both baseline and week 4 PET/CT scans were classified as low risk (Table 2). This is well below our target of 50%, classifying too many as false positive high risk, resulting in an algorithm that is neither practical nor scalable. We therefore adjusted the thresholds to be more specific at the cost of sensitivity to approach our 50% target.

Because prior clinical trials data already demonstrated that baseline cavity was a risk factor for poor outcomes^{8,10}, and because cavity size was the strongest predictor of poor outcome in our ROC curves (cavity air area-under-the-curve (AUC) > CT hard volume and PET TLG AUCs), we built the algorithm around this parameter first. We defined cavity as largest single cavity size rather than total cavity size because we posited that a single large cavity induced higher risk and may take longer to heal than multiple smaller cavities (i.e. one 30 mL cavity had a higher risk for poor outcome than two 15 mL cavities). In examining the baseline cavity size threshold, a 10.5 mL threshold captured 6/8 (75%) of treatment failures as high risk but only 46/73 (63.0%) cures as low risk. By increasing the threshold to 30 mL, the algorithm would miss one additional failure (now only 5/8 [62.5%] as high risk; increasing beyond 30 mL would lose more than one additional failure) but specificity would increase to 86.3% (63/73 cures now classified as low risk; Table 3). Overall, using 30 mL instead of 10.5 mL as the baseline cavity threshold increases the proportion of cured ($N=73$) and failure ($N=8$) patients defined as low risk from 48/81 (59.3%) to 66/81 (81.5%). Changing the week 4 cavity volume reduction threshold from 64% to 20% results in a similar sensitivity/specificity tradeoff. After applying both baseline and week 4 cavity change thresholds, only 29/81 (35.8%) would have been classified as low risk (before adding

Figure 1. ROC curves of radiological biomarkers in predicting failure vs. cure. Panel A–D: baseline; Panel E–H: %change of week 4 from baseline; AUCs along with 95% CIs were added as blue text at the bottom of each plot; optimal threshold and the corresponding sensitivity and specificity were labeled on the curve. ROC, receiver operating characteristic; AUC, area-under-the-curve; CI, confidence intervals; TLG, total lesion glycolysis.

Table 2. Using optimal AUC cutoffs of imaging features to predict Failure vs. Cure (N=81).

	Sensitivity (95% CI)	Specificity (95% CI)	% of Total Subjects
Baseline largest cavity air <10.5ml	75% (34.9%, 96.8%)	63% (50.9%, 74%)	48/81 (59.3%)
Baseline hard <135.6ml	62.5% (24.5%, 91.5%)	79.5% (68.4%, 88%)	61/81 (75.3%)
Baseline TLG <876.8	62.5% (24.5%, 91.5%)	68.5% (56.6%, 78.9%)	53/81 (65.4%)
Baseline combined criteria	87.5% (47.3%, 99.7%)	53.4% (41.4%, 65.2%)	40/81 (49.4%)
Baseline largest cavity air <10.5ml & %change <-64%	100% (63.1%, 100%)	39.7% (28.5%, 51.9%)	29/81 (35.8%)
Baseline hard <135.6ml & %change <-10.7%	87.5% (47.3%, 99.7%)	53.4% (41.4%, 65.2%)	40/81 (49.4%)
Baseline TLG <876.8 & %change <-10.7%	87.5% (47.3%, 99.7%)	49.3% (37.4%, 61.3%)	37/81 (45.7%)
Final combined criteria	100% (63.1%, 100%)	19.2% (10.9%, 30.1%)	14/81 (17.3%)

AUC, area-under-the-curve; CI, confidence interval; TLG, total lesion glycolysis.

Table 3. Compare different cavity air thresholds to predict Failure vs. Cure (N=81).

	Sensitivity (95% CI)	Specificity (95% CI)	% of Total Subjects
Baseline cavity air <30ml	62.5% (24.5%, 91.5%)	86.3% (76.2%, 93.2%)	66/81 (81.5%)
Baseline cavity air <10.5ml	75% (34.9%, 96.8%)	63% (50.9%, 74%)	48/81 (59.3%)
Baseline cavity air <30ml & %change cavity air <-20% (Predict baseline and one-month cavity criteria)	75% (34.9%, 96.8%)	76.7% (65.4%, 85.8%)	58/81 (71.6%)
Baseline cavity air <30ml & %change cavity air <-64%	100% (63.1%, 100%)	54.8% (42.7%, 66.5%)	40/81 (49.4%)

CI, confidence interval.

any other imaging criteria) using the 64% threshold but 58/81 (71.6%) were low risk with the 20% cavity reduction threshold (Table 3). In contrast to prior data on the risk from baseline cavities, quantitation of CT disease volumes and PET TLG has not previously been validated. Therefore, weighing cavity size to threshold about half (28.4% of total) of the target 50% as high risk seemed appropriate, allowing the remaining criteria (CT hard volume, PET TLG, Xpert cycle threshold, and adherence) to threshold the other half.

Similar to the cavity size thresholds, the cutoffs for hard CT volume and PET TLG were adjusted by decreasing sensitivity but increasing specificity from the optimal ROC parameters (Figure 1, Table 2) to stratify about 50% of the total cohort as low risk, aiming to capture as many treatment failure and retreatment patients as possible within the 50%. Figure 2 demonstrates the patient stratifications when baseline CT hard volume <200 mL and TLG <1500 units were used as low risk criteria at baseline. When combined with largest cavity air <30 mL, 60/81 (74.1%) patients were classified as low risk at baseline (Table 4). When applied to the entire Catalysis cohort, including the retreatment patients, 6/8 (75%) failures but only 1/11 retreatments were classified as high risk at baseline, suggesting that treatment failure patients may be more correlated with severity of baseline disease compared to

retreatment TB patients. Week 4 change criteria for CT hard volume and PET TLG were similarly adjusted and allowed for slight increases to account for potential paradoxical treatment responses that were ultimately still favorable. The final week 4 criteria selected allowed up to a 10% increase in hard volume and 30% increase in TLG at week 4 to remain low risk, resulting in 46/81 (56.8%) classified as low risk after applying both baseline and week 4 PET/CT criteria (Table 4). The week 4 criteria captured only one additional failure but two additional retreatments, suggesting that retreatment TB patients may be more correlated with poor treatment response at week 4 rather than severity of disease at baseline.

Finally, we included a measure of residual TB bacterial load in sputum in the early treatment completion criteria based on an analysis that Xpert cycle threshold around 30 at weeks 8 and 24 correlated with culture negativity and patient treatment outcomes²⁹. We incorporated this measure at week 16 as a safety mechanism to ensure that participants with higher sputum bacterial load (cycle threshold <30) did not stop treatment early. Combining the baseline and week 4 criteria stratified 41/81 (50.6%) of all cured and failure patients as low risk (Table 4). When applied to the entire Catalysis cohort, including retreatment patients, 1/8 (12.5%) failures and 6/11 (54.5%) retreatments were captured as low risk. The PredictTB early

Figure 2. Hard volume vs. total lesion glycolysis (TLG). Left: baseline; Right: % change at four weeks from baseline. Left plot: Six failures and one retreated were caught by baseline criteria (five failures and one retreated have cavity air ≥ 30 and two have $TLG > 1500$); Right plot: Two retreated subjects were selected by week 4 cavity air criteria (decrease of cavity air $< 20\%$). Two additional retreated cases and one failure were caught by Week 16 Xpert < 30 .

Table 4. Radiological markers to predict Cure vs Failure (N=81).

	Sensitivity (95% CI)	Specificity (95% CI)	% of Total Subjects
Baseline cavity criteria	62.5% (24.5%, 91.5%)	86.3% (76.2%, 93.2%)	66/81 (81.5%)
Baseline hard criteria	0% (0%, 36.9%)	91.8% (83%, 96.9%)	75/81 (92.6%)
Baseline TLG criteria	25% (3.2%, 65.1%)	93.2% (84.7%, 97.7%)	74/81 (91.4%)
Baseline Predict criteria	75% (34.9%, 96.8%)	79.5% (68.4%, 88%)	60/81 (74.1%)
Baseline and one-month cavity criteria	75% (34.9%, 96.8%)	76.7% (65.4%, 85.8%)	58/81 (71.6%)
Baseline and one-month hard criteria	25% (3.2%, 65.1%)	79.5% (68.4%, 88%)	64/81 (79%)
Baseline and one-month TLG criteria	37.5% (8.5%, 75.5%)	84.9% (74.6%, 92.2%)	67/81 (82.7%)
Final Predict Radiological criteria	87.5% (47.3%, 99.7%)	61.6% (49.5%, 72.8%)	46/81 (56.8%)
Final Predict criteria (+Week16 Xpert)	87.5% (47.3%, 99.7%)	54.8% (42.7%, 66.5%)	41/81 (50.6%)

CI, confidence interval; TLG, total lesion glycolysis.

treatment completion criteria when the trial started is shown in Table 5a.

The initial criteria were established as described above, acknowledging that early changes may be needed once the trial began if the actual proportion of PredictTB study participants stratified to the low- and high-risk arms were not close to the 50:50

target. Indeed, after about nine months of enrollment, only 23.4% of participants to reach week 16 were stratified as low risk (Arms B and C), with the remainder stratified to the high-risk arm (Arm A). This was less than half of the estimated 50% we expected to be low risk, which had major implications for the cost and duration of the study (total sample size, study duration, and cost would need to be increased to

Table 5a. Predict TB early treatment completion criteria at the start of the trial.

Early completion criteria:	Determined at Week 16 – unless known to have failed a radiologic criterion at baseline or week 4.
Radiologic criteria	Baseline PET/CT: <ul style="list-style-type: none"> • No total lung collapse of a single side, AND • No pleural effusion, AND • No single cavity air volume on CT scan >30 mL, AND • CT scan hard volume (-100 to +100 HU density) <200 mL, AND • PET total lesion glycolysis <1500 units Week 4 PET/CT: <ul style="list-style-type: none"> • All individual cavities decrease by >20% (unless cavity <2 mL), AND • CT scan hard volume does not increase by >10% unless the increase is <5 mL, AND • PET total lesion glycolysis does not increase by >30% unless the increase is <50 units
Bacterial load criterion	Week 16 Xpert cycle threshold $\geq 30^*$
Adherence criterion	Minimum of 100 doses received by week 16

*If the week 16 solid medium sputum culture is subsequently found to be positive for *Mtb* in a participant randomized to Arm B or C, this participant will be called in for evaluation and to provide sputum for a repeat culture. If the initial positive culture is confirmed by a second culture positive for *Mtb*, this participant will be considered to have met the study endpoint as a treatment failure and will be referred for continued treatment.

TB, tuberculosis; PET, positron emission tomography; CT, computed tomography; HU, Hounsfield unit.

achieve required sample size in Arms B and C) as well as the scientific relevance of the study (if trial successful, it would only apply to less than 25% of TB patients, diminishing relevance). After discussion with our study Data and Safety Monitoring Board (DSMB), we revised our study early treatment completion criteria.

Revising the early treatment completion criteria

We considered how to change both the Xpert cycle threshold cutoff and the PET/CT radiology thresholds. For the Xpert cycle threshold, the original cutoff at week 16 was based on a cohort study in South Africa with MGIT culture results, the only data available to us at the time. We adopted a stringent cycle threshold value of 30 based upon analysis of these data for subjects to be randomized to Arms B and C. Xpert detects bacterial DNA but does not determine the viability of detected DNA (i.e., detected bacteria may be dead). For PredictTB, however, LJ culture is used to determine primary study outcomes. For this re-analysis, we received unpublished results from TBTC study 29, which collected cycle threshold values and LJ culture results (Rada Savic, personal communication). In evaluating the change, we considered the chance of missing an LJ+ result, as well as the sensitivity and specificity of various cycle threshold cutoffs. In contrast to positive and negative predictive values, sensitivity and specificity do not depend on the underlying proportion of culture positive results, which varies over time and from study to study. That said, patient safety was a driving factor so we considered how many positive cultures might be missed for various cutoffs. This was defined as the probability of being LJ+ given a Xpert cycle threshold value less than the cutoff, i.e., $P(\text{LJ+} \mid \text{Ct-})$. We assumed what we considered were high proportions of LJ+ cultures (i.e., 10% and 5% at week 16 of treatment in the lower risk cohort of arm B/C) when making this decision.

In contrast to TBTC study 29, which randomized all-comers and did not stratify participants by risk, the PredictTB study further excludes poorly adherent participants, those with too severe disease at baseline, or those not responding appropriately to treatment at one month. As a result, the expected LJ+ rates of 10% and 5% were considered to be very high. Table 6 describes these proportions for the sensitivity and specificity estimates from TBTC study 29. Based on these estimates, a cycle threshold of 30 was expected to miss 2.1% of LJ+ results, while a threshold of 28 would miss 2.5%, assuming a 10% LJ+ rate. This translated to an increase in less than one participant being missed among those randomized to arm C. That is, if the background LJ+ rate was 10%, 3.3 (of 155 randomized to stop treatment at week 16) true LJ+ participants may be missed with a cycle threshold of 30, and 3.9 may be missed with a cycle threshold of 28. If the underlying LJ+ rate was 5%, this becomes 1.6 missed LJ+ participants with cycle threshold 30 and 1.9 missed LJ+ participants with cycle threshold 28. If the true underlying LJ+ rate was even lower (as we would expect it to be), the difference between the two cycle threshold values becomes even smaller. Of the 12 participants already enrolled in the PredictTB study with week 16 Xpert cycle threshold results at the time of this analysis, eight had negative results, two had cycle thresholds below 28 (18.2 and 25.2) and two had cycle thresholds above 28 (28.4 and 28.5). Thus changing the Xpert cycle threshold cutoff from 30 to 28 would potentially (depending on radiology criteria) have retained an additional two participants in arms B and C.

To further correct the arm imbalance, we also changed the baseline and week 4 radiologic criteria. Prior studies validated that cavity on baseline CXR is a risk factor for treatment relapse. In our analyses of prior data, cavity size was also

the strongest factor in predicting poor treatment outcome so we did not adjust this criterion. The data for CT hard volume and PET TLG as risk factors for poor treatment outcomes, however, were weak. [Figure 3a](#) shows the distribution of participants stratified to Arm A at baseline by the original radiology criteria. The numbers in the circles represent the number of participants that fell into arm A according to the defined criteria. The hard volume and total activity criteria were relatively well correlated in capturing participants, with only five participants moved to Arm A based on a single criterion, hard volume or PET TLG. Therefore, instead of arbitrarily increasing the hard volume and PET TLG cutoffs, we changed the

criteria from requiring both hard volume **AND** total activity to be below the thresholds to be considered low risk, to only requiring one criterion. That is, participants with either hard volume **OR** total activity below the threshold at both baseline and week 4 would be considered low risk. The thresholds themselves did not change. Applying this change to the PET/CT criteria results in the revised Venn diagram in [Figure 3b](#), which is the same as [Figure 3a](#) except for the five participants moved to Arm A based on hard volume or PET activity alone are no longer considered high risk. The revised early treatment completion criteria incorporating both Xpert cycle threshold and radiologic criteria changes are shown in [Table 5b](#).

Table 5b. Amended Predict TB early treatment completion criteria to correct arm imbalance between Arm A and Arms B/C. Changes are highlighted in yellow.

Early completion criteria:	Determined at Week 16 – unless known to have failed a radiologic criterion at baseline or week 4.
Radiologic criteria	<p>Baseline PET/CT:</p> <ul style="list-style-type: none"> No total lung collapse of a single side, AND No pleural effusion, AND No single cavity air volume on CT scan >30 mL, AND CT scan hard volume (-100 to +100 HU density) <200 mL OR PET total lesion glycolysis <1500 units <p>Week 4 PET/CT:</p> <ul style="list-style-type: none"> All individual cavities decrease by >20% (unless cavity <2 mL), AND CT scan hard volume does not increase by >10% unless the increase is <5 mL OR PET total lesion glycolysis does not increase by >30% unless the increase is <50 units
Bacterial load criterion	Week 16 Xpert cycle threshold $\geq$ 28 *
Adherence criterion	Minimum of 100 doses received by week 16

*If the week 16 solid medium sputum culture is subsequently found to be positive for *Mtb* in a participant randomized to Arm B or C, this participant will be called in for evaluation and to provide sputum for a repeat culture. If the initial positive culture is confirmed by a second culture positive for *Mtb*, this participant will be considered to have met the study endpoint as a treatment failure and will be referred for continued treatment.

TB, tuberculosis; PET, positron emission tomography; CT, computed tomography; HU, Hounsfield unit.

Table 6. Sensitivity and specificity estimates from TBTC study 29 for various Xpert cycle threshold cutoffs, along with estimates of missed LJ+ and missed LJ- results for assumed (week 16) culture-positivity rates of 10% and 5%.

Xpert cycle threshold Cutoff	Sensitivity: P(Ct<c LJ+)	Specificity: P(Ct>c LJ-)	Chance of missed LJ+ P(LJ+ Ct-) with 10% LJ+ rate	Chance of missed LJ+ P(LJ+ Ct-) with 5% LJ+ rate	Chance of missed LJ+ P(LJ+ Ct-) with 2.5% LJ+ rate
31	0.93	0.43	0.018	0.008	0.004
30	0.91	0.46	0.021	0.010	0.005
29	0.89	0.49	0.024	0.012	0.006
28	0.88	0.52	0.025	0.012	0.006
27	0.86	0.55	0.028	0.013	0.006
26	0.84	0.60	0.029	0.014	0.007
25	0.79	0.66	0.034	0.016	0.008

Figure 3a. Venn diagram of the original baseline PET/CT criteria by which participants were stratified to Arm A.

Figure 3b. Venn diagram of the *revised* baseline PET/CT criteria by which participants are stratified to Arm A.

The revised early treatment completion criteria were accepted by the NIAID DSMB on March 16, 2018 and implemented after local regulatory approvals at the Henan, China sites on May 19, 2018 and at the Western Cape, South Africa sites on June 15, 2018. Only 12 (3.9%) of the total sample size of 310 to the low risk arms were recruited under the original early treatment completion criteria. The revised criteria re-balanced the arm proportions very nicely, approaching 50:50 in Arms A and B/C. The data used to develop both the original and revised early treatment completion criteria are deposited on [Harvard Dataverse](#) (see [Data availability](#)).

Discussion

Previously conducted treatment shortening studies for DS-TB suggested that approximately 80–85% of patients are cured with four months of treatment^{1–4}. Shortening treatment only in lower risk participants who had no cavity on baseline chest x-ray and had sputum culture converted to negative by two months of treatment resulted in the four-month treatment success proportion increasing to 93% in one trial but this was still significantly worse than six months of treatment⁸. The PredictTB trial tests an alternate risk stratification criteria based on FDG-PET/CT disease burden at baseline, the change in PET/CT disease burden at week 4 of treatment, and a marker of residual bacterial load and adherence dose count at the end of treatment, hypothesizing that this combination will identify patients with tuberculosis who are cured with four months of standard treatment⁵. Risk signatures based on transcriptomics have recently been shown to correlate with treatment outcomes^{33,34}.

The development of the PredictTB early treatment completion criteria was based on a cohort of 92 DS-TB patients programmatically treated in Cape Town, South Africa on whom we had PET/CT scans at baseline and week 4 of treatment, Xpert cycle threshold data, and programmatic treatment outcomes (Catalysis cohort). Because these patients were treated programmatically, treatment was not directly observed and we were thus unable to determine the proportion of treatment failures due to poor adherence or differentiate true relapsed disease patients from those re-infected. The lack of these data confounded our attempts to develop early treatment completion criteria that captured treatment failure and true relapse patients with meaningful sensitivity and specificity. We resorted to developing criteria that stratified about 50% of patients as high risk, trusting that the most severely diseased patients at baseline, those with poor treatment responses at week 4, and those under the Xpert cycle threshold cutoff at week 16 were

captured as higher risk and therefore not eligible for treatment shortening.

A major limitation of developing our algorithm was the lack of sufficient relapse data to validate our early treatment completion criteria. This limitation is challenging to overcome, given the small numbers of available patient data anywhere with microbiological, strain-type confirmed relapses and the even smaller numbers of these with FDG-PET/CT scan data. We acknowledge the risk of overfitting our data (i.e., producing a risk model that may not be generalizable because it was fit only to the data on which it was developed) and in fact, after the study started, it became clear that our criteria were too conservative, stratifying >75% of participants as high risk and therefore not eligible for treatment shortening. Without immediate correction, we would likely run out of funding before the end of the trial due to the increased total sample size needed to achieve the required lower risk cohort sample size. Even worse, we would end up with a trial result that was applicable only to the 20–25% of patients stratified to the lower risk arms and therefore not relevant to the majority TB patients. Our amended early treatment completion criteria, however, have been stratifying participants at roughly 50:50 to the high- vs. low-risk arms.

The PredictTB trial early treatment completion criteria were developed to identify those with the most severe disease at baseline (potentially at higher risk for treatment failure) and with a poor week 4 treatment response (potentially at higher risk of relapse), along with a marker of residual bacterial load and an adherence dose count at treatment completion. These criteria are currently stratifying about 50% of patients to the higher risk arm and 50% of patients to the two lower risk arms, which is the target goal. Whether or not this will successfully identify a lower risk cohort that can be successfully cured with four months of standard therapy awaits the results of the trial, expected in 2022.

Data availability

Underlying data

Harvard Dataverse: Replication Data for PredictTB Early Treatment Completion Criteria. <https://doi.org/10.7910/DVN/97HYQ5>³⁵.

Data are available under the terms of the [Creative Commons Zero “No rights reserved” data waiver](#) (CC0 1.0 Public domain dedication).

References

1. Fox W: **Whither short-course chemotherapy?** *Br J Dis Chest*. 1981; **75**(4): 331–57.
[PubMed Abstract](#) | [Publisher Full Text](#)
2. Gillespie SH, Crook AM, McHugh TD, *et al.*: **Four-month moxifloxacin-based regimens for drug-sensitive tuberculosis.** *N Engl J Med*. 2014; **371**(17): 1577–87.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. Jindani A, Harrison TS, Nunn AJ, *et al.*: **High-Dose Rifapentine with Moxifloxacin for Pulmonary Tuberculosis.** *N Engl J Med*. 2014; **371**(17): 1599–608.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Merle CS, Fielding K, Sow OB, *et al.*: **A Four-Month Gatifloxacin-Containing Regimen for Treating Tuberculosis.** *N Engl J Med*. 2014; **371**(17): 1588–98.
[PubMed Abstract](#) | [Publisher Full Text](#)
5. Chen RY, Via LE, Dodd LE, *et al.*: **Using biomarkers to predict TB treatment duration (Predict TB): a prospective, randomized, noninferiority, treatment**

- shortening clinical trial. *Gates Open Res.* 2017; 1: 9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Imperiali MZ, Nahid P, Phillips PPJ, *et al.*: A patient-level pooled analysis of treatment-shortening regimens for drug-susceptible pulmonary tuberculosis. *Nat Med.* 2018; 24(11): 1708–15.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 7. Romanowski K, Balshaw RF, Benedetti A, *et al.*: Predicting tuberculosis relapse in patients treated with the standard 6-month regimen: an individual patient data meta-analysis. *Thorax.* 2019; 74(3): 291–297.
[PubMed Abstract](#) | [Publisher Full Text](#)
 8. Johnson JL, Hadad DJ, Dietze R, *et al.*: Shortening treatment in adults with noncavitary tuberculosis and 2-month culture conversion. *Am J Respir Crit Care Med.* 2009; 180(6): 558–63.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 9. Fox W, Sutherland I: A five-year assessment of patients in a controlled trial of streptomycin, para-aminosalicylic acid, and streptomycin plus para-aminosalicylic acid, in pulmonary tuberculosis. *Q J Med.* 1956; 25(98): 221–43.
[PubMed Abstract](#)
 10. Benator D, Bhattacharya M, Bozeman L, *et al.*: Rifapentine and isoniazid once a week versus rifampicin and isoniazid twice a week for treatment of drug-susceptible pulmonary tuberculosis in HIV-negative patients: a randomised clinical trial. *Lancet.* 2002; 360(9332): 528–34.
[PubMed Abstract](#) | [Publisher Full Text](#)
 11. Nettles RE, Mazo D, Alwood K, *et al.*: Risk factors for relapse and acquired rifamycin resistance after directly observed tuberculosis treatment: a comparison by HIV serostatus and rifamycin use. *Clin Infect Dis.* 2004; 38(5): 731–6.
[PubMed Abstract](#) | [Publisher Full Text](#)
 12. Yew WW, Chan CK, Chau CH, *et al.*: Outcomes of patients with multidrug-resistant pulmonary tuberculosis treated with ofloxacin/levofloxacin-containing regimens. *Chest.* 2000; 117(3): 744–51.
[PubMed Abstract](#) | [Publisher Full Text](#)
 13. Sonnenberg P, Murray J, Glynn JR, *et al.*: HIV-1 and recurrence, relapse, and reinfection of tuberculosis after cure: a cohort study in South African mineworkers. *Lancet.* 2001; 358(9294): 1687–93.
[PubMed Abstract](#) | [Publisher Full Text](#)
 14. Coleman MT, Maiello P, Tomko J, *et al.*: Early Changes by (18)Fluorodeoxyglucose positron emission tomography coregistered with computed tomography predict outcome after Mycobacterium tuberculosis infection in cynomolgus macaques. *Infect Immun.* 2014; 82(6): 2400–4.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 15. Lin PL, Coleman T, Carney JP, *et al.*: Radiologic responses in cynomolgus macaques for assessing tuberculosis chemotherapy regimens. *Antimicrob Agents Chemother.* 2013; 57(9): 4237–4244.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 16. Chen RY, Dodd LE, Lee M, *et al.*: PET/CT imaging correlates with treatment outcome in patients with multidrug-resistant tuberculosis. *Sci Transl Med.* 2014; 6(265): 265ra166.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 17. Malherbe ST, Shenai S, Ronacher K, *et al.*: Persisting positron emission tomography lesion activity and Mycobacterium tuberculosis mRNA after tuberculosis cure. *Nat Med.* 2016; 22(10): 1094–1100.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 18. Wallis RS, Kim P, Cole S, *et al.*: Tuberculosis biomarkers discovery: developments, needs, and challenges. *Lancet Infect Dis.* 2013; 13(4): 362–72.
[PubMed Abstract](#) | [Publisher Full Text](#)
 19. Horne DJ, Royce SE, Gooze L, *et al.*: Sputum monitoring during tuberculosis treatment for predicting outcome: systematic review and meta-analysis. *Lancet Infect Dis.* 2010; 10(6): 387–94.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 20. Phillips PP, Mendel CM, Burger DA, *et al.*: Limited role of culture conversion for decision-making in individual patient care and for advancing novel regimens to confirmatory clinical trials. *BMC Med.* 2016; 14: 19.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 21. Brennan AT, Maskew M, Sanne I, *et al.*: The interplay between CD4 cell count, viral load suppression and duration of antiretroviral therapy on mortality in a resource-limited setting. *Trop Med Int Health.* 2013; 18(5): 619–31.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 22. Kurbatova EV, Cegielski JP, Lienhardt C, *et al.*: Sputum culture conversion as a prognostic marker for end-of-treatment outcome in patients with multidrug-resistant tuberculosis: a secondary analysis of data from two observational cohort studies. *Lancet Respir Med.* 2015; 3(3): 201–9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 23. El Allili M, Vrijens B, Demonceau J, *et al.*: A scoping review of studies comparing the medication event monitoring system (MEMS) with alternative methods for measuring medication adherence. *Br J Clin Pharmacol.* 2016; 82(1): 268–79.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 24. Valencia S, León M, Losada I, *et al.*: How do we measure adherence to anti-tuberculosis treatment? *Expert Rev Anti Infect Ther.* 2017; 15(2): 157–65.
[PubMed Abstract](#) | [Publisher Full Text](#)
 25. Marx FM, Dunbar R, Enarson DA, *et al.*: The temporal dynamics of relapse and reinfection tuberculosis after successful treatment: a retrospective cohort study. *Clin Infect Dis.* 2014; 58(12): 1676–83.
[PubMed Abstract](#) | [Publisher Full Text](#)
 26. Friedrich SO, Rachow A, Saathoff E, *et al.*: Assessment of the sensitivity and specificity of Xpert MTB/RIF assay as an early sputum biomarker of response to tuberculosis treatment. *Lancet Respir Med.* 2013; 1(6): 462–70.
[PubMed Abstract](#) | [Publisher Full Text](#)
 27. Lange B, Khan P, Kalmambetova G, *et al.*: Diagnostic accuracy of the Xpert® MTB/RIF cycle threshold level to predict smear positivity: a meta-analysis. *Int J Tuberc Lung Dis.* 2017; 21(5): 493–502.
[PubMed Abstract](#) | [Publisher Full Text](#)
 28. Beynon F, Theron G, Respeito D, *et al.*: Correlation of Xpert MTB/RIF with measures to assess Mycobacterium tuberculosis bacillary burden in high HIV burden areas of Southern Africa. *Sci Rep.* 2018; 8(1): 5201.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 29. Shenai S, Ronacher K, Malherbe S, *et al.*: Bacterial Loads Measured by the Xpert MTB/RIF Assay as Markers of Culture Conversion and Bacteriological Cure in Pulmonary TB. *PLoS One.* 2016; 11(8): e0160062.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 30. Podewils LJ, Gler MT, Quelapio MI, *et al.*: Patterns of treatment interruption among patients with multidrug-resistant TB (MDR TB) and association with interim and final treatment outcomes. *PLoS One.* 2013; 8(7): e70064.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 31. Robin X, Turck N, Hainard A, *et al.*: pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics.* 2011; 12: 77.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 32. Sing T, Sander O, Beerenwinkel N, *et al.*: ROCR: visualizing classifier performance in R. *Bioinformatics.* 2005; 21(20): 3940–1.
[PubMed Abstract](#) | [Publisher Full Text](#)
 33. Thompson EG, Du Y, Malherbe ST, *et al.*: Host blood RNA signatures predict the outcome of tuberculosis treatment. *Tuberculosis (Edinb).* 2017; 107: 48–58.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 34. Penn-Nicholson A, Mbandi SK, Thompson E, *et al.*: RISK6, a 6-gene transcriptomic signature of TB disease risk, diagnosis and treatment response. *Sci Rep.* 2020; 10(1): 8629.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 35. Chen R: “Replication Data for PredictTB Early Treatment Completion Criteria”. Harvard Dataverse, V1, UNF: 6:wckzF/sNge+t4N9AnwnfpA== [fileUNF]. 2020.
<http://www.doi.org/10.7910/DVN/97HYQ5>

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 17 February 2021

<https://doi.org/10.21956/gatesopenres.14379.r29920>

© 2021 Menzies R. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Richard Menzies

McGill University International TB Centre, Montreal, QC, Canada

This is an interesting paper in which the authors describe development of risk stratification that they then apply to an ongoing randomized trial. The fundamental objective is to predict the subgroup of patients with active TB with high treatment success rates with shortened therapy, i.e. 4 months of therapy, and identifying those who require longer treatment.

Major comments:

The problem in treatment of drug-sensitive TB is the relatively high relapse rate with standard 6-month treatment, not failure. In most systematic reviews, the failure rate is less than 1%, but the relapse rate is around 4% or 5% in patients with DS-TB treated with the standard therapy of INH, Rifampin, PZA and Ethambutol for 2 months followed by 4 months of INH and Rifampin (2HRZE/4HR). And in unselected patients receiving 4 months - the relapse rate is much higher (10-15%). A number of risk factors for relapse have been identified: intermittent therapy, particularly in the first 2 months of therapy, cavities (on plain chest radiograph) present at initiation of treatment, or at 2 months or at 5 months, smear positivity at baseline or at 2 months, and culture positivity at 2 months. All of these risk factors predict higher rates of relapse; currently therapy in patients with these risk factors is prolonged until 9 months. One would imagine that the counterfactual would be true, i.e. patients without these risk factors should benefit from shorter treatment, but this remains a hypothesis that has not yet been borne out in trials.

There are two major problems with the objectives of this study: (1) The investigators are assessing different information in an attempt to refine the prediction of relapse. This is good, since prediction of relapse is imperfect at present. BUT, they have not made use of the available prediction tools that are already known, but rather have used ONLY the different prediction tools - findings on CT scan and PET reactivity. So we do not know if CT or PET findings actually improve accuracy of prediction when added to the microbiologic and radiographic information that is currently available to clinicians in all settings. (2) They focused on treatment failure rather than treatment relapse. The authors justified this because relapse was more difficult to diagnose under programmatic conditions. But unfortunately this does not change the fact that an analysis of

treatment failure is simply looking at the wrong outcome for treatment shortening. The study should remain focused on the key problem for TB treatment shortening - which is relapse after the end of treatment.

Specific comments:

There are a number of points of information that would be helpful in order to be able to better interpret this paper. This information includes:

1. Patients had drug-sensitive TB, but how many of them had prior treatment? This is an important variable to consider, because it may predict adherence behaviour but also because there is some evidence that previously treated patients may be at greater risk of failure because of some degree of drug resistance, which is not always evident with phenotypic DST.
2. Were all isolates from these patients tested for all first-line agents or only for Rifampin? If only for Rifampin, they could miss important resistance to other first-line drugs which would substantially increase the risk of both failure and relapse, but also make the results of less interest because treatment shortening should not be considered if there is any first-line drug resistance.
3. What was the regimen given? I'm assuming it was 2HRZE/4HR but this should be specified. Also, was it daily or intermittent? Was DOT used and, if so, how was DOT given? Were patients hospitalized for any significant time at the beginning of therapy?
4. Patient characteristics are missing. In particular, we need to know if they were HIV-infected and, if so, if they were on effective ART.
5. Adherence – the authors mentioned that pill counts are done, which they acknowledge has limitations. Again, this raises the issue of DOT and why DOT was not used? If they were relying on pill counts, what was the percent adherence based on pill counts? This would be important particularly to look at the difference in those who were cured, failed or had retreatment.
6. Failure – very high failure rates are noted here. 8 out of 92 failed which is 9%. As noted earlier, generally drug-sensitive patients with standard regimen have less than 1% failure rate. Why so many failures?
7. Was there evidence of acquired drug resistance? Could this have been in fact re-infection or new infection with drug-resistant strains?
8. Relapse – also very high. 13%, or 11 out of 84, relapsed compared to usually 4-5% in most series. I understand they were not all culture confirmed, but how many were culture confirmed, and of these, how many have acquired drug resistance or evidence of re-infection? Even the basic information of whether they were HIV-infected or not would be helpful in order to better interpret this information.

Recommendations:

1. First provide the missing information as described above.

2. Some explanation as to the high failure and relapse rates would be important.
3. In Table 1, add microbiologic findings, particularly smear and culture status at baseline or pre-treatment and at 2 months. Also add plain chest radiography results so that we can see whether CT or PET findings are adding to these other known predictors.
4. Incorporate the known predictive findings (plain chest Xray, smear and culture) into all predictive algorithms. That should be the starting point. Then assess whether adding CT and PET findings improves the prediction.
5. PET findings, I think it would be helpful to have a simple summary of the PET findings at baseline and if the repeat simply is improving or not, or whether the PET has improved in some places but worsened in others. These qualitative interpretations would be helpful because that is the info in reports that clinicians receive, and would give a better sense of whether PET scan findings are in fact predictive.

Is the rationale for developing the new method (or application) clearly explained?

Partly

Is the description of the method technically sound?

No

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Clinical trials, Epidemiology - in TB

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 16 February 2021

<https://doi.org/10.21956/gatesopenres.14379.r30223>

© 2021 Chadha V. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Vineet K. Chadha

National TB Institute, Bengaluru, Karnataka, India

In the 'Background' section, the authors as a justification for their analysis have referred to high treatment success rates with shorter drug regimen but have conveniently ignored the higher relapse/recurrence rates with these regimens as observed during the same studies as well as in many other studies. The Predict TB trial wherein patients were stratified into shorter and standard duration therapy based on the criteria as proposed by the authors in the current manuscript has already been discontinued for the same reason of significantly higher rates of recurrence in the shorter regimen arm. This is a further testimony to the poor justification for the analysis as well as arbitrariness in the repeated mathematical jugglery to reach 50% allocation target.

The small sample size of 81 patients used for the present analysis is another issue. Besides, the method used by the authors to ascertain recurrence based on symptomatic status after 2 years without any systematic effort by regular screening and sputum examination during the post treatment period also propagates a wrong method for ascertaining recurrence. I don't find any novelty or learning for other researchers in adopting their method in any of their endeavors.

The manuscript should therefore not pass peer-review.

Is the rationale for developing the new method (or application) clearly explained?

Partly

Is the description of the method technically sound?

No

Are sufficient details provided to allow replication of the method development and its use by others?

No

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Epidemiology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 25 January 2021

<https://doi.org/10.21956/gatesopenres.14379.r30195>

© 2021 Vorster M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Mariza Vorster

Department of Nuclear Medicine, University of Pretoria, Pretoria, South Africa

In this paper the authors set out to explain how image findings on ^{18}F -FDG PET/CT and GeneXpert cycle threshold data have been analyzed in order to develop the early treatment completion algorithm that is currently applied in the ongoing PredictTB trial. The PredictTB trial aims to prospectively identify the sub-group of patients that are at a lower risk and who could successfully complete treatment at 4 months.

^{18}F -FDG PET/CT was acquired at baseline and again at week 4 of the treatment regimen for the 92 patients with pulmonary TB. Patient adherence and residual TB bacterial load was assessed by means of Sputum Xpert MTB/RIF cycle threshold at week 16. Patients included were followed up for a period of at least one year to assess outcomes. Measurements included determination of baseline disease burden as well as treatment response evaluation.

Semi-quantitative parameters included Total Lesion Glycolysis (TLG) on PET and Hounsfield units and air volumes on CT. The software used and other aspects of methodology were described in sufficient detail to allow replication. Details on the statistical analysis are adequate.

The results indicated that 73 patients (out of 92 included) were asymptomatic at 2 years after completion of therapy, with 8 treatment failures and 11 retreated patients. Baseline imaging parameters demonstrated statistically significant differences between cure and treatment failure and between failure vs re-treatment with regards to the largest cavity air volume and total cavity air volume. None of the other imaging features could reliably differentiate the various treatment response groups from one another.

With regards to the statistical analysis, it may be problematic to compare one large group of patients with two other groups that are significantly smaller. Therefore, important differences could potentially emerge with bigger patient numbers in the other groups.

In addition, ROCs were drawn for each variable and thresholds adjusted in order to determine the best possible sensitivity vs specificity trade-off. The authors detail the effects of varying thresholds in various figures and tables. These indicated that any algorithm with high sensitivity for high-risk patients (with unfavourable outcomes), suffered from low specificity for low-risk patients (and vice

versa). The algorithm was therefore adapted to stratify 50% of patients to the higher risk arm and 50% to the lower risk arm.

Study limitations included the lack of directly observed treatment, with subsequent inability to differentiate poor adherence from true treatment failure and true relapse from re-infections and are detailed in the discussion.

The developed algorithm is used in the ongoing PredictTB trial and its prediction performance remains to be seen.

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Gallium-68 based PET, Theranostics, Infection and Inflammation imaging

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 03 Feb 2021

Ray Chen

We thank the reviewer for these positive comments. With respect to the smaller number of unfavorable outcomes, we use statistical methodology that conditions on outcomes (e.g., methods appropriate for case-control studies), which appropriately accounts for different sample sizes in the two groups. With respect to the total sample size, as stated in our limitations paragraph, we acknowledge the challenge of drawing definitive conclusions from small numbers. This became evident early on in the trial by the imbalance in arm distribution, requiring a modification to the early treatment completion criteria. However, these were the only data we had when we began the trial. With the data generated in the

PredictTB trial, we may find that additional modifications to the early treatment completion criteria are necessary.

Competing Interests: None.