Immunodiagnosis of latent tuberculosis : new answers to an old question?

Franken, W.P.J.

8

Variation in T-SPOT.*TB* spot interpretation between independent observers of different laboratories

Willeke P.J. Franken[1], Steven Thijsen[2], Ron Wolterbeek[3], John J.M. Bouwman[2], Hanane el Bannoudi[1], Sandra V. Kik[4,5], Jaap T. van Dissel[1], Sandra M. Arend[1].

[1]Department of Infectious Diseases, Leiden University Medical Center, Leiden, The Netherlands
[2]Department of Medical Microbiology and Immunology, Diakonessenhuis Utrecht/Zeist, The Netherlands
[3]Department of Medical Statistics, Leiden University Medical Center, Leiden, The Netherlands
[4]KNCV Tuberculosis Foundation, The Hague, The Netherlands
[5]Center for Infection and Immunity Amsterdam, Academic Medical Center Amsterdam, The Netherlands

SUBMITTED FOR PUBLICATION

ABSTRACT

Background

T-SPOT.*TB* is a specific assay for the diagnosis of tuberculosis. The assay needs to be performed with freshly isolated cells and interpretation requires training. T-SPOT.*TB* has been used in various clinical-epidemiological settings, but so far no studies evaluated the effect of freezing and thawing of blood cells before testing or inter-observer variation in test reading.

Aim

To compare T-SPOT.*TB* results obtained with freshly isolated or frozen cells and to evaluate variation between different observers in reading T-SPOT.*TB* results.

Materials & Methods

The study was nested within an ongoing cohort study, in which part of the T-SPOT.*TB* had been performed with frozen material. Culture plates were read visually by four different observers from two laboratories, and by two automated readers.

Results

Of 313 T-SPOT.*TB* assays, 235 were performed with fresh and 78 with frozen cells. No significant difference was found between results obtained with fresh or frozen cells. The percentage of positive results varied between readers by maximally 17%; 5/6 raters were within a 6% difference in positive results. Analysis of the observed inter-rater differences showed that some individuals systematically counted more spots. Because test interpretation includes subtraction of background values, this systematic variance had little influence on inter-individual differences.

Conclusion

The findings of this study suggest that it is possible to use thawed blood cells for T-SPOT.*TB*. The test result as positive or negative varied between independent raters, mainly due to samples with values around the cut-off. This warrants further study regarding determinants affecting the reading of T-SPOT.*TB*.

INTRODUCTION

Roughly a century after the introduction of the tuberculin skin test (TST), the recent development of interferon-gamma release assays (IGRA) for specific detection of infection with *Mycobacterium tuberculosis* has realized a new class of immunodiagnostic tests that has extensively been evaluated both for detection of active tuberculosis (TB) and of latent TB infection (LTBI) (1-4). T-SPOT.*TB*® and QuantiFERON-TB® Gold in-tube are the commercially available and approved IGRA formats, being based on culture of isolated peripheral blood mononuclear cells and of whole blood, respectively. Numerous studies that evaluated the use of IGRA have been published in the past years showing their particular value for detection of LTBI in populations with high rates of false-positive TST due to BCG vaccination or exposure to nontuberculous mycobacteria (5;6). T-SPOT.*TB* is based on the ELISPOT technique in which cells responding with interferon-γ production after antigen stimulation are visualized as spots, which must be enumerated. The assay is performed in four wells with different stimulations: medium as negative control, PHA as positive control and peptides of the TB specific antigens ESAT-6 (panel A) and CFP-10 (panel B). One of the disadvantages of T-SPOT.*TB* is that it must be performed with fresh material which may not always be convenient. As the assay is based on single well culture for each stimulus, random variability cannot be detected. Another disadvantage is that counting the spots might lead to variation when read by different observers or automated readers.

Thus far, no studies have addressed the inter-observer variability of the T-SPOT.*TB*. In the present study, these issues were addressed by using material obtained within an ongoing cohort study in the Netherlands in which part of T-SPOT.*TB* assays was performed with frozen material for logistical reasons (blood arriving in the laboratory on a Friday was frozen since the assay needs to be completed 20 hours later). We compared results of T-SPOT.*TB* obtained with freshly isolated to those with frozen and thawed cells. Secondly, we evaluated the reading of the T-SPOT.*TB* plates in two laboratories by different observers.

SUBJECTS AND METHODS

Materials and data for this analysis were obtained from an ongoing cohort study in the Netherlands which aims to assess the predictive value of the TST and IGRA for development of active TB among immigrants who are close contacts of a smear-positive TB patient (unpublished data). The baseline report of this study is submitted for publication.

T-SPOT.TB

T-SPOT.*TB* was performed following the manufacturers instructions (http://www.oxfordimmunotec.com). When blood was obtained on a Friday, cells were isolated and frozen at minus 152°C until testing. The cells were frozen in RPMI medium containing 10% DMSO and 10% Fetal Calf Serum (FCS) and stored for a maximum of two weeks at minus 70°C until they were transferred to minus 152°C.

The number of spots was scored visually using a magnifying glass by four independent observers, two from the department of Medical Microbiology of Diakonnessenhuis Utrecht, and two from the department of Infectious Diseases of Leiden University Medical Center. None of the observers had knowledge of TST or T-SPOT.*TB* scores of the other raters. All four observers had received individual training in reading T-SPOT.*TB*. In addition, spots were counted by two automated readers, the Biosys Bioreader 3000 pro and the Biosys Bioreader 5000. Interpretation of test results was according to the criteria defined by the manufacturer; a sample was defined positive if the number of spots in panel A minus Nil and/or in panel B minus Nil exceeded 5. If the number of spots in the Nil well was 6 to 10 the sample was considered reactive if the spot count in panel A or B was more than twice the number of spots in the Nil. If the Nil spot count was 11 to 20 spots, the spot count in panel A or B needed to be at least three times the spot count in the Nil for the sample to be considered responsive. If the spot count in the Nil was more than 20, the sample was considered inconclusive. For the analysis all samples with spot counts in panel A or B over 20 are reported as 20, since two raters did not quantify spot numbers over 20.

Statistical analysis

Differences between results obtained with fresh and frozen cells and different observers and readers were calculated using mixed models. Differences in percentage of positive results were analyzed with chi-square test. Since two raters did not quantify spot numbers above 20 spots, all analyses have also been performed on the selection spot count >20; samples where two or more raters obtained values >20 were excluded. Out of the six raters, one was randomly appointed as reference rater. Analyses were performed using SPSS 14.0 for Windows (Chigaco, IL, USA). Two-sided P values ≤ 0.05 were considered statistically significant.

RESULTS

In total, T-SPOT.*TB* measurements of 313 subjects were available. The assay was performed 235 times with freshly isolated PBMC's and 78 times with frozen PBMC's (maximum interval between freezing and thawing was 207 days with an average of 95 days). In figure 1a and 1b spot counts in panel A minus Nil and panel B minus Nil are depicted for all six observers. In Table 1 the final T-SPOT.*TB* results of all six raters are shown. All but one rater scored between 51% and 58% positive results, the other rater reporting 41% positive results.
The first question was whether the difference in positivity could be explained by the fact that part of the samples were tested with frozen material which may lead to non-specific background stimulations.

Table 1. T-SPOT.*TB* results according to 6 independent raters for all samples (N=313), for samples tested with freshly isolated PBMC's (N=235) and for samples tested with frozen and thawed PBMC's (N=78)

	All		Fresh		Frozen	
Rater	Positive N / 313 (%)	Inconclusive N/ 313 (%)	Positive N / 235 (%)	Inconclusive N/ 235 (%)	Positive N / 78 (%)	Inconclusive N/ 78%)
1	181 (57.8)	4 (1.3)	136 (57.9)	1 (0.4)	45 (57.7)	3 (3.8)
2	164 (52.4)	4 (1.3)	125 (53.2)	1 (0.4)	39 (50.0)	3 (3.8)
3	160 (51.1)	9 (2.9)	122 (51.9)	3 (1.3)	38 (48.7)	6 (7.7)
4	139 (44.4)	3 (1.0)	112 (47.7)	1 (4.7)	27 (34.6)	2 (2.6)
5	172 (55.0)	24 (7.7)	131 (55.7)	11 (4.7)	41 (52.6)	13 (16.7)
6	179 (57.2)	29 (9.3)	135 (57.4)	16 (6.8)	44 (56.4)	13 (16.7)

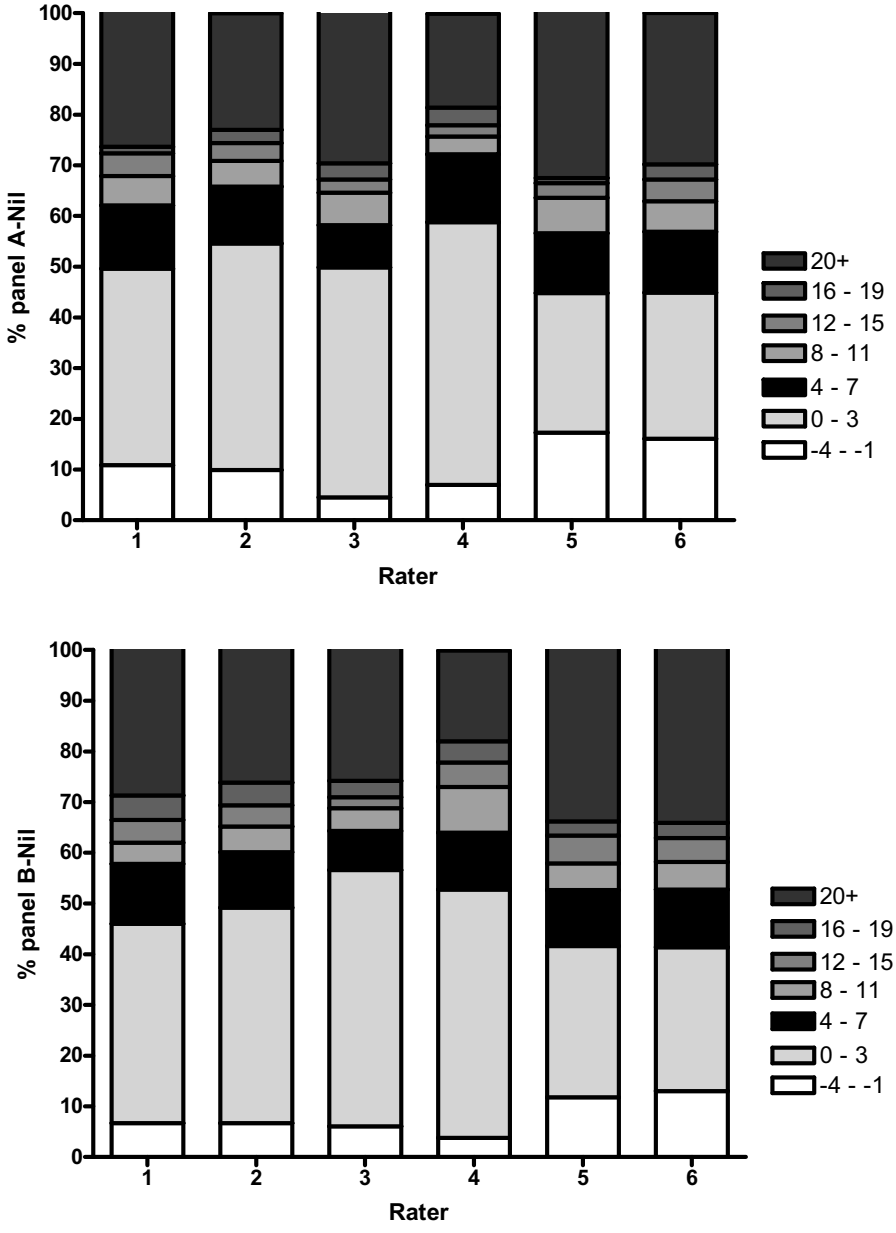

Figure 1. Spot count of 6 individual raters for panel A minus Nil (a) or panel B minus Nil (b).

Figure 1a. Distribution of spot counts for panel A minus Nil of 6 independent raters

Figure 1b. Distribution of spot counts for panel B minus Nil of 6 independent raters
Panel A = ESAT-6
Panel B = CFP-10
Nil = background stimulation

Fresh versus frozen and thawed samples

Results of all fresh samples were compared with results of all frozen samples. The average of all six observers was calculated. Average spot counts in the Nil, panel A and panel B were lower for freshly tested samples compared to the frozen samples, although only significantly lower for the Nil count (Table 2a and 2b). Since the final result of T-SPOT.*TB* is determined by subtracting the spots counted in the Nil well from those in panel A and panel B, the analysis was also performed for the spot count in panel A minus Nil and panel B minus Nil, showing no statistically significant difference in test outcome between fresh and frozen samples for both panels (p=0.909 resp. 0.268). When repeating the analysis excluding the 127 samples with high spot counts above 20 spots per well, similar results were obtained with no significant difference (p=0.105 resp. 0.761).

From the above analysis it appeared that the number of spots scored for panels A or B minus Nil was not influenced by the status of the PBMC's used (frozen or not). This should imply that the percentage of positive results does not differ between fresh or frozen cells. For all raters except number 4 this was indeed true (Table 1).

Table 2a. Influence of fresh or frozen material on T-SPOT.***TB*** results from mean of all six raters (n=235 for fresh and n= 78 for frozen samples, total n=313)

Panel	Material	Mean Spot count (SD)	Difference (95% CI)	p-value
Nil	Fresh	1.2 (0.2)	-2.4 (-3.3 to -1.6)	<0.001
	Frozen	3.6 (0.4)		
A	Fresh	8.1 (0.5)	-1.6 (-3.6 to 0.3)	0.104
	Frozen	9.7 (0.8)		
B	Fresh	8.9 (0.5)	-0.6 (-2.5 to 1.4)	0.555
	Frozen	9.5 (0.9)		
A-Nil	Fresh	7.3 (0.5)	0.1 (-2.0 to 2.2)	0.909
	Frozen	7.2 (0.9)		
B-Nil	Fresh	8.2 (0.5)	1.2 (-0.9 to 3.3)	0.268
	Frozen	7.0 (0.9)		

Panel A = ESAT-6
Panel B = CFP-10
SD = standard deviation
CI = confidence interval

Table 2b. Influence of fresh or frozen material on T-SPOT.*TB* results from mean of all six raters, excluding samples with more than 20 spots (n=140 for fresh and n=46 for frozen samples, total n=186)

Panel	Material	Mean Spot count (SD)	Difference (95% CI)	p-value
Nil	Fresh	1.0 (0.2)	-1.3 (-2.0 to -0.5)	0.001
	Frozen	2.2 (0.3)		
A	Fresh	3.0 (0.3)	-2.0 (-3.3 to - 0.7)	0.002
	Frozen	5.0 (0.6)		
B	Fresh	3.4 (0.3)	-1.3 (-2.6 to - 0.1)	0.049
	Frozen	4.7 (0.6)		
A-Nil	Fresh	2.0 (0.3)	-1.1 (-2.5 to 0.2)	0.105
	Frozen	3.2 (0.6)		
B-Nil	Fresh	2.4 (0.3)	-0.2 (-1.6 to 1.2)	0.761
	Frozen	2.6 (0.6)		

Differences between six independent raters

In Figure 2, the mean spot count and absolute differences in spot count are depicted for all six raters. The most important observation was that each individual rater had his or her own consistent preference for counting spots, some raters counting more or less spots than others, but do this consistently in all the wells of a particular sample, eventually resulting in no significant difference between the final calculation of counts in panel A or B minus Nil. Only scores of rater 3 and 4 for panel B minus Nil were significantly lower than those obtained by the reference rater and of rater 4 also for panel A minus Nil. When analyzing only the samples with spot counts less than 20, it appeared that only rater 4 was significantly lower than the reference rater (Figure 2; panels B and D).

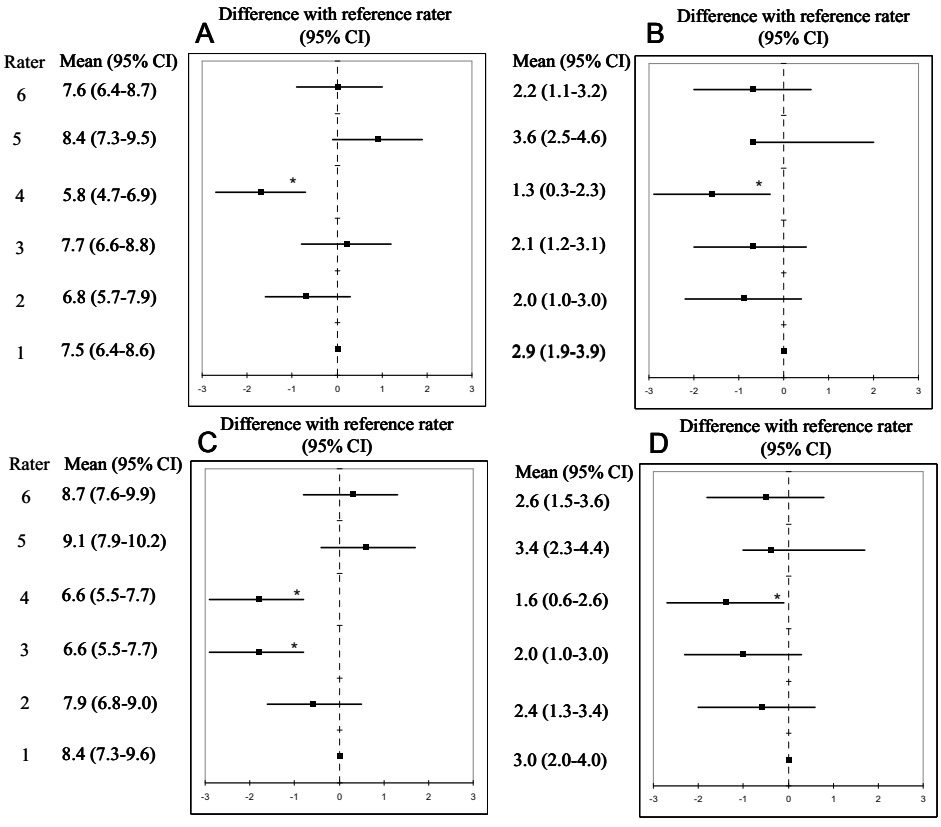

Figure 2. Difference in spot count between 6 different raters compared to one reference rater.
Panel A represents the results of panel A minus Nil for all samples; panel B represent panel A minus Nil for samples with counts <20. Panel C represents results of panel B minus Nil for all samples and panel D represents the results of panel B minus Nil for samples with counts <20.
Panel A = ESAT-6
Panel B = CFP-10
Nil = background stimulation

DISCUSSION

This study was initiated to determine the effect of different human and automated readers on the test results of T-SPOT.*TB*. In addition, we compared tests performed using freshly isolated cells with tests using frozen and thawed cells. The main finding was a maximum difference of 15% in final test interpretation (i.e., positive or negative) between the six raters in this study cohort that was characterized by

a high overall rate of positive tests. The second main finding was no difference between proportion final positive samples between fresh and frozen material.

The observed significant difference in spot counts between the six raters, with a maximally 17% difference in positive results, is an important finding. When results of T-SPOT.*TB* are used for clinical decision making, the test result should be objective and not affected by variations between different raters. According to the manufacturer it is allowed to either count spots visually using a magnifying glass or by use of an automated spot reader. Since the results of five out of six raters, including both automated readers, produced between 51% and 58% positive results, it was most likely that counts by rater 4 were falsely low. However, in the absence of a gold standard for latent TB infection the opposite cannot be refuted.

There was a significant difference in the absolute number of spots counted in the Nil, panel A and panel B when performed with freshly isolated PBMC's compared to the assay performed with frozen and thawed cells (excluding the samples with spot count over 20). However, after subtracting the number of spots in the Nil there was no difference in the final results. Although the average spot count was not influenced by the status of the PBMC's (freshly isolated or frozen and thawed), in one of the six raters the reported percentage of positive results was significantly lower in samples that had been frozen. Smith et al showed in 2007 that use of thawed cells did not influence the results of an in-house ELISPOT assay if freezing was done using the standard protocol (7). The results of the present study confirm that finding for the commercially available T-SPOT.*TB*. This could be important in a setting where pooling of samples is preferable or unavoidable, as e.g. for research purposes or when the number of clinical samples is small. For daily practice in routine laboratories we think that it is not advantageous to freeze samples before testing because test results generally need to be available on a short time basis.

A limitation of our study was the small number of samples that was tested both as fresh and frozen and thawed cells. As an alternative we compared all assays performed with fresh material with those performed with frozen material. A direct comparison in a larger number of samples tested both with fresh and with frozen material should be performed before definite conclusions can be drawn. Our study only addressed the reading of already processed plates and did not study inter-laboratory variation in overall performance of the assay, which could contribute to variation in final test result as well. Further research could include the distribution of blood samples to several laboratories. Of note, the population on which this study was based included an extraordinarily high rate of latently infected individuals,

which should be realized when interpreting the observed absolute differences in positive test results. In routine laboratory settings the positivity rate will in general be much lower and as result the overall agreement between raters can be expected to be higher than that reported in our study. Therefore the inter-observer relative difference of 6-15% of the number of positive results, as was observed in our study, should be taken as the starting point.

In conclusion, our study demonstrates that significant variation in results of the T-SPOT.*TB* can occur between independent observers. This finding warrants further study into determinants of inter-observer variation.

Acknowledgements
The authors wish to thank all the participants and the staff of the participating municipal health services GGD Amsterdam, GGD Den Haag, GGD Eindhoven, Hulpverleningsdienst Flevoland, Hulpverlening Gelderland Midden, Hulpverleningsdienst GGD Groningen (locations Groningen and Assen), GGD Hart voor Brabant, GGD Hollands Midden, GGD Regio Nijmegen, GGD Rotterdam e.o., GGD Regio Twente, GGD Utrecht GGD West-Brabant, GGD Zuid-Holland West, GGD Zuidoost-Brabant. Dr. S. Verver and Mrs. M. Mensen we thank for carefully reading the manuscrip and. Dr. van der Burg of the department of Clinical Oncology for allowing us to use his automated spot reader.

REFERENCE LIST

(1) Menzies D, Pai M, Comstock G. Meta-analysis: New Tests for the Diagnosis of Latent Tuberculosis Infection: Areas of Uncertainty and Recommendations for Research. Ann Intern Med 2007 Mar 6;146(5):340-54.
(2) Bothamley GH. IFN-g-release assays in the management of tuberculosis. Expert Rev Resp Med 2007;3(1):365-75.
(3) Ewer K, Deeks J, Alvarez L, Bryant G, Waller S, Andersen P, et al. Comparison of T-cell-based assay with tuberculin skin test for diagnosis of Mycobacterium tuberculosis infection in a school tuberculosis outbreak. Lancet 2003 Apr 5;361(9364):1168-73.
(4) Arend SM, Thijsen SF, Leyten EM, Bouwman JJ, Franken WP, Koster BF, et al. Comparison of two interferon-gamma assays and tuberculin skin test for tracing tuberculosis contacts. Am J Respir Crit Care Med 2007 Mar 15;175(6):618-27.

(5) Franken WP, Timmermans JF, Prins C, Slootman EJ, Dreverman J, Bruins H, et al. Comparison of Mantoux and QuantiFERON TB Gold tests for diagnosis of latent tuberculosis infection in Army personnel. Clin Vaccine Immunol 2007 Apr;14(4):477-80.

(6) Brodie D, Lederer DJ, Gallardo JS, Trivedi SH, Burzynski JN, Schluger NW. Use of an interferon-gamma release assay to diagnose latent tuberculosis infection in foreign-born patients. Chest 2008 Apr;133(4):869-74.

(7) Smith JG, Joseph HR, Green T, Field JA, Wooters M, Kaufhold RM, et al. Establishing acceptance criteria for cell-mediated-immunity assays using frozen peripheral blood mononuclear cells stored under optimal and suboptimal conditions. Clin Vaccine Immunol 2007 May;14(5):527-37.