Variation in T-SPOT.TB Spot Interpretation between Independent Observers from Different Laboratories\textsuperscript{\textdagger}

Willeke P. J. Franken,\textsuperscript{1} Steven Thijsen,\textsuperscript{2} Ron Wolterbeek,\textsuperscript{3} John J. M. Bouwman,\textsuperscript{2} Hanane el Bannoudi,\textsuperscript{1} Sandra V. Kik,\textsuperscript{4,5} Jaap T. van Dissel,\textsuperscript{1} and Sandra M. Arend\textsuperscript{1*}

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

Received 4 December 2008/Returned for modification 18 February 2009/Accepted 20 August 2009

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 have evaluated the effect of interobserver variation in test reading. Our aim was to evaluate variation between different observers in reading T-SPOT.TB results. 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. Of 313 T-SPOT.TB assays, 235 were performed with fresh cells and 78 were performed with frozen cells. No significant difference was found between results obtained with fresh cells and those obtained with frozen cells. The percentage of positive results varied between readers by maximally 15%; five/six raters were within a 6% difference in positive results. Analysis of the observed interrater differences showed that some individuals systematically counted more spots than others did. Because test interpretation includes subtraction of background values, this systematic variance had little influence on interindividual differences. The test result as positive or negative varied between independent raters, mainly due to samples with values around the cutoff. This warrants further study regarding determinants affecting the reading of T-SPOT.TB.

Roughly a century after the introduction of the tuberculin skin test (TST), the recent development of gamma interferon release assays (IGRA) for specific detection of infection with Mycobacterium tuberculosis has realized a new class of immunodiagnostic tests that have extensively been evaluated for detection both of active tuberculosis (TB) and of latent TB infection (1, 2, 4, 7). 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 (PBMCs) and of whole blood, respectively. Numerous studies that evaluated the use of IGRA have been published in the past several years, showing their particular value for detection of latent TB infection in populations with high rates of false-positive TSTs due to Mycobacterium bovis BCG vaccination or exposure to nontuberculous mycobacteria (3, 5). T-SPOT.TB is based on the enzyme-linked immunospot assay technique in which cells responding with gamma interferon production after antigen stimulation are visualized as spots, which must be enumerated. This can be done by use of an automated spot reader or by using a magnifying glass. The assay is performed in four wells with different stimulations: medium as a negative control, phytohemagglutinin as a positive control, and peptides of the TB-specific antigens ESAT-6 (panel A of the assay) and CFP-10 (panel B of the assay). 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 results are read by different observers or automated readers.

Thus far, no studies have addressed the interobserver 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 the T-SPOT.TB assay was performed with frozen material for logistical reasons (blood arriving in the laboratory on a Friday was frozen since the assay needed to be completed 20 h later). We evaluated the reading of the T-SPOT.TB plates in two laboratories by different observers and by two automated readers. Next, we compared results of T-SPOT.TB obtained with freshly isolated cells to those obtained with frozen and thawed cells.

MATERIALS 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 (the baseline paper was published recently [6]; the follow-up paper has been submitted for publication). T-SPOT.TB was performed following the manufacturer’s instructions (Oxford Immunotec). When blood was obtained on a Friday, cells were isolated by using a density gradient and frozen at \(-152\)C until testing. The cells were frozen in RPMI medium containing 10% dimethyl sulfoxide and 10% fetal calf serum. When cells were thawed for performing the assay, they were thawed

\textsuperscript{\textdagger} Corresponding author. Mailing address: Department of Infectious Diseases, C5-P 39, Leiden University Medical Centre, P.O. Box 9600, 2300 RC Leiden, The Netherlands. Phone: 31 71 5264915. Fax: 31 71 5266758. E-mail: s.marend@lumc.nl.

\textsuperscript{*} Published ahead of print on 26 August 2009.
in RPMI medium with 50% fetal calf serum. When the cells were thawed, they were directly centrifuged and then resuspended in AIM-V medium used for the assay. The cells were rested for about 1 h and then counted and used directly in the assay.

The number of spots was scored visually with a magnifying glass by four independent observers, two from the Department of Medical Microbiology of Diakonessenhuis Utrecht and two from the Department of Infectious Diseases of Leiden University Medical Center. None of the observers had knowledge of the 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 (raters 5 and 6). Interpretation of test results was according to the criteria defined by the manufacturer: a sample was defined as positive if the number of spots in panel A (ESAT-6) minus Nil and/or in panel B (CFP-10) minus Nil exceeded 5. Nil represents the results from the negative control, cells stimulated only with medium. 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 well. 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 well for the sample to be considered positive. If the spot count in the Nil well was more than 20, the sample was considered inconclusive.

Statistical analysis. Differences between results obtained with fresh and frozen cells and with different observers and readers were calculated using mixed models with the reference rater defined as a fixed effect and the sample number defined as a random effect. Differences in percentages of positive results were analyzed with the 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 in the Nil well was more than 20, the sample was considered inconclusive.

RESULTS

In total, T-SPOT.TB measurements of 313 subjects were available. The assay was performed 235 times with freshly isolated PBMCs and 78 times with frozen PBMCs (maximum interval between freezing and thawing was 207 days with an average of 95 days). In Fig. 1a and b, 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 remaining rater reporting 44% positive results.

Differences between six independent raters. In Fig. 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 counted more or fewer spots than others but did 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 and those in panel B minus Nil. Only scores of raters 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 only the samples with spot counts less than 20 were analyzed, it appeared that only rater 4 was significantly lower than the reference rater (Fig. 2B and D).

Agreement between six independent raters. In Table 2, agreement between the independent raters is expressed in

TABLE 1. T-SPOT.TB results according to six independent raters for all samples, for samples tested with freshly isolated PBMCs, and for samples tested with frozen and thawed PBMCs

<table>
<thead>
<tr>
<th>Rater</th>
<th>All (n = 313)</th>
<th>Fresh (n = 235)</th>
<th>Frozen (n = 78)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Inconclusive</td>
<td>Positive</td>
</tr>
<tr>
<td>1</td>
<td>181 (57.8)</td>
<td>4 (1.3)</td>
<td>136 (57.9)</td>
</tr>
<tr>
<td>2</td>
<td>164 (52.4)</td>
<td>4 (1.3)</td>
<td>125 (53.2)</td>
</tr>
<tr>
<td>3</td>
<td>160 (51.1)</td>
<td>9 (2.9)</td>
<td>122 (51.9)</td>
</tr>
<tr>
<td>4</td>
<td>139 (44.4)</td>
<td>3 (1.0)</td>
<td>112 (47.7)</td>
</tr>
<tr>
<td>5</td>
<td>172 (55.0)</td>
<td>24 (7.7)</td>
<td>131 (55.7)</td>
</tr>
<tr>
<td>6</td>
<td>179 (57.2)</td>
<td>29 (9.3)</td>
<td>135 (57.4)</td>
</tr>
</tbody>
</table>
kappa values. All kappa values are above 0.6, indicating that the overall agreement varies from moderate to good.

**Fresh versus frozen and thawed samples.** Since the number of samples tested with both fresh and frozen material was limited, not allowing direct comparison, as an alternative the results of all fresh samples were compared with results of all frozen samples. The absolute number of spots in the Nil well and panel A and panel B of frozen cells was higher than that in fresh samples (data not shown). Since for the final result the Nil count is subtracted from both panels, the increase was corrected, and it appeared that there was no statistically significant difference in test outcome between fresh and frozen samples for both panels (data not shown). The percentage of positive test results also did not differ between fresh and frozen cells, except for rater 4 (Table 1).

**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 finding was no difference between the proportions of final positive samples between fresh and frozen material.

The observed significant difference in spot counts between the six raters, with a maximally 15% 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 permissible to count spots either 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 numbers of spots counted in the Nil well and panel A and panel B when the assay was performed with freshly isolated PBMCs compared to that when the assay was performed with frozen and thawed cells (excluding samples with spot counts of over 20). However, after subtracting the number of spots in the Nil well, there was no difference in the final results. Smith et al. showed in 2007 that use of thawed cells did not influence the results of an in-house enzyme-linked immunospot assay if freezing was done using a standardized protocol (8). The results of the present study support 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-term basis.

We have not compared the interrater variation of the T-SPOT.TB with the interrater variation of the other commercially available IGRA, the QuantiFERON-TB Gold in-tube. It would also be interesting to analyze the reproducibility of the latter enzyme-linked immunosorbent assay-based test in different laboratories using different spectrophotometers. However, we believe that this variation will be smaller than the variation that we have described in this paper for the T-SPOT.TB, since the enzyme-linked immunosorbent assay results are interpreted by software and thus are free from subjective human interpretation.

A limitation of our study was the small number of samples that were tested both as fresh and as 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 material and with frozen material should be performed before definite conclusions can be drawn. Our study addressed the reading of only already processed plates and did not study interlaboratory variation in overall performance of the assay, which could contribute to variation in final test results as well. Further research could thus include the distribution of blood samples to several laboratories simultaneously. Of note, the population on which this study was based included an extraor-}

dinarily high rate of latently infected individuals, which should be taken into account 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 a result, the overall agreement between raters can be expected to be higher than that reported in our study. Therefore, the interobserver relative difference of 6 to 15% of the number of positive results, as 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 interobserver variation.

ACKNOWLEDGMENTS

We 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 West Brabant, GGD Zuid-Holland West, and GGD Zuidoost-Brabant. We thank S. Verver and M. Mensen for carefully reading the manuscript and S. H. van der Burg of the Department of Clinical Oncology for allowing us to use his automated spot reader. T-SPOT.TB kits were kindly provided by Oxford Immunotec.

This study was funded by unrestricted grants from The Netherlands Organization for Health Research and Development (ZonMw).

The manufacturer had no influence on study design or protocol, data collection, data analysis, or preparation and submission of the manuscript of this study.

REFERENCES