Jornal Brasileiro de Pneumologia
Brazilian Journal of Pulmology

ISSN 1806-9312 (on-line)
19/12/2008

1111 - Vol. **34** / Ed. **5** in **2008**

Section: **Original Article**

Pages: **304** to **311**

Methodology for characterizing proficiency in interpreting sputum smear microscopy results in the diagnosis of tuberculosis

Author(s):

Francisco Duarte Vieira, Julia Ignez Salem, Antônio Ruffino-Netto, Susana Alles de Camargo, Regina Ruivo Ferro e Silva, Lúcia Cristina Corrêa Moura, Meire Jane Vilaça, José Vitor da Silva

Keywords:

Tuberculosis; Microscopy; Reproducibility of results.

Abstract:

Objective: To propose a methodology for characterizing proficiency in sputum smear microscopy for acid-fast bacilli (AFB) in the diagnosis of tuberculosis and to determine the number of microscopies necessary to establish this proficiency, as well as the quality of the transcription of results, the causes of the discrepancies in the readings (rater or microscope used), and the criterion for classification of microscopy results that poses the most difficulty in characterizing proficiency. Methods: Four hundred sputum smear microscopies for the diagnosis of tuberculosis were analyzed through double-blind readings by six professionals who usually read/supervise microscopies performed in public health care facilities. The sample was stratified to obtain, at least, a reliability of 90% in the double-blind readings, an α error of 5%, and a precision of 3%. The results were analyzed using observed reliability and the kappa index. Results: Thirteen errors (0.27%) were found in the transcription of results. Reliability increased when the three distinct categories of positive results (AFB+, AFB++, and AFB+++) were grouped or when inconclusive results were excluded from the analysis. The quantification of the bacterial load was the classification criterion that posed the most difficulty in establishing proficiency. Using higher quality microscopes increased reliability. Reliability values stabilized only from the reading of 75 microscopies onward. Conclusions: Double-blind sputum smear microscopy readings using a panel containing 75 slides (36 negative, 4 inconclusive, and 35 positive) proved to be appropriate for characterizing proficiency in sputum smear microscopy for the diagnosis of tuberculosis when such proficiency is intended to reproduce laboratory routine.

Introduction

The use of microscopy to test for acid-fast bacilli (AFB) in clinical samples of sputum, routinely known as sputum smear microscopy, is the primary tool for the diagnosis and monitoring of cases of pulmonary tuberculosis. It is a rapid, easily performed, affordable technique. However, there is variation in the results of visualization and quantification of AFB when sputum smear microscopy is performed by different professionals (designated raters), since "the degree and the frequency of the error - due to excess or to defect - vary from one person to another and, over time, within a given individual."(1) Therefore, it is essential that a system to ensure the quality of sputum smear microscopy be implemented.

According to the recommendations made by the Association of Public Health Laboratories/Centers for Disease Control and Prevention (APHL/CDC),(2) the quality control system for sputum smear microscopy for tuberculosis includes three components: internal quality control, quality improvement, and external quality control. External quality control makes it possible for the participating laboratories to evaluate their capability by comparing their results to those obtained at other laboratories in the network (central and intermediate laboratories). This process comprises three tools: proficiency testing (examination of panels containing sputum smears); blind second readings of sputum smear microscopies at a hierarchically superior laboratory; and supervision by a professional from a referral laboratory (designated supervisor) in order to review the quality of the sputum smear microscopies and the readings performed in the laboratories under its jurisdiction.

In Brazil, external quality control includes only second readings of sputum smears by comparing the results obtained by local and regional laboratories with the reevaluation of slides performed by Central Laboratories of Public Health, which is also known as indirect supervision. However, it is logical to state that, in order to execute this component of the external quality control satisfactorily, the raters or supervisors responsible for the readings must have proven proficiency in reading sputum smear microscopies, that is, they must have passed a proficiency test in sputum smear microscopy for tuberculosis.

For proficiency testing, the APHL/CDC guidelines(2) suggest the use of one of four models of panels, each one with 10 sputum smear slides. The variation occurs in the number of distinct diagnostic classifications for sputum smear microscopy results (negative, inconclusive, AFB+, AFB++, and AFB+++). Nevertheless, some authors(3,4) point out that, among those involved in quality of sputum smear microscopy for tuberculosis, there is no general consensus regarding these quantifications.

In view of the controversy surrounding the issue, the principal objective of the present study was to propose a methodology that makes it possible to characterize proficiency in sputum smear microscopy for AFB in the diagnosis of tuberculosis, as well as to determine the number of microscopies necessary to establish this proficiency, through the analysis of intra-rater and inter-rater reliability in double-blind readings of sputum smear microscopies. The methodology used also made it possible to evaluate the quality of the transcription of results, to determine whether the discrepancies in the readings of sputum smear microscopies depended on the raters (lack of technical skill) or on the equipment (microscope), and to identify which criterion for classification of sputum smear microscopy results poses the most difficulty in characterizing proficiency.

Methods

The present study was approved by the Ethics in Research Committee of the National Research Institute of Amazônia (Protocol no. 006/2004). Four hundred sputum smear slides for the diagnosis of tuberculosis were analyzed through double-blind readings by six professionals who were readers or supervisors of sputum smear microscopies performed at public health care facilities. The study sample was stratified to obtain a reliability of at least 90% in the double-blind readings, an α error of at least 5%, and a precision of at least 3%.

In order to simulate the routine of professionals working in the field of tuberculosis diagnosis, the sputum smear slides were prepared and stained (5) using 400 sequential sputum samples from anonymous patients with respiratory symptoms and suspected of having pulmonary tuberculosis. These sputum samples were collected in public health care laboratories in the city of Manaus, Brazil. A preliminary reading of the microscopies used for detecting AFB was performed in order to determine whether the sputum smear slides met all of the Brazilian criteria for classification of results.(5) Preliminarily, we found that 46% of the sputum smear microscopies presented negative results, 4% presented inconclusive results, and 50% presented positive results.

The six participating professionals, designated raters, were readers or supervisors of sputum smear microscopies performed at public health care facilities. Of those six, two were representatives of the central-west region, and each one of the remaining four was a representative of a different geographic macro-region of Brazil (north, northeast, southeast, and south). These raters were identified as A, B, C, D, E, and F, and the results of the double-blind readings were identified as A1/A2, B1/B2, C1/C2, D1/D2, E1/E2, and F1/F2, respectively. A flowchart of the procedures is presented in Figure 1.

Slides with an opaque edge, onto which sequential registration numbers were recorded using graphite, were used in the preparation of the sputum smear microscopies. Since each sputum smear slide would be subjected to 12 microscopy readings (six double-blind readings) using an immersion lens and different pieces of equipment, all slides were coverslipped with Entellan to protect them from contamination and loss of staining due to cleaning (removal of immersion oil), which are factors that could have an adverse effect on the second readings of the sputum smear microscopies.(6)

The sputum smear microscopies were sent to the raters according to the current biosafety guidelines for the shipment of biological material via airmail.(7) In order to identify technical deficiencies or prevent equipment-related reliability discrepancies, all raters used the same microscope in their workplace to perform the double-blind readings (first and second readings) of the sputum smear microscopies. All raters received instructions and procedure guidelines for the reading of the microscopies. Every day, 25 slides were analyzed; this number being recommended by Van Deun & Portaels(8) for daily reading by experienced microscopists. Therefore, the 400 sputum smear microscopies analyzed were divided into 16 blocks of 25 slides each.

While each sputum smear slide was being read, the number of AFB found in each microscopic field examined was recorded on graph paper by the raters in order to determine the means and report the results.(5) The results were transcribed to a form designed for reporting the results of readings of sputum smear microscopies. This form included fields for the slide number and the date each block of slides was read. The same procedure was adopted in the second reading. To ensure a blinded evaluation, the registration numbers of the slides were changed between the two readings. This step was performed by a professional who did not participate in the reading of the sputum smear microscopies. This professional maintained the database confidential until the end of the investigation.

In order to ensure the reliability of the results obtained in the first and second readings, the information transcribed from the graph paper to the form for transcription of results was checked, and the errors detected were registered for later analysis. It was the results on the graph paper that were entered into an electronic spreadsheet that was created using the program Epi Info for Windows, version 3.3.2. The most consistent result obtained for each slide by the different raters was considered the standard reading for that slide and was identified as P1 (first reading) or P2 (second reading).

In order to determine whether the discrepancies in the readings of the sputum smear microscopies depended on the raters (lack of technical skill) or on the equipment used (microscope), the observed intra-rater reliability and kappa index of the results of the double-blind readings of the sputum smear microscopies were calculated. Inter-rater reliability and rater/standard reading reliability were also calculated. Calculations were made using the program Epi Info for MS-DOS, version 6.04d, and the interpretation of kappa was established according to the recommendations made by Pereira.(9)

In order to analyze which criterion for the classification of sputum smear microscopy results poses the most difficulty in characterizing proficiency, the distinct diagnostic categories of microscopy results were grouped, and observed reliability was calculated according to the results obtained by each rater.

In order to determine the minimal number of microscopies necessary to evaluate proficiency in sputum smear microscopy for the diagnosis of tuberculosis, the observed reliability in blocks of 50, 75, 100, and 125 microscopies was analyzed with the objective of establishing the time point at which the raters would obtain values equal to or greater than 90% in three sets analyzed for each block.

Results

The analysis of the transcription of results from the graph paper to the form provided revealed 13 transcription errors, corresponding to 0.27% of the 4800 readings of sputum smear microscopies. Of those 13 errors, 7 were considered highly significant, since they created false-negative results.

Intra-rater reliability was initially analyzed by diagnostic classification of sputum smear microscopy results, which is standardized into five categories: inconclusive, negative, AFB+, AFB++, and AFB+++.
Subsequently, considering that the distinct categories of positivity present differences only in the number of AFB and that these differences are not considered diagnostic discrepancies,(2) these categories were grouped, and reliability results were obtained for three categories (negative, inconclusive, and AFB+) and two categories (negative and AFB+). The values found are presented in Table 1.

Table 2 shows the observed and kappa values for intra-rater reliability, inter-rater reliability, and rater/standard reading reliability in the 400 sputum smear microscopies analyzed using only three diagnostic result categories (negative, inconclusive, and AFB+).

The analysis of which criterion for classification of sputum smear microscopy results poses the most difficulty in characterizing proficiency generated the data shown in Table 3.

The observed reliability in the blocks of 50, 75, 100, and 125 sputum smear microscopies for the determination of the minimal number of microscopies necessary to evaluate proficiency is presented in Table 4.

Discussion

Highly significant errors (false-negative or false-positive) in the transcription of the results of readings of sputum smear microscopies are considered serious errors by the APHL/CDC.(2) Related scientific studies do not mention these serious errors, or, at most, mention them without statistical data and as personal information or information by another author, as is found in the editorial by Van Deun.(10) However, this type of error can even result in legal action against the responsible analyst and the respective institution. A false-positive result causes human suffering and incurs financial costs, whereas a false-negative result incurs costs on society, causes harm to the patient due to the delay in diagnosis, and makes physicians lose faith in the services offered.(11)

While determining whether the discrepancies in the readings of the sputum smear microscopies depended on the raters, we observed that reliability increased (Table 1) when the three distinct categories of positive results (AFB+, AFB++, and AFB+++) were grouped and when inconclusive results were excluded from the analysis. When only the kappa values obtained with the five categories of sputum smear microscopy results were analyzed, it was found that all raters presented good intra-rater reliability in their interpretation (kappa ranging from 0.61 to 0.80), except for rater A, who presented excellent reliability (kappa ranging from 0.81 to 0.99). It is of note that rater A read the sputum smear microscopies using a modern microscope with resolution and field-of-view superior to those of the equipment used by the other raters. Consequently, the quality of the microscope used seems to be important for obtaining higher reliability values, since the only differential among the raters was the type and the quality of the equipment.

The observed reliability values obtained without grouping the categories of positivity were also analyzed (Table 1). Nevertheless, none of the raters obtained the 90% reliability predicted in the present study. However, the mean observed reliability (81.2%) is similar to that reported in the study conducted by Martinez-Guarneros et al.(83%).(4) The difference in relation to maximum reliability (100%) might be related to a lack of technical skill on the part of the raters in quantifying AFB or to inconsistency in the reproducibility of the sputum smear microscopy due to limitations inherent to the technique itself. Among the limitations, the difficulty in reading the same microscopic fields at two different time points stands out,(3) even if there is standardization determining the initial positioning of the first microscopic field to be analyzed and the direction that should be followed in reading 100 fields.(5)

Further analysis of Table 1 reveals that excluding the variable quantification of AFB, represented by the number of plus signs in the sputum smear microscopy results, increases intra-rater reliability. Although the distinction indicated by the plus signs is not an essential condition for the diagnosis of tuberculosis, it is important for the follow-up treatment, since it provides the health professional with information about the effectiveness of the medications prescribed.(12)

The analysis of Table 2 reveals that raters D and E presented the highest inter-rater reliability and rater/standard reading reliability. Although raters A and E presented equal and better intra-rater reliability, it was raters D and E who presented the highest rater/standard reading reliability, with rater A presenting the third highest value. Since rater A was the only one who read the sputum smear microscopies using a modern microscope with resolution and field-of-view superior to those of the equipment used by the other raters, the rater A results contributed less to composing the standard reading, and this might explain the fact that rater A presented lower rater/standard reading reliability than did raters D and E. Based on these results, we can conclude that, for characterizing proficiency, the microscope is a variable that affects the results, especially if such proficiency is based on panels that use the report issued by the organizer of the panels as the result of reference.

According to the APHL/CDC guidelines,(2) reliability in double-blind readings of sputum smear microscopies is expected to be near 95% for highly positive smears (AFB++ and AFB+++), and from 30 to 50% for inconclusive smears (1-9 AFB/100 fields). Values lower than those of the reliability reported indicate technical deficiency suggestive of which criterion or criteria of classification of sputum smear microscopy results pose the most difficulty in characterizing proficiency. Therefore, the reliability values presented in Table 3, which are related to the grouping of AFB++ and AFB+++, the grouping of AFB+, AFB++, and AFB+++, and the grouping of AFB+, AFB++, AFB+++, and inconclusive results, all of which were considered positive results for the purpose of diagnosis, demonstrate a quantification error in the reading of the microscopies. These errors can result from lack of technical skill on the part of the rater or from the fact that the raters did not follow the instructions and procedure guidelines provided for the reading of the sputum smear microscopies.

Systematization can be related to the initial positioning of the first microscopic field to be analyzed, which consequently modifies the other fields due to the deviation of direction, as well to the count of the number of AFB in each field analyzed. It is most likely that the problem really is one of systematization, since reliability over 90% was obtained in results that were bound to be false-negative, as is the case of the reliability obtained between AFB+ and negative results (97.5%) and between inconclusive and negative results (92.2%). Therefore, the raters demonstrated competence in detecting AFB, even when present in small quantities only.

The hypothesis regarding systematization becomes even more convincing when the concordance between AFB+ results and inconclusive results is observed (Table 3). Although reliability did not reach the standard of 90%, these are results that indicate the absence of false-positive results, being considered only errors of quantification of the number of bacilli. As previously stated, such errors can result from different initial positioning of the first microscopic field in the first and second readings.

Further analysis of Table 3 reveals that reliability between inconclusive results is low, its mean (30.5%) being at the lower limit of what is established in the APHL/CDC guidelines.(2) Nevertheless, this mean was more closely related to modifications of results for AFB+ criteria, since the mean reliability between inconclusive and negative results was 92.2%, and the mean reliability between AFB+ and negative results was 97.5%. Therefore, this indicates that the number of discrepancies indicative of false-positive results is low. This fact is supported by the reliability obtained between AFB+ results, inconclusive results, and negative results, whose mean was 86.5%, higher than those obtained in the analyses of the groupings of different degrees of positive results, or in that of the grouping of different degrees of positive results and inconclusive results. These results further the hypothesis that reliability is related to the systematization of the reading of sputum smear microscopies in double-blind studies, or even in those in which the slides are read a second time by other raters.

In summary, through the analysis of the data presented in Tables 1 and 3, we found that the quantification of the number of bacilli was the factor posing the most difficulty in characterizing proficiency. Therefore, any one of the criteria for positivity poses difficulty in determining proficiency.

The results presented in Table 4 indicate that only raters A, D, and E, as well as the standard reading, had, at most, 3 sets of sputum smear microscopies with observed reliability values equal to or lower than 90%. Since this was the criterion established in the proposal of the present study, the number of sputum smear slides necessary for characterizing proficiency was analyzed based on the results obtained by those raters. Rater A presented consistent means in all blocks examined, indicating that, for that rater, 50 slides would be sufficient for establishing proficiency. However, rater D presented consistent means only from the reading of 75 slides onward. Rater E presented an increase in mean reliability in the block of 75 slides as compared with that of 50 slides, and, in the block of 100 slides, presented the same mean reliability as that seen for the block of 50 slides. The standard reading showed a tendency toward stabilization from the reading of 100 slides onward, presenting values higher than those seen for the reading of 75 slides.

Therefore, mean reliability stabilized or increased from the block of 75 slides onward, except in the case of rater E. The data lead to the conclusion that a panel for characterizing proficiency in sputum smear microscopy for the diagnosis of tuberculosis should have at least 75 slides, including approximately 48% negative slides, 5% inconclusive slides, and 47% positive slides. In the present study, these percentages corresponded to 36 negative, 4 inconclusive, and 35 positive (10 AFB+, 12 AFB++, and 13 AFB+++) sputum smear microscopies. The quantification of the number of bacilli was a determining factor of higher or lower intra-rater reliability, whereas inconclusive results indicated the competence of raters in dealing with debatable false-negative and false-positive results.

Despite the composition suggested, intra-rater reliability will ultimately characterize proficiency. Therefore, in order to ensure an optimal standard of proficiency, consistent with that of the raters evaluated in the present study, reliability should be no lower than 90% in any panel created for this purpose, regardless of the number of slides or the composition of the panels. However, if such proficiency is intended to reproduce laboratory routine, it is advisable that 75 sputum smear microscopies be used, 36 of which should be negative, 4 of which should be inconclusive, and 35 of which should be positive.

Acknowledgments

We would like to express our gratitude to the following sources of funding: the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior/Demanda Social (CAPES/DS, Coordination of the Advancement of Higher Education/Social Demand) and the Fundação de Amparo à Pesquisa do Estado do Amazonas/Ministério da Saúde/Conselho Nacional de Desenvolvimento Científico e Tecnológico (FAPEAM/MS/CNPq, Foundation for the Support of Research in the State of Amazonas/Brazilian National Ministry of Health/National Council for Scientific and Technological Development; Grant no. 981/05). We also wish to thank the Amazonas State Central Laboratory of Public Health, represented by the biochemical pharmacologist Oneide Silva. In addition, we would like to thank Maria de Fátima Barbosa, of the Federal University of Amazonas Department of Pathology and Legal Medicine, for her technical assistance in coverslipping the slides.

References

1. Toman K. Toman's Tuberculosis: case detection, treatment, and monitoring: questions and answers. 2nd ed. Geneva: World Health Organization; 2004. p. 334.

2. APHL/CDC - Association of Public Health Laboratories / Centers of Disease Control and Prevention. External Quality Assessment for AFB Smear Microscopy. Washington, DC: APHL; 2002.

3. Paramasivan CN, Venkataraman P, Vasanthan JS, Rahman F, Narayanan PR. Quality assurance studies in eight State tuberculosis laboratories in India. Int J Tuberc Lung Dis. 2003;7(6):522-7.

4. Martinez-Guarneros A, Balandrano-Campos S, Solano-Ceh MA, Gonzalez-Dominguez F, Lipman HB, Ridderhof JC, et al. Implementation of proficiency testing in conjunction with a rechecking system for external quality assurance in tuberculosis laboratories in Mexico. Int J Tuberc Lung Dis. 2003;7(6):516-21.

5. Brasil. Ministério da Saúde. Tuberculose - diagnóstico laboratorial - baciloscopia. Brasília: Série TELE-LAB; 2001.

6. Van Deun A, Roorda FA, Chambugonj N, Hye A, Hossain A. Reproducibility of sputum smear examination for acid-fast bacilli: practical problems met during cross-checking. Int J Tuberc Lung Dis. 1999;3(9):823-9.

7. International Air Transport Association, and International Civil Aviation Organization. Dangerous goods regulations. Montreal: International Air

Transport Association; 2003.

8. Van Deun A, Portaels F. Limitations and requirements for quality control of sputum smear microscopy for acid-fast bacilli. Int J Tuberc Lung Dis. 1998;2(9):756-65.

9. Pereira MG. Epidemiologia: teoria e prática. 6th ed. Rio de Janeiro: Guanabara Koogan; 2002.

10. Van Deun A. External quality assessment of sputum smear microscopy: a matter of careful technique and organisation. Int J Tuberc Lung Dis. 2003;7(6):507-8.

11. WHO/IUATLD. International Course on the Management of Tuberculosis Laboratory Networks in Low-Income Countries. Ottawa, Canadá. 2000.

12. Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. Guia de vigilância epidemiológica. 6. ed. Brasília: Ministério da Saúde; 2005.

Study carried out in the Mycobacteriology Laboratory of the National Research Institute of Amazônia, Manaus, Brazil.
1. Biochemical Pharmacologist in the Federal District of Brasília Central Laboratory of Public Health, Brasília, Brazil.
2. Full Researcher at the National Research Institute of Amazônia, Manaus, Brazil.
3. Full Professor in the Department of Social Medicine. University of São Paulo at Ribeirão Preto School of Medicine, Ribeirão Preto, Brazil.
4. Biochemical Pharmacologist in the Rio Grande do Sul Central Laboratory of Public Health. Rio Grande do Sul State Foundation for Health Science Research, Porto Alegre, Brazil.
5. Biomedical Researcher at the Adolfo Lutz Institute, Laboratory I, Santo André, Brazil.
6. Biochemical Pharmacologist in the Paraíba Central Laboratory of Public Health, João Pessoa, Brazil.
7. Pathology Technician in the Amazônia Central Laboratory of Public Health, Manaus, Brazil.
8. Pathology Technician in the Federal District of Brasília Central Laboratory of Public Health, Brasília, Brazil.
Correspondence to: Julia Ignez Salem. INPA/CPCS, Av. André Araújo, 2.936, CEP 69060-001, Manaus, AM, Brasil.
Tel 55 92 3643-3058. Fax 55 92 3643-3061. E-mail: salem@inpa.gov.br
Submitted: 17 April 2007. Accepted, after review: 2 August 2007.