Semiquantitation of Bacteria in Sputum Gram Stains

PAUL N. VALENSTEIN

Department of Pathology, State University of New York at Stony Brook, Stony Brook, New York 11794-8691, and Laboratory Services, Northport Administration Medical Center, Northport, New York 11768

Received 21 March 1988/Accepted 13 June 1988

In many clinical laboratories, bacteria seen in Gram-stained sputum smears are reported semiquantitatively, using a three- or four-category scale consisting of ratings such as numerous, moderate, rare, and none seen. The consistency with which these categories are assigned was evaluated by repeatedly presenting coded smears to seven experienced microbiology technologists. Technologists rated the same smear twice, pairs of smears prepared from the same specimen, and smears prepared after specimen refrigeration. Agreement was assessed with the weighted kappa test. Semiquantitation of gram-negative rods, gram-positive diplococci, and gram-positive cocci in clusters all showed poor reproducibility (kappa = 0.32, 0.34, and 0.17, respectively). Twenty-four percent of paired ratings differed by two or more categories. Lack of reproducibility was due mainly to the inability of the technologists to render a consistent rating when viewing the same slide on separate occasions (P < 0.001). Variation in the rating styles of different technologists, differences between smears prepared from the same specimen, and specimen refrigeration tended to further decrease the consistency of ratings, but the reductions were not statistically significant. The quantity of potentially pathogenic bacteria in sputum smears is not estimated consistently with standard microscopy procedures and should not be reported.

The sputum Gram stain is widely used as a basis for making a rapid etiological diagnosis in bacterial pneumonia and as a reference point for interpreting the significance of sputum culture (7, 8). In many clinical laboratories, the various types of bacteria seen in Gram-stained smears are reported semiquantitatively by using a three- or four-category scale consisting of ratings such as numerous, moderate, rare, and none seen. In spite of this common practice, the reproducibility of bacterial semiquantitation has been the subject of very few investigations. Penn and Silberman (11) reported that bacteria and yeasts seen in sputum smears prepared from specimens that had been refrigerated for 20 h did not correlate with an initial Gram stain made from the same specimen. However, it is unclear from this report whether discrepancies were due to refrigeration or to an intrinsic unreproducibility of Gram stain evaluation. Bartlett and colleagues reported that when 10 technologists quantified bacteria in gram-stained smears prepared from stock culture suspensions, agreement with expected values averaged 83% for slides with single organisms and 70% for slides with mixed organisms (3).

The present study was designed to examine how reproducibly technologists quantitate bacteria in sputum smears prepared from clinical specimens and to identify the factors that contribute to any inconsistencies in enumeration. Reproducibility was assessed with the weighted kappa test, a statistical procedure that gives partial credit for ratings that are in close but not perfect agreement.

MATERIALS AND METHODS

Smear preparation. Three Gram-stained smears were prepared from 20 fresh early-morning sputum specimens collected in the course of caring for hospitalized patients with new pulmonary infiltrates; no specimen was more than 1 h old. All specimens had fewer than 10 squamous epithelial cells present per low-power (magnification, ×100) microscopic field, as determined by two separate assessments made by an experienced microscopist not participating in the study. Two smears were prepared immediately from the most visibly purulent areas of the specimen by using separate cotton-tipped wooden applicators and alcohol-cleaned glass slides. If only one purulent area was evident, it was used to prepare both smears. Smears were allowed to air dry for 10 min and then heat fixed and Gram stained. The remaining specimen was stored in a dark refrigerator at 4 ± 2°C for 24 h, after which time a third smear was prepared and stained in a manner identical to that used for the first two. All slides were coded numerically and were not otherwise identified.

Study design. Smears were presented to seven licensed microbiology technologists, each of whom had worked in the study laboratory for at least 1 year (mean, 4.2 years) and four of whom had a combined total of 21 years of previous experience in other microbiology laboratories. The technologists were not informed about the details of the study protocol until the investigation was completed. A total of 80 slides were presented to each technologist for interpretation in 4 sets of 20. The first set of 20 slides contained the first and second smears from the first 10 specimens, arranged in random order but consecutively numbered from 1 to 20. After all of the slides had been interpreted, they were removed from the laboratory. The following week a second set of 20 slides was presented. The second set contained the first and third smears from the first 10 specimens, randomly arranged but numbered from 21 to 40. In this way each technologist unknowingly interpreted the first smear twice, once in the first set of 20 slides and once in the second set. Smears from the remaining 10 specimens were presented in the same manner as the first 10 and were numbered from 41 to 80.

Smear interpretation. Technologists were instructed to rate the quantity of bacteria, yeasts, neutrophils, and squamous epithelial cells present in each smear and to record their assessments on a preprinted form. Quantities of gram-positive cocci in clusters, gram-positive diplococci, gram-negative bacilli, gram-positive cocci in chains, gram-negative cocci, and gram-positive bacilli were rated individually. Ratings of bacteria were made on a four-point scale after
TABLE 1. Quantity of gram-negative bacilli in 140 paired sputum smear interpretations

<table>
<thead>
<tr>
<th>Result of examination 2</th>
<th>No. of interpretations in examination 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>13</td>
</tr>
<tr>
<td>Few</td>
<td>8</td>
</tr>
<tr>
<td>Moderate</td>
<td>5</td>
</tr>
<tr>
<td>Numerous</td>
<td>5</td>
</tr>
</tbody>
</table>

examine multiple fields with a Nikon Labophot microscope (total magnification, ×1,000). The categories used were none (less than 1 per field), few (1 to 2 per field), moderate (3 to 15 per field), and numerous (more than 15 per field). When bacteria appeared in groups (chains, clusters, or pairs), technologists were instructed to count the number of groups and not the individual organisms.

To approximate the amount of time spent interpreting a Gram stain in clinical practice, it was recommended that slides be viewed for approximately 2 to 3 min, although more time could be devoted to a slide if desired. Rates were not required to examine a set number of fields or to formally tally and average on paper the number of cells or bacteria seen in each field.

Gram-positive cocci in clusters, gram-positive diplococci, and gram-negative bacilli were considered to represent potential pathogens; technologists' assessments of yeasts, neutrophils, squamous epithelial cells, and other bacteria were not analyzed.

Statistical analysis. Responses for each of the three bacterial morphotypes were coded as integers from 1 through 4. Agreement about the quantity of potentially pathogenic bacteria was measured with the weighted kappa statistic (4, 5). The kappa statistic ranges from 0 to 1, with 1 representing perfect agreement and 0 representing no more agreement than would be expected to occur on the basis of chance. Kappa values below 0.4 represent poor reproducibility; those above 0.8 represent good reproducibility. The weighted kappa test has two advantages over measurements of reproducibility based upon simple agreement. No credit is given for agreement equal to what would be expected to occur on the basis of chance alone, and partial credit is given for responses that are in close but not perfect agreement. In the present study, perfect agreement was given a weight of one, differences of only one category received a weight of two-thirds, and differences of two categories received a weight of one-third.

Kappa values and their standard errors were computed separately for each technologist and each bacterial morphology and were pooled where indicated by using the method of Fleiss (6). Intertechologist reproducibility was determined by pairing each rating of a smear by a technologist with that by another technologist chosen at random from the remaining six.

RESULTS

Table 1 shows the semiquantitative ratings of gram-negative rods made by different technologists interpreting two smears prepared from the same sputum specimen. Agreement between the two interpretations was poor (kappa = standard error, 0.32 ± 0.07). Of the 140 paired ratings, 29 (21%) differed by at least two categories. Intertechologist agreement about the number of gram-positive diplococci and gram-positive cocci in clusters is shown in Tables 2 and 3, respectively. Agreement for both of these morphologies was also poor (0.34 ± 0.07 and 0.17 ± 0.07, respectively). However, agreement for all three classes of bacteria was significantly better than that expected on the basis of chance alone (gram-negative bacilli and gram-positive diplococci, P < 0.01; gram-positive cocci in clusters, P < 0.05).

Lack of agreement between technologists who rated smears from the same specimen was potentially attributable to three separable elements: (i) differences between the rating styles of different technologists, (ii) variation between the two smears prepared from the same specimen, and (iii) an inability of individual technologists to rate a single smear consistently. The contribution of each of these elements to the overall reproducibility of semiquantitation is shown in Table 4. The inability of individual technologists to consistently quantify bacteria in a single smear was the factor that was most responsible for the poor reproducibility of interpretations. When the same observer rated the same smear twice, agreement was substantially less than unity for all three classes of bacteria (P < 0.001). When the same observer rated two smears prepared from the same specimen, agreement tended to be further reduced, but the reduction was not statistically significant. When two technologists rated smears from the same specimen, agreement tended to be reduced even further, but the reduction was statistically significant only for gram-positive cocci in clusters (P < 0.05).

The effect of specimen refrigeration on the reproducibility of interpretations is shown in Table 5. Interpretations of two smears made from a fresh specimen tended to result in better intraobserver agreement than did interpretations of a smear from a fresh specimen compared with a smear prepared from the same specimen after 24 h of refrigeration. Differences were small and not statistically significant.

DISCUSSION

Difficulty in determining the etiological basis of a patient's pneumonia is a familiar problem and one usually first approached by examining a Gram-stained sputum smear. In current medical practice, direct sputum examination is used not only in the initial evaluation of patients with pneumonia but also to monitor the effects of antimicrobial therapy, to
monitor patients at high risk for developing pulmonary infection, and in the laboratory as a guide for choosing which sputum specimens to culture (2, 10). Although the value of sputum examination has been criticized by a number of workers (1, 9), sputum specimens continue to be collected from a variety of patients and contribute substantially to the workload of most clinical microbiology laboratories.

The evaluation of bacteria in sputum smears has traditionally consisted of two activities, identification and quantitation. Bartlett and colleagues reported that the identification of organisms can be accomplished with reasonable accuracy by technologists examining Gram-stained smears (3). In particular, they reported that the identification of staphylococci, "bacteroides/haemophilus," and bacteria of mixed morphology can be made with an average of 75% accuracy in high-quality sputum and wound specimens.

The present study was designed to examine how consistently technologists quantify bacteria in sputum smears. Quantitation of bacteria in lower respiratory tract specimens is difficult to study for the following reasons. (i) Stable control material that adequately mimics clinical specimens is not available. (ii) Performance is markedly influenced by the skill of the microscopist. Microscopists may vary in their internal consistency and in their accuracy relative to one another. (iii) Strict quantitation of cells and bacteria in smears is not generally practical, and the number of bacteria is usually assessed semiquantitatively. This create the potential for substantial agreement to occur simply by chance and requires that concordance be measured with a statistical method that takes the possibility of chance agreement into consideration. (iv) Finally, reproducibility in a research setting is apt to be significantly better than in clinical practice, since in practice the attention of a microscopist must be attuned to all of the elements present on a slide, while in a research setting concentration may be focused on one constituent at a time.

The principal finding of this study was that potentially pathogenic bacteria in sputum smears are not quantified consistently, even when partial credit is given for ratings that are in close but imperfect agreement. Poor reproducibility was due mainly to inability of the technologists to render consistent ratings when viewing the same slide twice. Since the accuracy of quantitation has been reported to be adequate for smears made from bacterial suspensions (3), it is possible that the discrepant ratings noted in this study were due to the uneven distribution of bacteria in smears made from clinical specimens. Differences between smears prepared from the same specimen and differences in the rating styles of individual technologists both tended to further reduce the consistency of interpretations. Reductions attributable to the latter two sources were not statistically significant, but the power of the study to detect small differences in reproducibility was low. In clinical practice, it is likely that all three of these factors act in concert to limit the reproducibility of quantitative estimates.

Specimen refrigeration was not found to significantly reduce the consistency of bacterial quantitation. The level of agreement between smears prepared before and after specimen refrigeration was only slightly lower than the level of agreement between two smears prepared from the same specimen before refrigeration. In a previous study, Penn and Silberman reported that bacteria seen in sputum smears prepared from refrigerated specimens did not correlate with bacteria seen in Gram-stained sputum smears prepared from the same specimen before refrigeration (11). Findings in the present report raise the possibility that some of the discrepancies attributed to refrigeration might better be ascribed to the intrinsically poor reproducibility of sputum Gram stain interpretation itself. This issue requires further investigation.

In a recent study of patients with community-acquired bacteremic pneumonia, high-quality sputum specimens with more than 10 organisms of a single morphotype per oil-immersion field were not found to be significantly better predictors of blood culture isolates than high-quality specimens with fewer than 10 organisms per field (8). Although the authors did not explicitly examine the consistency of quantitative ratings, their finding supports the view that quantitation of bacteria in sputum smears is unlikely to be of clinical value. Evaluation of bacteria should accordingly be limited to identification of the predominant bacterial morphotype or to reporting that the smear reveals a mixture of morphologies when a predominant organism is not present. The quantity of bacteria in sputum smears should not be reported, since it cannot be estimated reproducibly with conventional procedures.

### ACKNOWLEDGMENTS

Roger Grimson (Department of Community Medicine, State University of New York at Stony Brook) reviewed the statistical analyses and made a number of helpful suggestions. Weighted kappa statistics were computed by using a computer program written by Suh Wu, a member of the same department.

### LITERATURE CITED


2. Bartlett, R. C. 1974. Medical microbiology: quality, cost, and...

AUTHOR’S RETRACTION

Semiquantitation of Bacteria in Sputum Gram Stains

PAUL N. VALENSTEIN

Department of Pathology, State University of New York at Stony Brook, Stony Brook, New York 11794-8691, and Laboratory Services, Northport Veterans Administration Medical Center, Northport, New York 11768


Paul Valenstein
Laboratory Service
Veterans Administration Medical Center
Northport, New York 11768