JOURNAL OF CLINICAL MICROBIOLOGY, Aug. 1991, p. 1559–1561
0095-1137/91/081559-03$02.00/0

Vol. 29, No. 8

GUEST COMMENTARY

The editors of the Journal of Clinical Microbiology *are introducing in this issue the first in a series of "Guest Commentaries on Contemporary Issues in Clinical Microbiology." Clinical microbiologists have been asked to contribute commentaries on selected topics. The papers will be critically reviewed prior to publication. The topics include:*

- *Evaluating Biochemical Identification Systems*
- *Evaluating Antimicrobial Susceptibility Instrumentation*
- *Evaluating Immunologic Detection Methods*

These commentaries are not published as how-to papers but rather are written to raise issues and to stimulate consensus studies by groups such as the American Academy of Microbiology or other private and professional standards organizations.

The readership of the Journal of Clinical Microbiology *is invited to respond to these guest commentaries in the "Letters to the Editor" section or to make recommendations for additional topics.*

Richard C. Tilton
Editor-in-Chief
Journal of Clinical Microbiology

Evaluating Biochemical Identification Systems

J. MICHAEL MILLER

Hospital Infections Program, Building 1, Room B341, C01, Centers for Disease Control, Atlanta, Georgia 30333

Most clinical laboratories in the United States use either manual or automated systems to identify a growing variety of microorganisms. The majority of these laboratories are unable or unwilling to evaluate fully an identification system before purchasing it. They rely instead on reports in the literature published by other clinical or reference laboratories or on simple demonstrations or comparisons conducted in-house and often supervised by a manufacturer's representative. Because of the impact instrumentation has on bacterial identification in our laboratories and because of the potential influence exerted by published reports of instrument evaluations, it is crucial that these studies be planned and conducted appropriately and that the data be expressed clearly and fairly. A review of the literature on instrument evaluation will reveal that while many studies were conducted appropriately and the results clearly presented, many others were poorly conceived and the results easily misinterpreted. The comments and opinions in this Guest Commentary will address the important issue of instrument evaluation and the subsequent publication of results and, hopefully, will challenge investigators to reach a consensus on a number of complex and pressing questions that will be presented.

The motivation for publishing on system performance must be to evaluate objectively and carefully the stated claims of the manufacturer, not to get a fast, easy publication or to assure the investigators of an abstract and poster at an annual meeting. The relationship between the investigator and the manufacturer is often delicate because the manufacturer may be funding the study, although this does not automatically compromise the investigators' objectivity. As with the publication of any scientific research, poorly designed studies invariably lead to premature or misleading conclusions and, often, poorly written papers. Performance evaluations of diagnostic microbiology systems require no less attention to a sound research plan. If one examines a sample of the hundreds of articles in the literature on system performance, there clearly is no consensus on study methods or protocols.

While this author does not mean to imply that all or even most articles or posters on the subject are misleading, some certainly are. We need to recognize the extent of the positive and negative impact these system evaluations make on laboratory choices and to develop consensus guidelines for study protocols and common formats for data expression in the published literature. Similar guidelines have been established for studies of antimicrobial agents.

The cost of the newer systems is often high, especially for today's health-care laboratories that are under constant budgetary constraints. Laboratory directors and those responsible for the acquisition of diagnostic systems for the laboratory can ill afford to purchase an expensive system only to find that its performance does not meet the needs or the expectations of the laboratory. They withhold their resources until they are convinced that the system they are considering will perform as expected. For many, the primary sources of performance information are journals that publish the results of evaluations and trials. The companies marketing the instruments also are similarly dependent on the published results of the performance evaluation studies. Clearly, the published report is highly significant and often very influential.

Resolving the complex problems and arriving at a consensus about testing parameters is very important but, until now, these problems have gone unsolved. Their resolution will require the efforts and expertise of many. Some of the

more basic areas that beg for consensus are: How are we to define "correct response" and "incorrect response"? Should test errors be repeated, and if so, which answer is reported? How many strains should be tested in order to report on accuracy at the genus and species level? Against what standard should instruments be evaluated? Parity with an existing instrument does not necessarily make the new one accurate when compared with reference standards unless the existing instrument is 100% correct. Are there degrees of accuracy when the reported probability of the correct response one time is 98% (highly accurate) and another time 84% (somewhat accurate)?

There are numerous issues that such proposed guidelines might address. For example, an **evaluation** should use a clearly defined protocol designed to assess the ability of the system, reagent, or method to meet fully the claims of the manufacturer. These studies should provide a description of the patient population, advantages and disadvantages of the instrument, system, or method, and a report on whether the instrument performed as expected. A **comparison** is a controlled study contrasting one or more systems to determine whether they are predictably equal in accuracy and performance. Among the results should be a statement of accuracy comparing the test system with the chosen reference method, an explanation of how discrepancies were resolved, and statistical support for the accuracy statement. Because comparisons of two systems may ignore errors common to both, the true accuracy of a system should be judged with an evaluation and not with simple comparative studies.

The report must clarify the type of study to which the instrument or method was subjected. In our laboratories at the Centers for Disease Control (CDC), we subject test systems to two types of analyses, a "stress test" and a "weighted laboratory profile". The stress test (as it is evolving) utilizes large numbers of strains from our extensive stock culture collection, including many species not routinely isolated in most laboratories. This group of strains is designed to test the limits of accuracy of any instrument. The weighted profile provides a means to reassess the data by using the types and percentages of isolates likely to be found in most day-to-day laboratory work. Because we are a reference laboratory, fresh clinical isolates are not always available, so evaluations using clinical isolates are best done by other institutions that have access to patient specimens. These clinical studies also must be defined for the reader. The results from the stress test and the clinical isolates may be very different, but each testing method is important. If a clinical laboratory wishes to conduct a stress test of its own, it should be clearly noted in the published results. Consequently, we should determine when a "clinical isolate" becomes a "stock culture" and how many transfers of a stock strain are necessary before subjecting it to testing. In most cases, if an instrument or method performs well in the stress test, it does at least as well in the tests that use clinical isolates.

A list of the test strains is essential. This listing can be combined with the test results and put into one table as in the following example.

Organism	No. correct to species/ no. tested	No. correct to genus only
Escherichia coli	26/28	1
Escherichia hermanii	2/4	2

The readers of these published studies recognize that there is not a consensus on the definition of terms that are common and frequently used. For example, when testing the ability of an instrument to identify an organism, how should we define "correct response" and "incorrect response"? Writers may incorrectly assume that all interested readers agree on these definitions; however, even the simplest definitions often mask complex issues. At first glance, we may define correct response as the accurate genus and species. Clearly, for many organisms, that response will be true, but to identify *Salmonella* and *Shigella* to the species level by biochemical tests may be impractical and unnecessary for appropriate patient management. Thus, in some instances, the genus designation alone could be considered a correct response. Species identification of *Cedecea*, *Kluyvera*, and certain gram-negative members of the non-*Enterobacteriaceae* family may also be unnecessary. In another case, if the report reads "No identification," meaning that the probability was too low for the data base to report an organism, is this an incorrect response? Even if the manufacturer claims to be able to identify this isolate, the instrument could not or did not do it, but it also did not report the wrong organism. If the instrument fails to identify the organisms listed by the manufacturer, then that nonresponse should be an error. However, this result should be clearly distinguished in a separate category from other errors. If the genus is reported correctly but the species is incorrect, is this an incorrect response? In most cases, a genus only report is usually consistent with acceptable standards of patient care, if accurate antimicrobial susceptibility results accompany the answer. Few investigators, however, report on the "patient-care value" or "patient-care consequences" of the results from these identification systems.

Manufacturers may be doing everyone, including themselves, a disservice by broadening their data bases or increasing the system's sensitivity to such a degree that the accuracy of the system could be compromised by the attempt to fully identify rare biotypes or strains with little clinical relevance. I assume that most clinical microbiologists likely would be satisfied with a report of *Cedecea* sp. if the option of performing additional tests to identify to species level were offered. I am aware of the epidemiologic importance of species designations. Certainly, our programming colleagues can issue instructions for the computer to report in one format for clinical utility and another for epidemiologic purposes.

Many authors correctly retest organisms that initially were incorrectly identified by the system under investigation. One retests either to validate responses or to clarify the reason for the initial error (technical versus mechanical). Retesting is especially important in evaluations, but it can be helpful in comparisons as well. In the clinical laboratory, however, how often will we know that the instrument's acceptable response was incorrect? True errors, like accurate responses, should be reproducible. But if, on retesting, the correct result was reported, the initial error could be due either to technique or to the inability of the instrument to identify consistently this isolate. One more test would be needed to resolve that question (assuming the best two of three responses would be considered). Reporting the results of retesting may confuse the reader and requires careful wording. On the other hand, all diagnostic systems should be tested for reproducibility, an important consideration in the decision to purchase an expensive, automated system. Using known, stable strains, or even the recommended quality control strains for the test, may be helpful in documenting

predictably reproducible responses in multiple, consecutive tests.

Another issue of concern is the lack of statistical analysis applied to the data in comparative studies. When an instrument is being compared with another or with conventional testing, the results are either significantly different (meaning the two identification systems are not equal in that study) or are not significantly different (meaning that the two identification systems performed similarly). While a cursory look at "percent accuracy" may show results that appear relatively close, the subsequent conclusions may not be supported by a statistical analysis. Conclusions made on the basis of data that are not supported by statistical analysis may only be assumptions. Yet many studies on instrument evaluations omit statistical analyses of their data. Commercial companies could be adversely affected if their instrument received a bad report in the literature because of incorrect or incomplete analysis. Laboratories also could be hurt by misleading conclusions drawn from otherwise accurate data that prompted a purchase later regretted.

A clear example of incorrect assumptions can be taken from an article previously published in a reputable journal. In this article, two microbiology instruments were being compared. Instrument B was correct at the species level for 94.6% of the isolates, and instrument A was correct for 91.1%. The conclusion was that instrument A "compared favorably" with instrument B, implying that the two instruments were essentially equal in their ability to correctly identify isolates. If only 100 isolates had been tested, that conclusion would have been valid. Instead, almost 1,500 isolates had been tested, and by applying chi-square analysis, the results were significantly different ($P < 0.001$), meaning the two instruments were not equal in identifying isolates at the species level, a conclusion opposite to that published. Instrument B, in this study, was clearly better than instrument A.

An appropriate statistical test should be applied to all such published data. McNemar's test, in which one sample is split in order to test two different instruments, or the chi-square test, in which similar but independent samples (suspensions) are applied to two instruments, may be appropriate. Investigators should consult an expert in statistical analysis to determine the most appropriate analytical method.

It is time to call for technical and editorial guidelines. Recognition of this need is not new. Many microbiologists involved in instrument evaluations have either perceived the need or expressed it. Indeed, plans are presently under way to gather and quantify the information needed to develop guidelines for this type of testing and to bring order to the evaluation and assessment of the data. Manufacturers of diagnostic reagents, devices, and instruments would do well to form a coalition for support and funding of such a process, because they would benefit from the publication of clearly written and accurately presented evaluations. Working together, representatives from the scientific societies, instrument manufacturers, and selected clinical laboratories could establish guidelines for us to follow.

In the meantime, the basic components listed below should be included in every study that evaluates or compares microbiology identification instruments.

1. List and explain the basic definitions used in the study.
2. Test the system only within the latest claims of the manufacturer, i.e., the latest data base revisions and software updates available.
3. Clearly specify the standard against which the system is being evaluated. If another system is used, how does it compare to reference methods? Is the chosen reference system assumed to be 100% accurate? Are conventional biochemicals used as the gold standard?
4. Apply appropriate statistical analyses to the data before drawing conclusions.
5. Clarify whether the report represents a routine clinical trial or a more stringent study. Use a group of test organisms that represents the expected relative percentages of organisms routinely isolated and the degree to difficulty usually expected at the study site.
6. Keep the statement of accuracy as simple as possible.
7. If the results differ greatly from those published by others, offer possible reasons for the discrepancy.
8. Discuss the positive and negative aspects of the system, including cost per test, technologist time, etc.
9. Arbitrate discrepancies by a reference method.

These comments are not intended to offer or impose standards for test protocols, to prescribe a single method for reporting the results, or even to answer questions. They are presented as a challenge to recognize the importance of our publications and to accept the responsibility for the potential influence our work has on our colleagues and on industry. Let us pursue rational, accurate, and timely examinations of these issues and provide leadership and a mechanism that will address them.