Better Laboratory Evaluations of Instruments and Kits Are Required

Callum G. Fraser and Ronald Singer

There are many published protocols for evaluating instruments and reagent-kit sets, and a plethora of papers describe the results of evaluations performed. It might be assumed, therefore, that this facet of clinical chemistry has no problems, but examination of evaluations of instruments published in the first half of 1984 shows flaws in some aspects of experimental design and execution, statistical analysis, and interpretation of data. We believe that clinical chemists, manufacturers and suppliers, and journal editors and referees can all contribute to improvement of this situation, and we suggest some guidelines for each of these groups.

Zweig and Robertson (1) suggest that new diagnostic tests should be subjected to well-designed clinical trials before being accepted for use, because the conclusions reached in some published studies are of uncertain validity, commonsense principles having been overlooked. This situation may have arisen because there have been few theoretical publications on this topic by clinical chemists.

In contrast, a great many publications by individuals, national professional bodies, and international groups describe protocols for evaluating instruments and reagent-kit sets designed for use in the clinical chemistry laboratory (2) and, in addition, certain journals provide instructions to authors on the information suggested or required in submitted papers concerned with evaluations of new instruments and reagent-kit sets. Moreover, development, assessment, and evaluation of analytical techniques are substantive roles for many clinical chemists. Thus it might be assumed that this facet of the discipline is now without major problems.

However, we suggest that published evaluations are not totally satisfactory and we believe that better evaluations of instruments and reagent kit sets are required. To examine our hypothesis, we studied all the evaluations of instruments published in the first half of 1984 that were available to us (Table 1), and we report our views on performance characteristics, statistical analysis, and interpretation of results.

Performance Characteristics

Published protocols generally recommend that practicability, precision, accuracy, specificity, and detectability be assessed.

Practicability. In the current economic climate, the clinical chemist who is planning to purchase an item of equip-

ment must carefully consider details of speed, costs, technical skill requirements, dependability, and safety. Only one of the published evaluations, evaluation 2, provided clear and concise details of mechanical and microbiological safety, a detailed costing study, and documentation of faults detected during the study.

Precision. All the evaluations in Table 1 estimated precision. Problems do exist, however, and as a prime example, we draw attention to evaluations 1a and 1b, in which analyses of three "control solutions" were performed 10 times each day for 10 days. According to published recommendations (2), this number of between-day replicate analyses is unsatisfactory. It is also clear that aqueous solutions were used to assess precision and, because these instruments are designed to assay glucose in blood and not in aqueous solution, we believe that more appropriate experiments, from a matrix point of view, should have been designed and executed. A further concern is that, in evaluations 1a, 1b, and 2, the precisions found on different days were found to be highly significantly different when the simple F-test was applied; these differences were not commented upon by the evaluators.

Accuracy and specificity. This assessment is concerned with the difficult problem of bias and involves detailed study of calibration standards, linearity, recovery, mode of blanking, specificity, interference, comparison of methods, and analysis of quality-control and -assurance materials. A feature of all the published evaluations is that comprehensive studies to investigate bias were not performed, resulting, in our view, in inadequate data being generated for objective conclusions. As examples, comparison of results obtained with the instrument under evaluation and with methods routinely in use in the laboratories of the authors was the sole indicator of accuracy in evaluation 4. Comparison of only 40 patients' results with a routine method, linearity studies with use of dilutions with water, and examination of possible interference with heparin, EDTA, and fluoride/oxalate were carried out in evaluation 5. It was of interest that, in contrast to the recommendations of many published protocols, in none of the evaluations was analyti-
cal recovery assessed, and analysis of materials with consensus or well-assigned values was not a significant feature of any of the studies.

Detectability. Detectability—the ability of a method to detect small amounts of analyte—was not investigated in any of the evaluations.

Statistical Analysis of Results

Published protocols vary in their recommendations as to the statistical treatment of raw data. Method comparisons were performed in all the evaluations studied. The need for more-uniform criteria for statistical treatment is perhaps demonstrated by the published data in that, in evaluations 1a, 1b, and 5, the correlation coefficient and regression equation were given; in evaluation 2, only the regression equation; and, in evaluation 4, the correlation coefficient, regression equation, and $S_{xy}$. Only in evaluation 3 was a full range of statistical parameters given. In contrast to most recently published recommendations, simple linear regression was used rather than the more complex but more appropriate Deming method, which does assume that the comparative method has error (8). In contrast to most published recommendations, evaluators did not attempt to statistically compare the means found for test and comparative methods. We noted some flaws in the statistical analysis of results in the published evaluations. For example, in evaluations 1a and 1b, instrument 1a was alleged to be "more precise" than instrument 1b, even though application of the simple $t$-test to the published data showed that the precisions did not differ significantly at the usually accepted level ($p = 0.05$).

Performance Standards

In previously published work (9–11), we have detailed that most published protocols do not delineate analytical goals, these being the numerical performance standards that are required to provide optimal patient care. Where goals have been referred to, they are considered to be dated, because they are based upon criteria proposed by Tonks in 1963 (12) or Barnett in 1968 (13). Both of these criteria have been criticized (9). We have previously strongly advocated that evaluations must be considered to be unsatisfactory unless the results obtained have been objectively compared to current consensus analytical goals for as many performance characteristics as possible (9–11). The subjective nature of the analysis of the data generated in evaluations is well illustrated by those cited here. For example, in evaluations 1a and 1b, precision was said to be "accurate and precise enough," and both machines were considered to offer accuracy and precision "at least as good as many available systems." The precision of the analyzer assessed in evaluation 2 was said to be "adequate for normal routine use," although "the low bias with one instrument at low concentrations of glucose gave cause for concern." The results of evaluation 5 were purported to show that the instrument was "precise, accurate and linear." As detailed earlier (9–11), these subjective comments are not totally correct and, as one example, only in evaluation 4 does the precision of glucose analysis meet the current analytical goal based upon intra-individual biological variation (14) at all analyte concentrations (11).

Conclusion and Recommendations

We believe this assessment of recently published work amply demonstrates that published evaluations are generally not totally satisfactory. The apparent shortcomings apply to all aspects of evaluation: experimental design, execution, statistical analysis, and assessment of results for their clinical value. We believe that our findings apply equally to evaluations of reagent-kit sets.

We suggest that a number of valuable lessons can be drawn from this analysis and we propose the following guidelines.

For clinical chemists. It must be more widely recognized that evaluation of instruments and diagnostic reagent-kit sets requires both skills and commitments to this facet of the discipline. An evaluation is a complex process, not to be lightly undertaken, because the laboratory must have sufficient staff, time, space, financial, and other necessary resources to do it properly. We have recently provided (15) a comprehensive but simple-to-follow guide for evaluators, which amply demonstrates the substantial efforts required. If evaluations are undertaken, surely a well-accepted published protocol should be followed and all aspects of the method investigated. We believe that analysis and interpretation of the results obtained are particularly unsatisfactory at present, and that much more effort is required in this area. Surely it is not effective to expend significant effort and resources in obtaining numerical experimental results without then giving them the appropriate statistical and other data analysis.

For manufacturers and suppliers. Instruments and reagent-kit set evaluations are often initiated by manufacturers and suppliers. Many laboratories request loans and (or) gifts of these and many will not make a firm purchase unless there has been a hands-on experience in the laboratory. It is perhaps pertinent to question whether this latter practice has come about because current published evaluations are rarely complete or comprehensive and the laboratory cannot obtain sufficient information from the indexed literature.

We suggest that manufacturers and suppliers should much more carefully select laboratories that are to perform evaluations. Ideally, such laboratories should have a proven track record in this facet of the discipline and should have sufficient resources to ensure that an excellent publication ensues, one that can satisfy even the most critical chemist. Moreover, it is usual—and indeed we consider it ethical—for evaluators who have been supported in their evaluation, either in whole or in part, by manufacturers or suppliers to submit the proposed publication to them for comment before it is dispatched to a journal. We advocate that manufacturers and suppliers be more critical at this stage, and we suggest that external, well-qualified clinical chemists be commissioned as referees should real in-house expertise be unavailable. We also suggest that manufacturers and suppliers insist on full protocols being made available to them by the laboratory before an instrument or reagent kit is provided to it for evaluation.

For journal editors and referees. The deficiencies of current published evaluations surely reflect badly on the standard of referees or reviewers of submitted manuscripts. We urge journal editors to select such individuals with care, attempting to choose clinical chemists known to have expertise in their field and known to be willing and able to provide a detailed and comprehensive review, complete with constructive criticism. We urge referees to consider carefully the main thrusts of this paper before agreeing to the publication of unsatisfactory evaluations.

Many of the points made in this paper have been stated previously in one or more of the published evaluation protocols. We hope this paper will re-stimulate interest in this aspect of clinical chemistry and dispel some of the complacency about evaluations that we believe currently exists in our profession.
References


Ed. note: Two persons asked to comment on this opinion have done so:

Dr. D. B. Tonks (Director of Clinical Chemistry, Hôpital General de Montreal): In principle, I must agree with this article, but in practice I would not follow such a rigid protocol unless (perhaps) I was evaluating a kit for general (outside) purposes. If I was going to use a kit in my own laboratory, I would evaluate each new set of results (comparisons) and decide after each run what I would do to follow up on it. I would not use professional intuition, gained after many years of experience.

As a matter of fact, I collect all of the NCCLS protocols and those of other bodies, but use them only for references and do not follow them in detail. Perhaps this is wrong, but I look at them for ideas rather than for procedural details to follow.

Dr. Basil T. Doumas (Medical College of Wisconsin, Milwaukee): Drs. Fraser and Singer submit that some evaluations on clinical instruments and reagent kits are less than satisfactory. Their criticisms, based on a rather small sample of publications (I wonder whether it is statistically significant), do not discriminate between good and bad.

Because I have no access to references 3–5, my comments will apply to references 6 and 7.

Reference 7, a Letter to the Editor, deals with a urea analyzer. It would be unreasonable to expect a very thorough evaluation and a full-length paper on a minor piece of equipment; no journal would publish it. I believe, however, that the information provided is useful to small laboratories which may wish to use this instrument. I do not understand why it is considered unacceptable to dilute specimens containing high urea concentrations with water.

The critique on evaluation 4 (ref. 6) is superficial, if not "picky." Example: "comparison of results obtained with the instrument under evaluation and with methods routinely in use in the laboratories of the authors was the sole indicator of accuracy." To begin with, the authors of ref. 6 made no statement regarding the accuracy of the methods adapted to the Technicon RA-1000; indeed, the word "accuracy" never appears in the text.

The statistical analysis was criticized because a full range of statistical parameters was not provided, means were not compared statistically, and the F-test was not applied to the estimates of precision. I think that we usually get enough statistical parameters to cause indigestion (some members of the editorial board complain that Clinical Chemistry has become the journal of("correlation and regression").

Missing in ref. 6 are the standard errors for the slopes and intercepts, but the $S_{xy}$ provides an estimate of the scatter of the data. Comparison of mean values is usually not very rewarding. If the comparison is done by the regular t-test (pooled variances), a significant difference may be missed because of the wide variation of analyte concentrations in the samples. On the other hand, the paired t-test is often misused for comparing means of such a large number of observations. The authors must be aware that a significant difference may be obtained by this test, even when it does not exist. A small bias between two methods may show a medically irrelevant significant difference. Application of the F-test to standard deviations would have added a few more tables to ref. 6. The large number (8) of control sera and methods creates a plethora of F-ratios, some of which are bound to be statistically significant.

Apart from satisfying statistical curiosity, what else can anyone do with such "important" findings? It is unrealistic to expect the same precision with all types of control materials; some are less turbid and more stable and homogeneous than others.

Recovery studies. These in most cases are nothing more than another test for linearity. If the amount of added analyte does not exceed the linear range of the method, the expected good-to-excellent recovery does not provide any direct information about the accuracy of the method. Recovery experiments have a place in methods requiring isolation steps—i.e., fractionation, extraction, etc.—and in RIA procedures.

Accuracy. This will remain elusive for the near and, perhaps, distant future. Even reference methods cannot claim absolute accuracy because in transferrability tests the materials (serum pools) analyzed derive from healthy individuals who are not expected to ingest a variety of drugs. Accuracy can only be established by extensive interference studies which are too costly and not to be included in a method. It is unreasonable to expect that all available drugs and foodstuffs be tested with every method available. By necessity such studies are limited to a few endogenous substances, hemoglobin, bilirubin, lipids, and to drugs selected on the basis of being likely to interfere.

At this point, it is tempting to play the devil's advocate. In the past, has a physician ever missed a diagnosis of diabetes because glucose was measured by the Folin–Wu method?

Cost analysis. This is not always a straightforward exercise, and it perhaps is better left to the individual users. Cost per test not only depends on the type of the instrument, but also on how the instrument is used. If specimens are batched, the cost will be considerably lower than for continuous 24 h/day operation.
Consensus materials with well-assigned values. Even when such materials are easily available, their usefulness in the evaluation of instruments and methods is limited.

For example, we have a very accurately characterized protein standard solution (SRM no. 927; National Bureau of Standards, Washington, DC), which, if properly used, can provide comparable (among laboratories) values for serum total protein. (By "properly" I mean that the biuret reaction is allowed to go to completion.) However, SRM 927, a bovine serum albumin solution, is not suitable for use as a calibrator for the biuret method in instruments involving short reaction times (from a few seconds to a few minutes), because, in the early phase of the reaction, bovine albumin reacts faster than human albumin.

Primary values for analytes are assigned to such materials by manual methods that are not the same as those adapted to automated analyzers. Because, quite often, different methods do not measure exactly the "same thing," we resort to Procrustean tactics; that is, changing (fudging) the primary assigned values of calibrators for the purpose of obtaining "acceptable" results for patients' samples. In this case, "acceptable" means no change in the established normal range. The question is: why use such costly materials when we pay no attention to the assigned values?

Analytical goals. Such goals, usually set by well-intended experts, are easier to prescribe than to achieve. The cost of accomplishing some of the perfectionists' goals is not trivial, and it is doubtful that the benefit to the patient is proportional to the expenditure. It is important that some thought be given to clinical requirements and medical relevance before such recommendations are issued.

The most recent set of analytical goals comes from the 1976 CAP Aspen Conference (ref. 14). Desirable CVs for some of the analytes considered at that conference are:

<table>
<thead>
<tr>
<th></th>
<th>Conc.</th>
<th>SD</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albumin, g/L</td>
<td>35</td>
<td>0.5</td>
<td>1.4%</td>
</tr>
<tr>
<td>Alk. phosphatase, U/L</td>
<td>60</td>
<td>1.1</td>
<td>1.8%</td>
</tr>
<tr>
<td>Creatinine, mg/L</td>
<td>10</td>
<td>0.22</td>
<td>2.2%</td>
</tr>
<tr>
<td>Sodium, mmol/L</td>
<td>140</td>
<td>0.56</td>
<td>0.4%</td>
</tr>
</tbody>
</table>

Although I am not against good precision (it would amount to being against God, motherhood, and apple pie), I would like to ask the question: once we've got such precise results, what are we or the physician to do with them?

Protocols, recommendations, etc. These tend to multiply like . . . rabbits. Local, national, and international organizations are producing them with ever-increasing frequency.

The fact that they have been duly approved by member countries or organizations does not equate them with the Ten Commandments; they are not binding on evaluators. A protocol may be suitable for one type of instrument but not for another. Example: the PSEP-4 protocol (National Committee for Clinical Laboratory Standards) is ideally suited for the Technicon SMA 12/60 or the SMA (it was, more likely, designed for them). The cost, however, for evaluating with this protocol the 40-some channels of the DuPont acc would be prohibitive. It would be equally costly to apply this protocol to a centrifugal analyzer. Protocols should provide general guidelines (principles) for evaluations, and the details left to the evaluator. Strict adherence to such documents stifles innovation and leaves no room for common sense. [Ed. note: This reminds one of the story about the research scientist who was accused of having no common sense. He responded to the effect that he was not one of those few so gifted by inheritance, and only had a formal education to rely on. It may be worth reminding our readers that NCCLS lists four "Evaluation Protocols," EP2-T, EP3-T, EP4-T, and EP5-T, in their last (July 1, 1984) list of publications.]