J Clin Pathol 1981;**34**:947-957

External quality assessment of clinical laboratories in the United Kingdom

TP WHITEHEAD, FP WOODFORD

From the Wolfson Research Laboratories, Queen Elizabeth Medical Centre, Birmingham B15 2TH and the Department of Health and Social Security, Hannibal House, London SE1 6TE

SUMMARY A review is given of the National External Quality Assessment Schemes (NEQASs) in various pathology disciplines in the United Kingdom, with a discussion of the relative roles of the DHSS, individual laboratory scientists, and the relevant professional bodies. Principles of operation and scientific problems in the design of NEQASs in different disciplines are described and contrasted, and some comparisons with the experience in other European countries and the USA are drawn.

At a working group convened by the WHO Regional Office for Europe in December 1979 we had the opportunity to consider in detail with colleagues from several European countries the arrangements for assessing the quality of clinical laboratory performance in countries with different health care systems. The arrangements differ in interesting ways both between countries and for different pathology disciplines. In many ways the United Kingdom has led the way in introducing national external quality assessment systems (NEQASs) and in probing the possibilities of extending them into the more qualitative and interpretative disciplines such as histopathology.

The report[1] of the working group contains recommendations for countries which have as yet no national systems for the quality assessment of clinical laboratories, and addresses the crucial question of whether the main responsibility for quality assessment lies with governmental regulatory agencies or with the relevant professional societies. The report concludes quite firmly that because governments are committed to providing optimal health care for their citizens, they are obliged to ensure that clinical laboratories are properly assessed, but that they should entrust to competent professionals the task of providing a scientifically valid method of assessment.

The authors are chairman and secretary, respectively, of the Advisory Committee on Assessment of Laboratory Standards (formerly LDAG Standards and Quality Control Sub-Group), an advisory body to the DHSS.

Accepted for publication 25 March 1981

Terminology

Readers may be surprised by the absence of the term "quality control" in the discussion so far, especially as one of us (TPW) has long advocated this term in clinical chemistry to denote what we now prefer to call "external quality assessment." It has become clear that we need to use the term *control* more carefully and to differentiate between "internal quality control," which is applied within the laboratory to determine whether the results should be released and therefore constitutes a real example of control in the conventional sense of the word, and "external quality assessment," which provides a retrospective check on laboratory results but in no way controls laboratory output at the time the tests were performed. Use of the composite term "internal and external quality control" has led some laboratory workers to the mistaken conclusion that use of internal quality control obviates the necessity for "external quality control" and vice versa. The two procedures are, of course, complementary: internal quality control primarily monitors day-to-day reproducibility—that is, precision, and detects frank errors in any one day's procedures, while external quality assessment primarily aims at detecting constant differences ("bias") between one laboratory's results and those of others. The two procedures can be defined as follows:

Internal quality control (IQC) is the set of procedures undertaken by the staff of a laboratory for the continuous monitoring of operations and results, in order to decide whether the results are reliable enough to be released.

External quality assessment (EQA) refers to a system of objectively checking laboratory results by means of an external agency. It includes comparison of a laboratory's results at intervals with those of other laboratories. The main object is to establish between-laboratory comparability.

Coupled to IQC and EQA procedures, which are concerned with the quality of laboratory procedures, there should be a further series of monitoring, education and training systems concerned not only with specimen examination but also with specimen collection, transport, and handling and with methods for the reporting of results. The sum of all these procedures is conveniently called a "quality assurance programme" which term does not imply that perfectly reliable results can be "assured" at all times, but that a systematic effort is being made towards achieving the highest possible quality of performance in patient investigation.

It should be noted—and is too often forgotten by those who resent or resist the assessment of laboratory performance—that there are several components of inadequate performance: equipment fallibility, reagent fallibility (including that of calibration materials) and operator fallibility. A most valuable outcome of EQA is the detection of instrument and reagent faults and the feedback of such information to manufacturers for correction.

Formal organisation of National External Quality Assessment Schemes (NEQASs)

In virtually all the European countries which have national EQA systems for clinical laboratory work, the schemes operate through some kind of co-operation between the government and groups of professional scientists. The governments usually act through their Department of Health, but the arrangements may also include compliance with standards set by the Department of Industry. For example, NEQASs were developed quickly in the Federal Republic of Germany by groups of scientists (under the aegis of the Academy of Medicine) in response to the so-called Calibration Law, which would otherwise have made all laboratory equipment subject to periodic testing by engineers unacquainted with the equipment's purpose and clinical function. Governments in all countries with active EQA schemes have accepted the principle that only experienced laboratory practitioners can design acceptable survey programmes. Indeed, most large-scale programmes have grown out of unofficial schemes launched by individual scientists or by scientists acting on behalf of professional societies or academies. But in many Western European countries, the Government has given impetus to the programmes in the last few years by imposing legal or financial sanctions on laboratories which virtually oblige them to submit to external proficiency assessment. In the UK, participation in NEQASs remains voluntary, but the government encourages participation by bearing the whole cost of the schemes so that participation by NHS clinical laboratories is free. It is perhaps arguable that participation in NEQASs ought to be mandatory in the UK for laboratories that issue results that affect the nation's health—whether through patient care, environmental health surveys or clinical research. This point will be considered later.

Strategies for bringing about improvement in laboratory performance

The ultimate objective in quality assurance programmes is improvement in analytical performance in support of improved patient care. To this end, EQA data may be used either for educational purposes or as a basis for the application of legal, financial or professional sanctions on laboratories showing unsatisfactory performance. In the UK, sanctions have been studiously avoided, and the main thrust has been towards self-education on the part of NEQAS participants through the consideration of their own results and of the interpretations offered by the organisers.

A considerable body of evidence indicates that if EQA schemes are to be effective in influencing participating laboratories to bring their own performance into line with that of others, analyses of survey results need to be returned to participants rapidly. Organisers of UK schemes have made this an important feature of their mode of operation, and most UK NEQASs report every 2-3 weeks, as opposed to the 3-4 months that is common elsewhere.

Professional societies in Britain have, with the encouragement of the DHSS, provided two NEQAS-linked aids to the continuing quality assurance programme. A careful study[2] by a team nominated by three UK professional bodies has identified laboratory factors which correlate with the performance scores obtained by the laboratories in two major EQA schemes (one of them commercial) for clinical chemistry. Practitioners in other disciplines can benefit from the report of this study, because of the finding that a strikingly large contribution to good performance is made by sound laboratory management, efficient laboratory organisation and high staff morale. If these factors loom large in such a highly mechanised discipline as clinical chemistry it would be surprising if they were not equally or more important in the pathology disciplines that are even more dependent on technical skill. The other contribution of the professional bodies has been the formation of Advisory Panels, which help laboratories that persistently yield poor EQA results to identify and rectify deficiencies. This help is offered in the strictest confidence, and the identity of the laboratories in question is never revealed to any health authority; in most schemes it remains unknown even to the scheme organiser. Participants in some endocrinology EQA schemes are, however, so conscious of the need for outside help and support that they have agreed to reveal their code numbers not only to the organiser but to all other participants. It is interesting that laboratories engaged in routine investigations, which in general are much less susceptible to failure than (for example) sensitive immunoassays, guard their identity so much more jealously.

The importance of keeping EQA results strictly confidential within the profession is unquestioned. However, the second principle on which the NEQAS system was erected, namely voluntary participation, becomes of dubious application once confidence in the scientific validity of the schemes has been established. Participation in the DHSS-sponsored schemes is free, the extra work load imposed by participation is usually negligible, and the worst possible consequence of participation is the identification of deficiencies that might otherwise pass unnoticed until they precipitated a mishap or crisis in patient management. Recognising that the voluntary participation constitutes perhaps the weakest structural feature in the UK national quality assurance programme, the Royal College of Pathologists has imposed its only sanction with regard to EQA, namely that participation in the relevant national scheme(s) is a necessary condition for approval of a laboratory for training of clinical laboratory staff.

Scientific basis of EQA

Ideally, assessment of a laboratory's performance would be based on its total output of results and would take into account the correctness of the result and its interpretation for every investigation, the speed of response, the absence of administrative and clerical errors, the completeness and good order of the laboratory records, and so on. Clearly, only a small fraction of these features can be assessed externally; some of them, including the aspect of day-to-day consistency, will always be better tackled by internal monitoring. The major objective of EQA is to evaluate between-laboratory comparability, and the main technique that has been adopted is the *survey* based on distributed specimens that mimic patients' specimens. Alternative techniques (pattern analysis of a laboratory's total output of results, or the use of inspectorates) have some theoretical advantages but are either not feasible for large numbers of laboratories, or inordinately expensive, or both.

If EQA is to be professionally acceptable as giving a fair reflection of day-to-day laboratory performance, several prior conditions must be met. First, the material distributed must be demonstrably homogeneous, stable during transit, and similar in properties to the corresponding patient specimens normally examined. These requirements may be mutually incompatible—for example, lyophilised serum must often be used to ensure stability of certain chemical substances or micro-organisms; the specimen then patently does not resemble a patient specimen, and will probably show similar properties to a patient specimen only if reconstitution procedures are meticulously specified and scrupulously observed. Stabilised blood preparations may behave in cell counters differently from fresh anticoagulated blood; and the requirement for homogeneity simply cannot be met for large numbers of tissue sections or blood films. Secondly, laboratories must agree to treat the survey specimens as nearly as possible in the same manner as patient specimens are handled. If they do not, or if they consult other laboratories before returning the result, the value of the whole operation in providing informative feedback to the laboratory that is relevant to its average performance is lost. Lastly, the reference point to which each laboratory's result is compared must be acceptable either as a good approximation to the absolute truth or as a result which it would be reasonable for laboratories to strive towards. The question of what constitutes such a reference point ("comparison value", "comparison result") and how it is arrived at is worth considering in more detail.

Selection or calculation of the point of reference in EQA surveys

The comparison value or comparison result is derived in one of two fundamentally different ways, one relying on the results of reference laboratories and the other on the consensus of all participants. The choice of approach may be limited by the nature of the investigation: the presence of a bacterium, virus or blood-group antigen in the distributed specimen is vouched for by the reference laboratory providing the specimen, and participants must acknowledge that failure to detect it is necessarily an incorrect result. Nevertheless, the reference result is checked by consensus since instability of the specimen might have made the substance undetectable by the time the specimen was received. In the EQA of other qualitative investigations—for example, cervical or sputum cytology, consensus is virtually the only available means of arriving at a generally acceptable reference point, unless participants are willing to accept the judgement of a single "expert" or of an expert group as superior to their own. However, in the quantitative assays undertaken in haematology and clinical chemistry there is a real choice: in clinical chemistry, for example, the Federal Republic of Germany's NEQAS has adopted the "reference laboratory" approach[3] whereas the UK scheme[4] relates participants' results to consensus means. Since the results of chemical determinations in biological fluids are unfortunately often method-dependent, both schemes have to specify *method-related* reference values or consensus means.

In the "reference laboratory" approach, between 3 and 10 laboratories repeatedly assay the specimen over a number of days and the results are pooled, with suitable statistical treatment to exclude outliers and to identify any systematic bias on the part of any of the reference laboratories. The main problems with this approach reside in the initial selection of reference laboratories and the lack of assurance that their performance will continue indefinitely at a high level. An unnoticed bias in the performance of one laboratory in a small group will disproportionately affect the reference result; enlarging the group to compensate for this effect can quickly involve the organiser in great expense in arriving at the reference value.

With the "consensus" technique, no extra expense is involved: the mean of all participants' results (grouped by method, if necessary), suitably corrected by the removal of outliers[5] is taken as the "correct result." The advantages of the consensus mean are that it is less influenced by individual aberrant results (because of the large number of participants) and that it reflects average performance in laboratories using current routine methodology. For the few analytes for which definitive assays are available (serum concentrations of ions and small molecules) and for which direct comparisons have been made for particular specimens, the consensus means from the 400 participants in the UK clinical chemistry NEQAS are in remarkably good agreement with the definitive values (Table 1). There is no logical reason for this agreement, but it is a comforting fact. Also of interest is the agreement between national consensus means for the same material in two countries (Table 2).

Table 1 *Comparison of definitive and consensus values for lyophilised bovine serum (WHO tentative reference serum lot 77/1). (Values are in mmol/l.)*

Substance	*1977*	
	Definitive value	*UK consensus value*
Sodium	142·24	142·25
Potassium	4·79	4·82
Calcium	2·27	2·30 2·28 2·23 *
Chloride	100·2	99·5
Glucose	5·40	5·53 5·45 *
Uric acid	0·273	0·277
Urea	5·76	5·79

*Consensus means for different commonly used methods.

Table 2 *Consensus means (all methods) for lyophilised bovine serum (Armtrol Lot 489) distributed in UK and Netherlands NEQASs*

Analyte	*UK (July 1980)*	*Netherlands (Dec 1979)*
Na (mmol/l)	141·4 ± 1·7*	141·9 ± 1·6
K (mmol/l)	4·38 ± 0·11	4·40 ± 0·09
Cl (mmol/l)	101·1 ± 2·3	101·1 ± 2·3
Glucose (mmol/l)	5·0 ± 0·4	5·57 ± 0·33
Ca (mmol/l)	2·44 ± 0·09	2·39 ± 0·09
Phosphate (mmol/l)	1·64 ± 0·09	1·67 ± 0·09
Fe (μmol/l)	32·2 ± 3·5	32·2 ± 4·5
Urate (mmol/l)	0·297 ± 0·021	0·296 ± 0·033
Creatinine (μmol/l)	119 ± 10	113·7 ± 9·8
Bilirubin (μmol/l)	23·6 ± 4·6	22·5 ± 3·0
Total protein (g/l)	70·8 ± 2·1	71·1 ± 2·3
Albumin (g/l)	34·9 ± 2·3	35·0 ± 3·4
Cholesterol (mmol/l)	4·79 ± 0·31	4·99 ± 0·39

*Mean ± SD (n = 225-382, UK; 232-402, Netherlands).

The disadvantage of the consensus mean is that it may not represent the results obtainable by the best currently available techniques, which although accessible in principle may nevertheless actually be used only by a minority of participants because of the expense of acquiring new equipment. In general, consensus techniques are satisfactory if a large enough number of laboratories is participating and the investigations involved are well established and standardised, eg determinations of blood haemoglobin and serum cation concentrations. For small numbers of laboratories and methods that are still in the developmental stage, the consensus mean is a more questionable point of reference.

EQA design in different disciplines

GENERAL

The laboratory discipline for which national EQA schemes are most widespread in Europe is general clinical chemistry—that is, clinical biochemistry excluding endocrinology, toxicology, or therapeutic drug monitoring. This is partly because participants' results on distributed specimens are quantitative and can therefore be readily handled mathematically in terms of their deviation from a reference point. A basic assumption in most schemes is that the results of frequent surveys of 15-30 commonly performed tests can be combined to yield an overall score which relates reasonably well to total laboratory performance. This principle,[6] though difficult to defend theoretically, has nevertheless been widely accepted. Without such acceptance it would have been impossible to make a start on assessing routine laboratory performance. Clinical chemistry is also the discipline for which it is easiest to produce large numbers of specimens that sufficiently resemble patient specimens and which are sufficiently stable to survive postal distribution.

Several European countries have also introduced NEQASs for haematology, including coagulation testing and to a lesser extent immunohaematology (blood grouping). Some countries also have NEQASs for microbiology (mainly bacteriology, but with some extensions into virology and parasitology), although Finland and Switzerland and to a lesser extent Sweden have simply opted for participation in the UK scheme. Endocrinology, clinical pharmacology, and toxicology are rather less well developed for EQA in continental Europe, except in the Federal Republic of Germany. EQA schemes for immunology, histopathology and cytology, where they exist at all, are still in the experimental stage and are virtually confined to the UK. Features of EQA peculiar to each discipline are described in the following sections.

CLINICAL CHEMISTRY

The principle usually employed is to calculate some index of performance on the basis of survey results. In the UK, this is called a "variance index score" (VIS), calculated by taking the difference between each participant's reported value for a given analyte and the consensus mean, expressing this difference as a percentage of the consensus mean, and expressing this result as a percentage of a "chosen coefficient of variation" (CCV). The CCV is different for each analyte: it is in fact the coefficient of variation observed for that analyte during NEQAS surveys in 1971. The VIS represents a composite of bias from the target value (namely, the systematic error, if any) and the day-to-day imprecision, and in itself gives no information as to the relative contributions of these two classes of error. Comparison with a sufficient number of other recent results in the same laboratory can, however, suggest whether performance of that assay is generally improving or deteriorating, and comparison of the result with the total spread of participants' results (provided by the organiser as an overall coefficient of variation) can also reveal gross errors or "blunders," either in transcription or specimen handling, which should be investigated without delay.

Calculation of the VIS reduces the result for each analyte to a common currency—namely, the level of performance usually attained in 1971—so that the results for all analytes can be averaged to give a *mean* VIS. This figure represents the laboratory's performance over the range of assays surveyed and hence, by extrapolation on the assumption mentioned above, the laboratory's performance in general. Happily, VIS's in most UK laboratories are now well below 100, with an average of 70, which means that the spread of results for common assays has been reduced to 70% of what it was in 1971. Finally, because performance varies (sometimes inexplicably) over time, and particularly with the nature of the specimen distributed, oscillations in the mean VIS with each distribution are damped by calculation of an "overall mean running variance index score" (OMRVIS) which is the average of the 40 most recent VISs for all analytes assayed by that laboratory. Since specimens are distributed every two weeks and each laboratory assays up to 14 analytes in each specimen, the OMRVIS averages the performance over the past 6-8 weeks. One theoretical problem which plagues EQA organisers is how to devise systems of calculation that will enable laboratory directors to discern long-term trends undistorted by transient errors, without disguising or concealing those errors. One method might be to present the highly discrepant result with full emphasis when it first occurs, but to calculate cumulative performance indices using exponential weighting so that the effect of events in the more distant past disappears more rapidly than when equal weighting is used (MJR Healy, personal communication 1980). Such a technique would be more valuable for a NEQAS covering a single analyte—for example, blood lead or a hormone—and with less frequent distributions than in the UK general clinical chemistry NEQAS.

Accusations of inappropriateness of the specimen are invariably raised by participants with the most discrepant results, and can usually be completely answered by reference to the tightness of distribution of other participants' results about the method mean. There is no doubt, however, that certain kinds of specimen are favourable to particular kinds of apparatus, and the organiser therefore needs to take care to distribute a wide variety of specimens (human or animal-derived, at both normal and pathological levels of analyte). The computer program for analysing the results must also contain routines to alert the organiser to an especially wide spread of results for a particular specimen or analyte or an especially low (or high) analyte concentration at which high accuracy is not needed in clinical practice and for which participants should not be penalised for obtaining mathematically poor results. One danger in clinical chemistry EQA is to let statistical procedures dominate the picture to the exclusion of clinical common sense.

HAEMATOLOGY

The field of blood cell counting is dominated by electronic cell counters which are calibrated, ultimately, by reference to manual methods performed under stringently defined conditions. Since virtually all laboratories use such counters it is more reasonable to use a consensus mean as reference point in EQA than to attempt to perform reference assays on each short-lived specimen that is distributed, in view of the strong desirability of returning results promptly to participants. The equipment also calculates a packed cell volume (PCV), which is checked by the manufacturers against a relatively complex reference method (no definitive method is possible since the PCV depends on arbitrarily defined conditions for packing the cells). For these measurements there is as yet no satisfactory stable primary reference material with assigned values. For haemoglobin estimation, a primary reference material and a definitive method of assay do exist, but again the consensus mean is more convenient to use.

The EQA of haemocytometry and haemoglobinometry[7] has as its major problem the provision of sufficient amounts of a whole blood preparation that is stable enough to survive postal distribution in good condition. This problem has been largely solved in the UK by the use of anticoagulated equine as well as human blood. Stabilised (fixed) leucocytes are added to some specimens, and donkey blood is used in many distributions to simulate abnormal human blood. In the calculation of performance indices, allowance is made for any deterioration of specimens during distribution that is apparent from the results received. Besides the full blood count, surveys are also frequently made of platelet and reticulocyte counts and of biochemical assays that form part of haematological investigations, eg serum iron, transferrin, vitamin B_{12} and folate, red-cell glucose-6-phosphate dehydrogenase and the percentage of minor haemoglobins (HbA_2 and HbF). Stained blood films for morphological examination are also circulated; participants perform a differential cell count and return numerical results, which are statistically analysed with reference to consensus means. However, because of the difficulty of providing large numbers of strictly comparable specimens and some doubt as to the validity of making comparisons of performance on this basis, the results of such surveys are not at present included in the calculation of over-all performance scores.

EQA in blood coagulation testing[8] involves four major tests: prothrombin time, activated partial thromboplastin time (APTT; including the evaluation of the anticoagulant effect of known amounts of heparin), fibrinogen concentration and Factor VIII estimation. In the UK, where manual coagulation testing is the rule rather than the exception, the wide dispersion of results for prothrombin time and APTT that was observable between laboratories in the 1960s has been narrowed by the central provision of lyophilised calibration materials (reference plasma), reagents (standardised thromboplastin and APTT reagent) and recommended procedures. The need for and the efficacy of these measures was demonstrated in the early 1970s by three-monthly surveys using lyophilised plasma pools from patients with naturally occurring or naturally acquired coagulation defects. In countries such as the USA with thousands rather than hundreds of laboratories which employ a wide variety of reagents and mostly automated instruments, EQA schemes have tended to use as test samples plasmas artifically depleted of one or other coagulation factor. Organisers of the UK EQAS consider such plasmas not sufficiently similar in behaviour to fresh plasma from patients with coagulation defects.

The prothrombin time and APTT are values defined by the conditions of assay and the reference plasma; consensus means have been used as the comparison point, although in principle a reference value based on reference methodology could have been provided. The National (UK) Reference Laboratory for Anticoagulant Reagents and Control organises the UK scheme; a concurrent international scheme, sponsored by the International Committee for Standardisation in Haematology, includes a small number of laboratories in each of 70 countries.

Two major factors limit the frequency of specimen distribution and therefore the effectiveness of such an EQA scheme: (a) the need for stringent quality control of the chemical composition, between-vial homogeneity, long-term stability and above all coagulation behaviour in the testing systems of the specimens to be distributed, and (b) the availability of sufficient quantities of suitable patients' plasma. The hope that plasmapheresis of patients with coagulation defects would provide large volumes of suitable test material has not been borne out by the results obtained on such plasmas in the UK scheme to date.

Despite these problems, EQA of coagulation testing in the USA has demonstrated, and in many cases clearly caused, dramatic improvements in between-laboratory agreement. In the UK, discrepancies between laboratories were not as apparent at the start of the scheme in 1970 because of the high degree of standardisation already brought about. The UK NEQAS has been particularly useful for the comparative assessment of commercial reagents and instruments because of the existence of a widespread familiarity with a standardised set of reagents and recommended procedures which participating laboratories can use alongside the reagent or instrument system that is to be tested.

In the EQA of blood group serology, one problem is the limited amount of appropriate blood available, although this problem has perhaps been overemphasised.[9] More crucial may be the provision of stable survey specimens, since deterioration in the specimen has been adduced to explain the immensely wide variation in sensitivity displayed by different laboratories to the presence of irregular antibodies. Aliquots of pooled sera and cell samples are distributed in the British scheme by means of Blood Transfusion Service vans designed for the urgent transport of refrigerated blood. The difficulties that some laboratories evidently experience with cross-matching, even of the ABO blood groups, as revealed by NEQASs in Britain and in France (for example), point to the need for more frequent surveys and faster feedback of results to participants and this is being arranged.

Common to all EQA schemes which survey manual testing procedures is the problem that the distributed specimen is usually examined by a single operator, possibly not the one responsible for a large proportion of patient specimens, and under non-routine conditions. (In haematology, this applies more obviously to blood coagulation testing, blood cell morphology and cross-matching procedures than to blood cell counting.) This constitutes a major weakness of the survey method, and the head of each laboratory should seek to minimise it by rotating the responsibility for dealing with the EQA specimens and ensuring their examination under conditions as near as possible to those obtaining for routine work.

HISTOPATHOLOGY

The provision of large numbers of near-identical specimens for EQA presents an even greater problem in the case of tissue sections for histology: indeed, it may be insoluble for many kinds of tissue. Moreover, there are deeper, conceptual problems in the EQA of histopathology, stemming from the fact that a histopathology result is usually a qualitative subjective judgement in a continuous spectrum of diagnostic possibilities. There is rarely an objective yardstick against which to measure a participant's results, and one must compare them either with those of an expert panel of assessors (corresponding to "reference laboratories") or with the consensus of the whole group, whichever is the more acceptable to the participants. The problems have been most lucidly discussed by Langley.[10] Partial solutions to these problems are considered below, under the heading "Cytology."

One aspect of histopathology seems less controversial, and more amenable to EQA: technical prowess in the processing, cutting, and staining of tissue sections. For the EQA of technical aspects of histopathology, large pieces of necropsy or surgical material can provide up to 50 tissue blocks; or, when the quantity of tissue is restricted, unstained sections can be used to assess quality of staining only. By simultaneously running 7 or 8 schemes, each with 40-45 participants, a single organiser can potentially cover all the laboratories in England and Wales. That, at any rate, is the rationale underlying a pilot scheme begun in 1980 that is based on a regional scheme previously operated successfully for 4 years in Wales.[11] Participants' returned slides are evaluated in four grades of satisfactoriness by two teams of four independent assessors, and the consistency and comparability of the assessors' opinions are then themselves evaluated both statistically and by a further team of checking assessors. The design is thus that of "reference laboratories," but an attempt is being made to overcome the subjectiveness of the judgements by having the assessors meet at intervals (after they have delivered judgement) to discuss and possibly resolve discrepancies.

A previous approach to EQA in histopathology, in which slides of moderate difficulty were despatched to participants, followed a month later by an interpretation of the slides by the experts who had selected them from their own recent cases, did not actually reach the point of *external* quality assessment since there was no feedback of participants' results to the organiser. This did not detract from the educational value of the scheme to many of the participants; it is an example of the application of the survey technique in the educational phase of a quality assurance programme.

As far as we know, these seem to be the only approaches in Europe to EQA in histopathology up to now, apart from short-term, small-scale surveys in Norway[12] and Denmark[13] which came to the conclusion that discrepancies among histopathologists' opinions on the same tissue sections showing early malignant changes were mainly attributable to disagreements about terminology. The same conclusion has been reached in pilot EQA studies in Britain in cytology, and the consequent reform of terminology represents the earliest and most direct beneficial educational effect of EQA in qualitative pathology investigations.

There seems to be no solution to the problem of specimen quantity for EQA of biopsy material, as opposed to necropsy or surgical material: this is serious, since histology of biopsy specimens constitutes the larger proportion and qualitatively perhaps the most crucial part of a histopathology laboratory's work. The use of photomicrographs to obviate the problem, proposed by the College of American Pathologists (CAP), is currently being evaluated by UK pathologists. Some biopsy specimens may yield up to 10 near-identical specimens, and groups of 10 laboratories were used in what Langley[10] understatedly calls "a rather complex model" by the CAP in 1970-71. Between-laboratory agreement and within-laboratory consistency (on repeat examination) was 78-83%.

A novel approach to EQA in histopathology, proposed by Codling in the UK and explored with financial support from the CAP,[14] uses pattern analysis of the incidence of various diagnoses emanating from each histopathology laboratory compared with national averages. It requires computerised recording of the total laboratory output of results, and cannot operate until a high proportion of laboratories introduces such recording. Interpretation of the results is complicated by the possible existence of demographic variability in disease states amongst the patient populations served by the laboratories surveyed. Nevertheless the method offers for the first time the potential of assessing a laboratory's work on the basis of its total output rather than on occasional survey specimens.

CYTOLOGY

Interlaboratory comparison of cytological investigations is subject to the same limitations as for histopathology—namely, that what constitutes the "correct" result is a matter of opinion—plus the fact that stained smears cannot be near-replicated as can adjacent histological sections. One possible solution, namely sending the same stained smear serially through a succession of laboratories for comparison of results is time-consuming and cumbersome; furthermore, the staining fades under repeated examination. One recent experimental approach in Britain has employed the exchange of slides between pairs of laboratories to determine at least the degree of discrepancy in findings, and has uncovered the need to define more closely the terminology used to denote morphological manifestations of malignant disease. A flaw in this experimental design is the absence of an agreed point of reference: with two laboratories there can be no "consensus," and with both laboratories equally likely to be right when there is disagreement, neither has any motivation for procedural change. A more satisfactory, if more complex, design uses the consensus of a "cluster" of laboratories, necessarily limited to 5 or 6 for the reasons given above, as the point of reference.[15] The consensus may even be strengthened to unanimity if the participants meet and discuss slides on which there has been discrepancy of opinion, although Wilson and Burke[16] warn against the false sense of security engendered by "forced agreements" brought about in this way.

Disagreements may possibly be further resolved by coupling a cluster of cytologists to a cluster of histologists who examine tissue sections from the same patients from whom the cytology material was obtained; or it may confuse matters further because of a disagreement amongst the histologists! A small pilot scheme to examine the outcome of this last design is under way in the UK. Whether a national scheme can eventually be developed as a network of linked clusters is a matter of speculation.

MICROBIOLOGY

The quality assessment of different aspects of a microbiology laboratory's work requires a variety of techniques. In bacteriology, the main kinds of investigation are to identify organisms in pure culture, to isolate and identify organisms in simulated patient specimens (sputum, faeces, throat and wound swabs), to determine antibiotic sensitivity, to perform bacterial serology and to assay serum antibiotic concentrations. Only the last-named investigation resembles clinical chemistry in yielding a quantitative result on a linear scale. In virology, testing for the presence or absence of a named virus, for example, hepatitis B, gives a yes/no answer, and assessing antibody titres give numerical answers on a logarithmic scale. Surveys of water and milk specimens that may contain pathogens are also made since these reflect another part of the public health laboratory's work.

The main problems of EQA in microbiology centre around the distributed specimen, its stability and viability, and public health safety during postal distribution. A national EQA scheme is conveniently based on the national reference laboratory, although this is not essential. The UK scheme, based at the Central Public Health Laboratory, serves 400 UK laboratories and more than 60 laboratories abroad; about 150 of these laboratories cover virology. The "intended result," based on the organising laboratory's formulation of the specimen, is checked by the consensus of returned results and by the organising laboratory's own assessment of specimens that have been through the postal distribution. A clinical "history" accompanies the specimen, together with detailed specification of the test(s) requested, and results are returned in the form in which they would be sent to a clinician, ie together with an interpretation. Scoring of the returns strongly reflects clinical adequacy as well as laboratory prowess: thus, the scoring takes into account the speed of response (telephoned and written), failure to respond, partially correct results, badly wrong results without the stated intention to refer the specimen elsewhere or carry out confirmatory tests, and inadequate emphasis in the report on the presence of pathogenic organisms. The assessment technique is, therefore, that of the reference laboratory tempered by comparison with consensus data (both qualitative and quantitative where applicable) and by clinical judgement as to the seriousness of error. Lyophilisation has been used increasingly to ensure viability and reproducibility of distributed specimens, which are commonly reconstituted in broth.

Major outcomes of the UK scheme have been the discovery of poor quality control (in the original, manufacturers' sense) in the production of culture media and of reagents (especially of diagnostic sera), and the inadequacy of many reagents and techniques for the assessment of antibiotic sensitivity. The last-mentioned finding has led to the exclusion of antibiotic sensitivity testing from the UK NEQAS pending the outcome of a separate multilaboratory study of these tests.

Because microbiology NEQAS surveys include many different kinds of investigation, and because the investigations are not automated, there is a strong temptation for laboratory staff to subject the EQA specimen to special treatment, and even to check their tentative conclusions with a reference laboratory before making the return. There is anecdotal evidence that this occurs to a not inconsiderable extent in the UK scheme. A solution to the corresponding problem in the German clinical chemistry Scheme is to circulate two *different* specimens (blind) to the various participating laboratories in any one distribution. In a Scheme not employing this manoeuvre, the educational benefits of participating are lost to a laboratory that does not treat specimens in a routine fashion, and the resources expended on the Scheme are largely wasted. The solution to this problem lies wholly in the hands of the head of the laboratory, who must prohibit both the above practices if he wishes his laboratory to benefit at all from participation.

The only other large-scale NEQAS for bacteriology in Europe is in France. The scheme covers 3800 laboratories and makes 3 or 4 distributions each year. A parallel scheme for parasitology has 1600 participating laboratories.

PHARMACOLOGY, TOXICOLOGY AND ENDOCRINOLOGY

Clinical pharmacology, toxicology and endocrinology all involve quantitative chemical determinations, so that EQA schemes for these disciplines could in theory be based on the same principles as for clinical chemistry. However, the provision of suitable specimens may be much more difficult than in general clinical chemistry.

Pharmacology and toxicology

In the EQA of assays of therapeutic drugs and of poisons, for example, the fact that patients' serum contains metabolites which are absent from a simulated specimen prepared by weighing in the substance under investigation may adversely affect the correlation between EQA results and laboratory performance. This difficulty has been set aside in the design of an international EQA scheme developed in Britain[17] for the assay of anticonvulsant drugs whose serum concentrations are commonly monitored because they must be maintained in a "therapeutic window." Specimens for distribution are made by weighing the (water-soluble) drugs into pooled outdated blood-bank plasma from drug-free donors; the calculated final concentrations are checked against consensus mean values. Marked improvement in interlaboratory agreement has resulted. The data have also shown biases between methods and the unsatisfactory nature of at least one method (spectrophotometry for phenytoin).

Endocrinology

The estimation of hormones in serum often depends critically on the presence and concentration of human serum proteins. Survey specimens based on animal serum are unsuitable, both because they lack human serum proteins and also because they contain abnormal interfering substances. Modifications of normal human serum can be satisfactorily made to produce specimens containing most of the appropriate concentrations, but the most difficult kind of specimen to produce is often the analyte-free one which is ultimately needed for testing the accuracy of an assay system. Sometimes, the only really appropriate specimen for EQA may be serum from a particular kind of patient, which is rarely available in sufficient quantity for EQA. Fortunately, the number of laboratories performing serum hormone assays (apart from thyroxine and cortisol) is relatively small, but this in turn leads to a different kind of difficulty, namely the statistical treatment of small numbers of values. This difficulty is easier to overcome than the other. Again, freeze-drying aimed at inducing stability may alter the properties of relevant proteins or the hormone itself, though this can be tested and the lyophilisation conditions can often be suitably adjusted. Finally, immunological methods used for hormone assay that depend on the use of antisera which differ greatly in composition from one another are necessarily reagent-dependent, so that results from different laboratories may be difficult to compare. However, evidence is now accumulating that "bias due to antiserum differences" (which arise because of the production of antibodies to different portions or forms of the circulating hormone) may be a figment used to cloak inadequacies of assay technique. Although assays of parathyroid hormone and gastrin, for example, are truly subject to such bias, the organiser of an EQA scheme would do best to assume that antiserum bias is absent until its existence is independently demonstrated.

Difficulties of EQA in endocrinology, coupled with the relatively small number of hormone-assay laboratories, may be responsible for a perceptible reluctance in many countries to introduce NEQASs in this field. However, at least one commercial "quality control" service (that of Wellcome Reagents Ltd) has included thyroxine and cortisol among its surveyed analytes for many years. This company is about to launch an international immunoassay EQA service with specimens containing several hormones in a human-serum base (there having been some scepticism expressed about the validity of using bovine serum matrices for the EQA of hormone immunoassays in the general scheme).

American pathologists have not shrunk from EQA of at least some hormones: decreases, some of them dramatic, in the average between-laboratory coefficients of variation have been shown by the CAP programme[18] for insulin, thyroxine, triiodothyronine and thyrotropin during the period (1975-8) when radioimmunoassays for these hormones were coming into widespread use. EQA in endocrinology has been developed rather extensively in the UK, mainly because of the formation in 1973-4 of the Supraregional Assay Service for hormones in England and Wales and the corresponding Inter-Area Immunoassay Support Service in Scotland. Some 15 of the constituent laboratories were committed to providing large numbers of hormone assays to high (and comparable) standards, and they therefore organised a number of small-scale informal EQA systems amongst themselves which have, with DHSS assistance, been opened to wider participation as the number of laboratories performing the assays has increased. The consequence is that the UK now has national EQA schemes for 14 hormone assays (progesterone, oestradiol-17β, cortisol, testosterone, T4, T3, TSH, LH, FSH, prolactin, HPL, serum oestriol, pregnancy urinary oestrogens, and insulin). There has been NEQAS coverage of a similar range of hormones in the Federal Republic of Germany since 1976,[19] but the frequency of distribution (3-4 per year) is much lower than in the UK (once every 3-4 weeks).

The UK schemes differ from the general clinical chemistry scheme in one fundamental respect: their organisers communicate frequently and actively with the participants, on the premise that the assays in question are far from mature and themselves need optimisation before variations in proficiency in performing them can be assessed. In an attempt to improve interlaboratory comparability before methodology becomes fossilised, the organisers have concentrated on detecting and eliminating methodological bias in particular methods or laboratories rather than tackling problems of imprecision as their prime consideration. *Bias* here means deviation from a reference group value, the reference group laboratories having been selected on the basis of (a) their ability to assay zero-analyte specimens correctly, (b) their results in recovery or dilution experiments and (c) their high precision on specimens repeatedly distributed "blind." In some instances concentration on laboratories' performance with analyte-free specimens produced such an improvement in interlaboratory agreement in accuracy—that is, elimination of bias—that it eventually became possible to use the over-all mean of participants' results as the point of reference for calculation.[20] With steroid hormones, where the biologically active molecule to be assayed is a chemically well-defined molecule of known composition, definitive assay by physical means—namely, mass spectrometry—has become possible and a method-independent reference point for a hormone is in principle feasible.[19] All problems are not yet eliminated, however: the definitive method is itself open to some degree of error (chiefly instrumental variability in the case of mass spectrometry), and there may be residual argument as to whether the biologically active, and hence clinically significant, constituent is the total, the conjugated or the non-protein-bound form of the molecule in the serum. Analyte-free sera are still theoretically necessary to establish the accuracy of the definitive method.

INTERNATIONAL EQA SCHEMES

Because EQA has developed more rapidly in some subdisciplines in the UK than in other countries, there are naturally requests and suggestions that a given scheme be made available internationally—for example, throughout Europe. However, except for certain assays for which the scheme itself would benefit from having larger numbers of data to which statistically valid procedures could be applied, supranational schemes have several disadvantages compared with a series of national schemes, coordinated by limited mutual participation. Quite apart from the postal and administrative difficulties of dealing with geographically scattered laboratories and the different systems for regulating and licensing laboratories in different countries, the essential educational function of EQA is exercised most effectively by vicinal groups of competent peer professionals. In most European countries, the relevant professional groups are well enough developed, recognised, and organised that they can be charged with the task of helping the less accurate and precise laboratories to come up to standard, or else to develop better terminology and definitions to bring about interlaboratory agreement, if that is what is needed. That is not to deny the value of existing programmes of technical co-operation, supported by EQA schemes, between developed and developing countries in (for example) general clinical chemistry, assay of reproductive hormones and haematology.

Scope for European co-operation

Above and beyond this, there seems to be further scope for mutual learning about EQA systems both within Europe, between European countries and the USA, and among disciplines. For example, some of the apparent problems of obtaining enough specimen for distribution to the relatively small numbers of laboratories (300-400) in the UK seems to have been solved in other countries where there is distribution to many thousands of laboratories, admittedly at much lower frequency. Discussion amongst the national organisers might prove fruitful. Again, the evaluation of cell morphology in blood films is fairly commonly accepted as part of EQA for haematology in many European countries but has not yet been incorporated into the performance score provided by the UK EQAS for haematology. Haematologists have nevertheless welcomed the blood film surveys, and this may encourage histopathologists to support corresponding efforts on their behalf. EQA in haemocytometry based on highly artificial mixtures of animal cells and animal blood seems to have gained acceptance, while EQA of cervical and pulmonary cytology seems rigidly confined to scarce and variable human specimens. Perhaps microbiology, much of which is non-numerical and dependent on pattern recognition, has lessons to impart to the pathology disciplines that claim immunity from EQA on the grounds that their character, so different from clinical chemistry, is non-quantitative and subjective. And perhaps the UK has much to offer in developing further its strategic designs in the EQA of histology, cytology, and endocrinology in conjunction with its European partners.

References

[1] *External quality assessment of health laboratories.* Copenhagen: WHO Regional Office for Europe, 1981.

[2] Report of the Working Party on "Factors affecting analytical performance in clinical chemistry laboratories" (1980). Available from Dr PRN Kind, Dept of Clinical Chemistry, St Thomas's Hospital, London SE1 7EH.

[3] Hansert E, Stamm D. Determination of assigned values in control specimens for internal accuracy control and for interlaboratory surveys. *J Clin Chem Clin Biochem* 1980; **18**:461-90.

[4] Whitehead TP, Browning DM, Gregory A. A comparative survey of the results of analyses of blood serum in clinical chemistry laboratories in the United Kingdom. *J Clin Pathol* 1973;**26**:435-45.

[5] Healy MJR. Outliers in clinical chemistry quality control schemes. *Clin Chem* 1979;**25**:675-7.

[6] Whitehead TP. *Quality control in clinical chemistry.* New York: Wiley, 1977.

[7] Lewis SM, Coster JF, eds. *Quality control in haematology.* London: Academic Press, 1975.

[8] Poller L. Quality control in blood coagulation. In: Thomson JM, ed. *Blood coagulation and haemostasis.* Edinburgh: Churchill Livingstone, 1980;331-59.

[9] Myrhe BA, Mullen S, Polesky HF, Van Schoonhoven P, Walker R. The comprehensive blood bank survey program of the College of American Pathologists—1977. *Am J Clin Pathol* 1979;**72**:352-7.

[10] Langley FA. Quality control in histopathology and diagnostic cytology. *Histopathology* 1978;**2**:3-18.

[11] Barr WT. Technical quality control in histopathology. *J Clin Pathol* 1978;**31**:996-8.

[12] Iversen OH, Sandnes K. The reliability of pathologists. *Acta Pathol Microbiol Scand* 1971;**79**:330-4.

[13] Ringsted J, Amtrup F, Arklund C *et al.* Reliability of histo-pathological diagnosis of squamous epithelial changes of the uterine cervix. *Acta Pathol Microbiol Scand* 1978;**86**:273-8.

[14] Henson DE, Codling BW, Macartney JC. *Interlaboratory histological evaluation: a new approach to quality control in anatomic pathology.* Skokie, Illinois: College of American Pathologists, 1976.

[15] Evans DMD, Shelley G, Cleary B, Baldwin Y. Observer variation and quality control of cytodiagnosis. *J Clin Pathol* 1974;**37**:945-50.

[16] Wilson EB, Burke MH. Some statistical observations on a co-operative study of human pulmonary pathology, I. *Proc Natl Acad Sci USA* 1957;**42**:1073-8.

[17] Griffiths A, Hebdige S, Perucca E, Richens A. Quality control in drug measurement. *Therapeutic Drug Monitoring* 1980;**2**:51-9.

[18] Hansell JR, Haven GT. Changes in level of precision of common ligand assays during a seven-year interval. *Am J Clin Pathol* 1979;**72**:320-5.

[19] Röhle G, Breuer H. External quality control for hormone determinations in the Federal Republic of Germany. *Horm Res* 1978;**9**:450-4.

[20] Hunter WM, McKenzie I. Quality control of radioimmunoassays for proteins: the first two and a half years of a national scheme for serum growth hormone measurements. *Ann Clin Biochem* 1979;**16**:131-46.

Requests for reprints to: Dr FP Woodford, Department of Health and Social Security, Hannibal House, London SE1 6TE, England.