Response analysis in histopathology external quality assessment schemes

P N Furness, I Lauder

Abstract

Aims—To develop a computerised method for analysing the results of histopathology external quality assessment (EQA) schemes which can provide confidential personal reports to individual participating pathologists.

Methods—A program was developed using the OMNIS database system, running on Apple Macintosh or IBM compatible computers.

Results—The program produces a general report of participants' responses to each case, and a choice of two types of personal report. One of these provides a list of the participant's diagnoses with a list of the most popular (Consensus) diagnoses for comparison. The other provides automatically calculated scores for the pathologist's performance along with simple statistical evaluation. The scores can be calculated by comparison with the consensus of the group or with correct diagnoses if they are known. A histogram indicating the distribution of performance within the group can be produced. The program can accept uncertainty in the form of differential diagnosis lists from participants. Potentially dangerous diagnostic errors can be identified and handled separately. Participants are identified only by code numbers and confidentiality can easily be enforced. The program is currently being used in the national renal pathology EQA scheme and in the local general histopathology scheme in the East Midlands.

Conclusions—This program offers solutions to problems which have bedevilled the organisers of histopathology EQA schemes. It offers confidential advice to pathologists and will help to identify areas where an individual might benefit from continuing career grade medical education. It raises the possibility of the development of nationally agreed standards of performance in the reporting of pathological specimens, and it may be applicable to other specialties where textual reports are produced.

(J Clin Pathol 1993;46:357–363)

External quality assessment (EQA) schemes in clinical chemistry, microbiology, and laboratory haematology have been running for many years on a national basis. They use computerised methods of analysing the results and the statistical methods in these analyses permit confident decisions to be taken on the need for remedial action when a laboratory falls persistently below a predefined standard.

The situation in histopathology is radically different. EQA schemes are relatively new. They are usually small and organised locally. The analysis of responses rarely goes beyond a laborious manual tabulation of the numbers of pathologists supporting each of a number of suggested diagnoses for each case. There is usually no further statistical analysis; an assessment of the performance of individual histopathologists is rarely attempted. In the absence of any personal type of report the participant has to assess him or herself in a completely subjective way, often by memory, from diagnoses made months before.

Without any objective assessment it is impossible for criteria of acceptable performance to be developed. Without such criteria the EQA scheme cannot inform a pathologist whether an adequate standard is being maintained.

It could be argued that histopathology EQA schemes assess the performance of individuals, not laboratories, so a comparison with other laboratory disciplines is invalid. The primary purpose of an EQA scheme is to detect when a laboratory is providing a seriously substandard service. In the histopathology laboratory the output depends on many processes, but its essence is the final opinion of the histopathologist. There is widespread agreement that EQA schemes are useful in histopathology. If this is agreed, then the quality of the diagnoses provided by individuals must be assessed. We suggest that the subjective self assessment provided by schemes at the moment is quite inadequate for this purpose. Objective, reproducible methods are needed to give pathologists the information with which to assess themselves without question of bias, and to permit the development of definitions of acceptable standards. Without such information, histopathology EQA schemes may have educational value, but they lack any real element of quality assessment.

The evaluation of responses in histopathology EQA schemes has been limited by problems which do not arise in other laboratory disciplines.

(1) The responses are textual, not numerical, and so are not easily amenable to
With most diagnoses there is a potential for synonyms, and, more difficult still, partial synonyms which may convey subtle additional information.

The range of possible diagnoses is vast. If a list of diagnoses was to be provided for participants to choose from, its size would rival the SNOMED code book.

Uncertainty in responses needs to be allowed. A differential diagnosis list is often essential, especially if participants are provided with small biopsy specimens and limited information on further investigations.

The true diagnosis may never be known. Indeed, a review of the history of medicine would probably indicate that even those diagnoses which today seemed obvious might be viewed differently in years to come.

Not all diagnostic errors are equally important. To misclassify one completely excised benign tumour as another may not matter, but an error between a benign and malignant diagnosis may be disastrous.

The participants are individuals; confidentiality is therefore paramount, and notification of performance must be done with great sensitivity.

These problems have been addressed only to a limited extent. In some specialist schemes (for example, cervical cytology and breast pathology in the United Kingdom) a list of potential diagnoses for the whole specialty has been drawn up. Participants have to choose from the list. This facilitates analysis, but it is not applicable to general histopathology. In the large scheme organised by the American College of Pathologists, a list of potential diagnoses for each case is drawn up by a panel of experts who have also provided the correct diagnosis (personal communication, Dr H Travers, chairman, Surgical Pathology Committee, College of American Pathologists). This would be too laborious for the small local schemes which exist in the United Kingdom. It can also be criticised for departing substantially from normal pathological practice; biopsy specimens rarely arrive in the laboratory with a shortlist of diagnoses, one of which is known without doubt to be correct. In contrast, the scheme organised by the Royal College of Pathologists of Australasia asks participants to report the cases as they would do normally (personal communication, Dr J Maynard, coordinator, RCPA Quality Assurance Program, Anatomical Pathology). This is a commendable ideal, but the analysis of lengthy verbal reports results in long lists of proffered diagnoses which often overlap in meaning. A personal analysis cannot be provided. An attempt to provide personal reports has recently been described in Scotland, but the assessment involves responses being marked as 0, 0.5, or 1 by the organiser—who is presumably also a participant—in the light of a preferred diagnosis, which is usually the diagnosis of the original report. This dependence on one person to make subjective judgements of his or her colleagues is clearly not ideal.

The computer program described here attempts to provide an objective method of assessment for the responses to histopathology EQA schemes. It does not attempt to define what is an unacceptable or dangerous level of performance; such questions are ones which must be addressed by the profession as a whole, once suitable methods of measurement have been provided.

The program was originally designed for the analysis of responses in the United Kingdom National Renal Pathology EQA scheme, where problems with inadequate biopsy specimens and differential diagnosis lists are particularly severe, but it has proved to be readily applicable to the United Kingdom model of small local general histopathology EQA schemes. To facilitate the development of such schemes on a national basis the Department of Health has provided funding for the establishment at Leicester of a National Coordinating Centre for Histopathology EQA and it is intended that further development, support, and distribution of the new software will be undertaken at the National Coordinating Centre.

The program uses the OMNIS database system (Blyth Software Ltd). It will run on any Apple Macintosh with a hard disk, or an IBM compatible computer running Windows 3, although it is somewhat sluggish on an IBM with less than a 386 processor.

**Methods and results**

The overall pattern of an EQA scheme using this method of analysis is shown in fig 1. Data are entered into labelled boxes in user-friendly screens which are called up from the data menu. When cases have been identified, the “enter clinical data” screen (fig 2) allows a secretary to enter relevant details of each case before the circulation starts. An option of the output menu then prints out the data on sheets on which the participants can enter their responses (fig 3). These sheets are designed to facilitate the subsequent analysis.

Participants are each allocated a confidential code number by a secretary, who should be the only person who holds a list of all the codes and participants' names. It is essential that one person has this information, to allow

---

**Histopathology EQA analysis**

1. Obtain cases
2. Enter clinical data
3. Generate response forms; send slides, forms and confidential codes to participants
4. Receive completed response forms and enter data into computer
5. Calculate "popularity" of each diagnosis
6. Print and circulate case analyses
7. Hold participants' meeting to discuss cases; possibly consult clinical colleagues on what constitutes a dangerous diagnosis
8. Amend data in the light of (7)
9. Print and distribute personal analyses and distribution histogram

---

*Figure 1.*
Figure 2  Screen for entering the clinical information which is to be circulated to participants with the sections. The control boxes containing apparently illegible text are greyed to indicate that those functions are not available at the current stage in the data entry process.

personal reports to be posted correctly, but this person need not be a pathologist and need have no access to the database. The personal reports can be folded before being passed to this person for posting, so that even the secretarial staff are unaware of individual pathologists’ scores. When the participants return their opinions, they may enter their code number, if they wish to receive a personal analysis, or they may submit the response anonymously.

The form (fig 3) asks for a single main diagnosis for each case, but if a differential diagnosis list is necessary, this can be accommodated. The participant is simply asked to give an assessment of the probability of each diagnosis on the list by allocating a numerical weighting to each, such that the total adds up to 10. For example, if a colonic biopsy specimen was thought to represent a tubulovillous adenoma this diagnosis alone would be offered; but if a participant found an area which raised a possibility of invasion, the response might be: Tubulovillous adenoma (n = 9); Adenoma with focal invasive adenocarcinoma (n = 1).

This is the area of the response on which the personal analysis will occur. Secondary diagnoses, subtle additional points, further investigations and other comments can be recorded elsewhere on the form.

When responses are returned to the scheme coordinator they can be recorded via another data entry screen (fig 4). When the case number is entered, any diagnoses which have previously been suggested are displayed. Further diagnoses can be added if necessary. If a differential list is given, the probability accorded to each diagnosis is recorded by entering values (0 to 10) in the boxes on the right.

Comments, finer points of classification (such as the degree of dysplasia in the adenoma quoted above), additional diagnoses and other information can be entered in the box at the bottom of the screen. This supplementary information is held with the record of the case rather than with the record of the participant. It can therefore accumulate as analyses are performed, much like conventional manual EQA analysis, but it is not available for subsequent statistical analysis.

The data entry process involves building a list of diagnostic categories for each case which are meaningful, informative, and clinically relevant but which do not overlap. This must involve a degree of interpretation of the actual text provided by the participants, and requires judgments about what aspects of each case are actually of importance. Items of secondary importance are relegated to the box at the bottom of the screen, and will appear on reports without further analysis. For example, a skin biopsy specimen with a heavy lymphocytic infiltrate was circulated. This produced a wide variety of diagnoses, but the distinction of major clinical importance was is this lymphoma or not? The diagnostic categories chosen for analysis were therefore Benign lymphocytic infiltrate and Lymphoma. The numbers of pathologists suggesting insect bite, sarcoidosis, subtypes of lymphoma, benign infiltrate NOS, etc., were recorded separately. This is an extreme exam-
ple, but to do otherwise would leave the one pathologist who suggested insect bite being categorised as equally wrong as the one who diagnosed lymphoma with complete confidence, despite a consensus that the lesion was benign.

It is clear that this process must be carried out by a pathologist, not a secretary. We have found this to be no great problem even if the pathologist in question has little experience of using computers. The amount of pathologist’s time taken in entering data decreases with experience; after having entered three circulations of the Renal Pathology EQA scheme, the process takes less than 20 seconds per case response. This is comparable to a simple manual analysis, but the subsequent potential for data analysis is much greater.

The program provides various automatic and manual data verification procedures. When the data entry is complete, selection of an option from one of the menus causes the arithmetic mean of the scores allotted to each diagnostic category to be calculated. A case by case analysis can then be printed, giving a breakdown of the “popularity” of each diagnosis (fig 5). At this stage, the report is comparable to conventional EQA analysis.

Before personal reports can be produced, decisions are needed on how to proceed with the analysis (indeed, whether to proceed at all). These decisions are probably best made by a group of participating pathologists rather than by the organiser acting alone. It is therefore appropriate to have a meeting of participants to discuss the cases at this point, at which the following questions can be answered in the light of the analysis so far.

(1) Are any of the cases inappropriate for the EQA scheme? A case may be considered to be too obscure or specialist, such as a rare neuropathological diagnosis in a general EQA scheme, or it might have been circulated with inadequate or misleading information. If so, there is a facility to delete the case before proceeding to analysis of individual performance.

(2) Do any of the diagnostic categories entered by the organiser overlap unduly? The system which generates the personal reports assumes that a participant who has made one diagnosis with complete confidence (score 10) considers all the other diagnoses in the list to be wrong. It is therefore important that each diagnosis in this list is sufficiently broad so that it effectively excludes all the others. The example of the lymphocytic skin infiltration quoted earlier is relevant here. There is a facility to merge overlapping diagnoses under a new heading, along with all the associated data.

(3) Should any of the diagnosis scores be changed? The calculated scores represent a consensus of the participants, but further information might show one diagnosis to be unequivocally correct (or, indeed, incorrect). The organiser can amend the scores as necessary. In general, however, the consensus answer is probably best unaltered. It represents what a group of pathologists can be expected to make of the material provided. If subsequent information proves the majority to be wrong, the case was probably inappropriate (see Question 1) and should be deleted.

(4) Are any of the preferred diagnoses clearly wrong and clinically dangerous? A clear benign/malignant error would usually come into this category. If so, this can be recorded. Any participant who has suggested such a diagnosis with over 50% confidence will then be specifically informed in the analysis.

This question may be considered by some pathologists to be the most important part of the analysis, even though it involves a subjective judgment. As an alternative approach, we are currently asking our clinical colleagues to
Response analysis in histopathology EQAs

indicate, as the potential recipients of these reports, which diagnoses might constitute a serious error—assuming that the consensus diagnosis can be accepted as correct. Preliminary results indicate a considerable divergence between pathologists and clinicians on what might be considered dangerous.

If personal reports are required, the organiser makes any amendments in the light of the decisions of the participants’ meeting. The program can then be instructed to carry out the necessary calculations. A score is calculated for each pathologist’s response to each case. The maximum score is obtained either by having made (with complete confidence) the diagnosis which proved to be most popular, or by having given a list of differential diagnoses with probabilities which exactly match the consensus of the group. Lesser degrees of correlation, or giving one diagnosis which is less popular, give proportionately lower scores. Two types of personal report can then be printed.

The first (fig 6) lists for each case in a circulation the participant’s main single diagnosis and the most popular diagnosis as defined by the analysis of the group. This serves to remind the participant of what he or she said about each case, and to identify not only cases where low scores were recorded, but why a low score was recorded. In this report dangerous diagnoses are indicated with an asterisk.

The second type of personal report (fig 7) provides a more detailed numerical breakdown of a participant’s response. It again provides the participant with his or her own score for each case, but in addition the highest, lowest, and mean scores for the whole group are given for comparison. This serves to indicate which cases produced disagreement in the group. The arithmetic mean of the participant’s scores for the current circulation is also presented. The mean of the participant’s accumulated scores in the most recent circulations (up to a maximum of 5) is also presented, along with the highest, lowest, mean and standard deviation of the cumulative scores for the whole group. This accumulation over five EQA circulations means that the cumulative score comes to represent an analysis of a very large body of data, such that an occasional diagnostic aberration will have little effect, unless that aberration was recorded as a dangerous diagnosis. These are listed separately at the bottom of the personal report.

Study of these printouts provides a considerable amount of information. Individuals who are overconfident might attain a reasonable score, but an occasional overconfident wrong diagnosis would show in the dangerous diagnosis list. Excessively indecisive individuals who present lengthy differential lists for every case might never record a dangerous diagnosis, but the indecision would lead to a low overall score. If an individual’s score is providing cause for concern the analyses will identify whether there is one particular type of case or area of pathology which is causing a problem.

Responses which were returned anonymously obviously cannot be analysed in this way. Instead the data are pooled for all anonymous responses and participants for whom an identified response was not received can be sent an analysis of this pooled information. Participants who feel threatened by this departure analysis can therefore opt out.

The facility also exists to output the cumulative scores to a disk, without any identification data, in forms which can be read by a number of graph drawing programs (including Microsoft Excel). An example is given in fig 8.

Conclusions

We may now review the list of problems outlined in the introduction, and consider the extent to which they have been overcome.

1) Textual data are converted into numerical data, permitting statistical manipulation, the allocation of individual scores, and graphical demonstration of the spread of the group.

2) Synonyms are recognised by the pathologist who enters the data. Subtleties of diagnosis cannot be used for assessment, but can be recorded and reported.

3) The list of diagnoses for each case is built up from the opinions of the participating pathologists, who are therefore as free to form opinions as they are in

---

**National Renal Pathology EQA Scheme**

Report for participant 999 on cases scoring 1.00 or below in circulation A

Note that your score relating to each case is dependent on numerous factors, including the ‘weight’ you assigned to your favoured diagnosis, the weight you assigned to any other diagnoses you made and the degree to which opinions differed within the group. Consequently the score cannot be deduced from the data presented in this report.

<table>
<thead>
<tr>
<th>Case No</th>
<th>Score</th>
<th>Favoured main single diagnosis</th>
<th>Favoured main diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.10</td>
<td>*Alport’s disease/hereditary nephropathy</td>
<td>Focal segmental glomerulosclerosis</td>
</tr>
<tr>
<td>3</td>
<td>0.04</td>
<td>Scleroderma</td>
<td>Hypertension/Ischaemia</td>
</tr>
<tr>
<td>4</td>
<td>0.08</td>
<td>*Leptospirosis</td>
<td>Acute tubular necrosis</td>
</tr>
<tr>
<td>5</td>
<td>0.08</td>
<td>Diabetes</td>
<td>Lupus glomerulonephritis</td>
</tr>
</tbody>
</table>

Asterisks (if any) indicate responses recorded as dangerous diagnoses.
National Renal Pathology EQA Scheme
Personalised Response Analysis

If you sent a response back to me with your code number included, this analysis sheet should bear that code number and the analysis relates solely to you. If you did not return your response sheet, or did so without entering your code, this analysis relates to the mean of all the anonymous responses I received (and is therefore of relatively little value to you). A non-personalised breakdown of the diagnoses offered for each case, and their relative popularities, is included. The coded response sheets and your addressed envelope should have been put together by my secretary. If the code below is either not yours or zero (anonymous), then we have made a mistake; please return this sheet to me!

Dr Peter Furness, PhD, MRCPath, EQA Scheme Organiser, Department of Pathology, Clinical Sciences Building, Leicester Royal Infirmary, PO Box 65 Leicester.

Response analysis for participant code: 0
(Code 0 indicates sum of anonymous responses)

<table>
<thead>
<tr>
<th>CASE</th>
<th>Highest score:</th>
<th>Your Score:</th>
<th>Mean score:</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1-00</td>
<td>0.71</td>
<td>0.77</td>
</tr>
<tr>
<td>3</td>
<td>1-00</td>
<td>1.00</td>
<td>0.95</td>
</tr>
<tr>
<td>4</td>
<td>0.03</td>
<td>0.71</td>
<td>0.72</td>
</tr>
<tr>
<td>5</td>
<td>1-00</td>
<td>0.82</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Your score for this circulation: 0-809

Highest cumulative score: 0.00
Lowest cumulative score: 0.402
Mean cumulative score: 0.806
Standard deviation: 0.147

'Dangerous Diagnoses' recorded by you in this circulation for case(s):

Figure 7  Sample personal response analysis. This particular example, identified with participant code zero, represents the mean of all the responses which were returned anonymously. Up to 20 cases per circulation can be accommodated.

Routine practice.

Differential diagnosis lists can be readily accommodated.

The participants themselves can decide whether there is a true diagnosis, and whether to use it as the basis of the personal evaluation. Comparison with the peer group is available as an alternative, and provides a valid form of assessment even in cases where the diagnosis is contentious.

Diagnostic errors which are particularly dangerous can be identified and the pathologist responsible can be made aware, although this is not used as part of the statistical analysis. Diagnostic dif-

Figure 8  Histogram illustrating the distribution of cumulative scores within a group (data generated artificially for illustrative purposes only).
Response analysis in histopathology EQAs

To maintain confidentiality, the correlation between numerical codes and participants’ names need only be known by one person. If that person has nothing to do with the rest of the system, he or she can be asked only to put folded papers into appropriate envelopes, and so be kept unaware of the score of any individual pathologist. If a participant falls below acceptable standard (however that may be defined in the future), a “Dear colleague” letter could be sent by a similar procedure along with a breakdown of cases where a low score or a dangerous diagnosis was recorded. This should enable the participant to pinpoint the cause of the poor performance and act accordingly.

Histopathology external quality assessment is one form of medical audit and it is clearly the responsibility of the profession as a whole to define acceptable standards of performance. Until now this has been extremely difficult in histopathology but over a period of several EQA circulations using the computerised analysis outlined in this paper it would be possible for bodies such as the Histopathology Quality Assurance Advisory Panel and the Royal College of Pathologists’ Histopathology Specialty Advisory Committee to define an agreed minimal level of performance. For individual pathologists over a period of time specific areas of weakness for that particular histopathologist can be identified. This could obviously be very useful in the selection of the most appropriate continuing career grade medical education courses. Procedures will also have to be evolved to deal with the situation where individual pathologists are persistently making serious errors in their histological diagnoses. In assessing satisfactory performance it will probably be necessary to take into account both the overall score and the presence of serious errors; otherwise individual participants could avoid the possibility of making serious errors by the simple technique of giving deliberately vague and strictly neutral diagnoses.

To facilitate comparability on a national basis the National Coordinating Centre is actively considering the possibility of exchanging case material between schemes from diverse parts of the United Kingdom. If each circulation of, say, 15 slides were to include two or three cases which had been used in several other centres then over a period of time it might be possible to define a national standard of satisfactory performance, perhaps as falling no more than two standard deviations below the national mean score. There would obviously be a need for considerable discussion within the profession as to what constitutes a satisfactory level of performance and it is hoped that the suggestions contained in this paper will engender a constructive debate on this very difficult problem.

With adequate safeguards of confidentiality this method of histopathology EQA analysis should not be unduly inquisitorial, particularly if the option to return responses anonymously is maintained. In practice it has been accepted in the National Renal Pathology EQA Scheme and in the local histopathology EQA scheme in the East Midlands. Relatively few anonymous responses are being received.

This program has the endorsement of the Histopathology National EQA Steering Committee for the analysis of general histopathology EQA schemes.

Addendum

Since submission of this paper the program has been developed further. There is now a facility to identify cases as being of a specific type: skin, gynaecological, gastroenterological, etc. A report can then be produced which gives participants a personal breakdown of their mean “score” in each of the subspecialties, along with the mean and standard deviation of the group for comparison. The potential for guiding the use of continuing medical education (CME), and for monitoring its effectiveness, is obvious.

We are grateful for the advice and support of the Histopathology National EQA Steering Committee and many other pathologists. Invaluable statistical advice was provided by Dr Paul Burton. The program uses Omnis 7, a product of Blyth Software. It was developed on computers provided by the Department of Health (for the National Renal Pathology EQA scheme) and from Trent Regional Health Authority Research Committee.