Kidney International, Vol. 60 (2001), pp. 1998–2012

International variation in the interpretation of renal transplant biopsies: Report of the CERTPAP Project[1]

PETER N. FURNESS and NICHOLAS TAUB, for the Convergence of European Renal Transplant Pathology Assessment Procedures (CERTPAP) Project

International variation in the interpretation of renal transplant biopsies: Report of the CERTPAP Project.

Background. The Banff working formulation of renal transplant pathology is intended to have international application. There remains a need to develop methods to harmonize the application of such grading systems between laboratories. Banff grades do not always permit precise management decisions to be made. Alternative schemes have been devised for the diagnosis of acute rejection, but there have been no independent tests of the different approaches.

Methods. Sections from 55 renal transplant biopsies were circulated around the laboratories of 22 major transplant units for the Convergence of European Renal Transplant Pathology Assessment Procedures (CERTPAP) Project. Participating pathologists were asked to grade 32 different histological features, without any clinical information. After each circulation of five cases, feedback was provided to participants. Statistical evidence of improvement in interobserver variation was sought. At the end of the study, correlations with the original clinicopathological diagnosis were sought.

Results. Interobserver variation was greater than has previously been reported. For every feature studied, some pathologists consistently under-grade or over-grade. There was relatively little evidence of improvement in interobserver variation as a result of the feedback system. No single feature permitted a reliable diagnosis of acute rejection. Applying the Banff and CCTT schemas to the histological grades showed no clear diagnostic advantage for either system, but a simple computer-based inference network, which combined data from 12 histological features, out performed either approach. Within the "protocol" biopsies studied, long-term survival correlated better with "acute" than with "chronic" histological features.

Conclusions. These results do not undermine the value of the Banff classification, but they demonstrate a need for caution when translating biopsy results between institutions. It is obvious that evaluation of biopsies in multicenter trials must be done in one center. In the management of individual patients, the need to interpret Banff grades in the light of local experience and clinical information is stressed.

The Banff working classification of kidney transplant pathology was developed with two distinct aims: "*to guide therapy in transplant patients and to help establish an objective rejection end point in clinical trials*" [1]. Evidence of its success in making progress towards these goals is provided by its almost universal acceptance in clinical trials and research publications where the interpretation of renal transplant biopsies is required.

The definitions within the Banff classification mostly represent discrete points imposed on a natural biological continuum, such as the severity of tubulitis or interstitial fibrosis. Consequently, it is meaningless to speak of a "true" grade for a biopsy; the system is an artificial human construct, and the "correct" grade is merely that which is agreed by international consensus. Hence, the two most important attributes of any scheme of histological grading are clinical relevance and reproducibility. Numerous publications have confirmed the clinical relevance of the Banff classification [2–6]. A smaller number have tested its reproducibility and have found it to be acceptable, if not ideal [7, 8].

However, all of the published studies of reproducibility of the Banff classification have been performed by small groups of dedicated transplant pathologists who have worked closely together and who therefore may be expected to have reached a degree of consensus on how the Banff classification should be applied. It can be argued that this is not a sufficiently rigorous test. If a scheme is to be used globally, then it should be tested

[1] A complete listing of the participants who all contributed equally to the work of the CERTPAP Project includes **Peter N. Furness** (*Leicester, UK*); **Nicholas Taub** (*Leicester, United Kingdom*); **Karel J.M. Assmann** (*Nijmegen, The Netherlands*); **Giovanni Banfi** (*Milan, Italy*); **John N. Botelis** (*Athens, Greece*); **Marta Carrera** (*Barcelona, Spain*); **Jean-Pierre Cosyns** (*Brussels, Belgium*); **Anthony M. Dorman** (*Dublin, Ireland*); **Dominique Droz** (*Paris, France*); **Claire M. Hill** (*Belfast, N. Ireland*); **Bela Iványi** (*Kossuth, Hungary*); **Silke Kapper** (*Mannheim, Germany*); **Erik N. Larsson** (*Uppsala, Sweden*); **Aryvdas Laurinavicius** (*Vilius, Lithuania*); **Niels Marcussen** (*Aachus, Denmark*); **Anna Paula Martins** (*Lisbon, Portugal*); **Michael J. Mihatsch** (*Basel, Switzerland*); **Lydia Nakopoulou** (*Athens, Greece*); **Volker Nickeleit** (*Basel, Switzerland*); **L-H Nöel** (*Paris, France*); **Timo Paavonen** (*Helsinki, Finland*); **Agnieszua K. Perkowska** (*Warsaw, Poland*); **Heinz Regele** (*Vienna, Austria*); **Rafail Rosenthal** (*Riga, Latvia*); **Pavel Rossmann** (*Prague, Czech Republic*); **Wotgech A. Rowinski** (*Warsaw, Poland*); **Daniel Seron** (*Barcelona, Spain*); **Stale Sund** (*Oslo, Norway*); **Eero I. Taskinen** (*Helsinki, Finland*); **Tatjana Tihomirova** (*Riga, Latvia*); and **Rudiger Waldherr** (*Mannheim, Germany*).

Key words: Banff classification, CERTPAP Project, kidney transplantation, pathology, histological grading, morphological grading, CCTT criteria.

globally. When we interpret publications from different countries we need to know whether the Banff grades quoted are directly equivalent to our own experience.

To develop a more rigorous test experienced renal transplant pathologists were recruited from 22 major transplant centers, scattered over most of the countries of Europe, for the Convergence of European Renal Transplant Pathology Assessment Procedures (CERTPAP) Project. The participants were asked not to make diagnoses, but to undertake pure morphological grading of histological features. Clinical information was deliberately withheld, as we did not wish the results to be influenced by skills of clinical interpretation.

During an earlier study of the Banff classification, limited to the United Kingdom, participants had requested better training in the application of the Banff classification than was available merely by reading the literature. In anticipation that with the design of the present study initial reproducibility was likely to be low, from the onset, a system was implemented whereby results were fed back to participants at intervals, to allow them to compare their grading of the relevant histological features with the average of the entire group. We argued that this would facilitate "convergence" of grading criteria, and hoped that this could provide a mechanism whereby such ongoing training could be offered to larger numbers of renal transplant pathologists.

Although clinical information was withheld from participants it was collected in some detail. This allowed us to test the correlation between each histological feature and the diagnosis of acute rejection or the subsequent rate of decline of graft function. It also permitted a simple test of the definitions of acute rejection that are implied in the Banff classification and in the more recent CCTT classification [9], and it allowed us to test other ways of integrating the histological grades into a clinically meaningful diagnosis.

METHODS

Cases

To make the study as representative as possible of routine work, the microscope sections were provided by the participants. Cases were selected according to the following criteria:

(1) Acute rejection. Biopsies taken within six months of transplantation, where subsequent clinical review showed clearly that the transplant either (*a*) was definitely suffering from acute rejection (defined as an increase in serum creatinine of at least 15% of baseline in the week preceding the biopsy, followed by a fall to within 5% following treatment, or loss of the graft to rejection, with no other changes to explain the changes in creatinine), or (*b*) was definitely not suffering from acute rejection (this is, either a "protocol" biopsy in a graft with a stable creatinine, or a biopsy for graft dysfunction where the problem was subsequently shown to be something other than rejection, and responded to treatment of that problem).

Table 1. Clinical information collected on each case

Age of donor
Age of recipient
HLA matching
Cause of renal failure (if known)
Date of transplantation
Duration of any delayed graft function
Immunosuppressive protocol
Date of biopsy
Lowest serum creatinine (with date)
Most recent serum creatinine (with date)
Number of acute rejection episodes, and how they were treated
Any other relevant complications

(2) Chronic rejection. A "protocol" biopsy taken from a stable graft at any time from six months to two years after engraftment. These biopsies should have been taken at least five years ago, to provide a reasonable length of follow-up to allow a meaningful correlation with subsequent clinical outcome.

Clinical information was collected as defined in Table 1. From each biopsy twelve sections were cut, six stained with hematoxylin and eosin (H&E) and six with periodic acid Schiff (PAS). Groups of five cases, with one H&E and one PAS section per case, were circulated by post around small groups of participants, to allow each participant two weeks to view the sections and post them on, while allowing all participants to view the cases in two months. In some centers the responses were agreed by a small team of observers, sometimes including nephrologists, in accordance with usual practice at that centre, but a single response was requested per institution. The postal circulation of the slides was maintained using reminder letters generated by a software package previously written for the UK National Renal Pathology External Quality Assessment Scheme (*http://www.le.ac.uk/pa/pnf1/eqa/*). The number of sections was deliberately less than the Banff classification recommends, as it would not have been possible to cut sufficient replicate sections from one biopsy. We recognized the risk that different participants were viewing significantly different sections, an assessment of the impact of this effect was made, as discussed later in this article.

Participants were asked to return a response sheet for each case, giving their grading of 32 different histological features as defined in Table 2. This list was developed by discussion with participants. We included all features that have definitions in the Banff 97 [10] or CCTT [9] classifications, together with a selection of other features of personal interest to participants. More detailed definitions than those given in Table 2 were provided to all participants.

Table 2. Histological features assessed, with abbreviated definitions

Feature	Summary of definition
Tubulitis—grade	As in Banff 97 (t)
Tubulitis—extent	Number per 10 HPF [9]
Luminal neutrophils	Present or absent
Isometric vacuolation	Present or absent
Anisometric vacuolation	Present or absent
Other forms of acute tubular damage	Present or absent
Tubular atrophy	As in Banff 97 (ct), but expressed as % affected
Glomeruli: number present	Number in one section
Early type of allograft glomerulitis	As in Banff 97 (g)
Number of glomeruli completely sclerosed	Number in one section
Mesangial matrix increase	Number showing "moderate increase" (Banff mm) in one section
Glomeruli with segmental sclerosis	Simple count
Chronic allograft glomerulopathy	As in Banff 97 (cg)
Interstitial edema	% of cortex showing edema
Mononuclear cell interstitial infiltration	As in Banff 97 (i), but expressed as % affected
Interstitial fibrosis	As in Banff 97 (ci), but expressed as % affected
Large lymphocytes	Number per single HPF
Plasma cells	Number per single HPF
Eosinophils	Number per single HPF
Neutrophils	Number per single HPF
Interstitial hemorrhage	Present or absent
Infarction	Present or absent
Number of arterial cross sections	Number in one section
Arteriolar hyaline thickening	As Banff 97 (ah)
Endothelial cell activation—arterial	Graded 0 to 3
Endothelial cell activation—venous	Graded 0 to 3
Neutrophils in peritubular capillaries	Present or absent
Intimal arteritis	As Banff 97 (v)
Arteriolitis	Present or absent
Fibrous intimal thickening	As Banff 97 (cv)
Breaks in arterial elastica	Present or absent
Inflammatory cells in intima in chronic fibrosis	Present or absent

In this way, a total of 55 cases were studied, in 11 groups of 5 cases, over a period of approximately two years. Participants were asked to contribute sections that were technically adequate by the Banff criteria, but some centers found this difficult to achieve, and in retrospect some of the sections were found to be below this standard, though none were inadequate. Inevitably sections from different centers also had different staining characteristics. These problems were felt to be irrelevant to the evaluation of reproducibility, as the material available was the same for all participants, but they do impinge on any assessment of diagnostic accuracy, as discussed below.

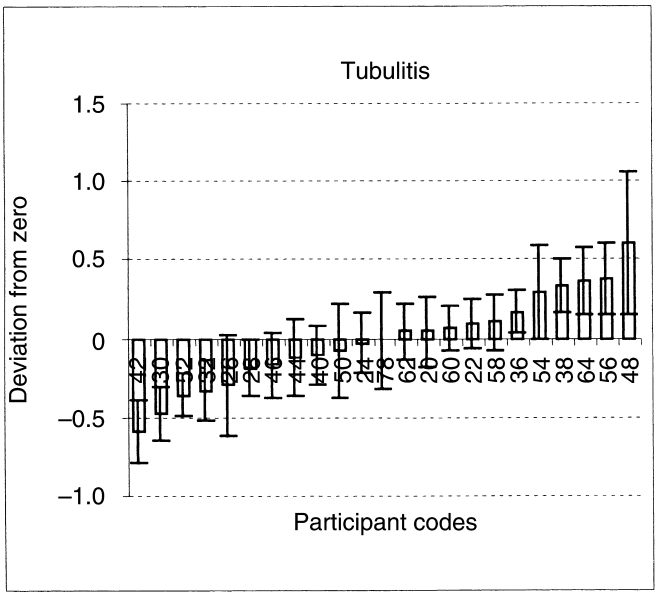
Feedback to participants

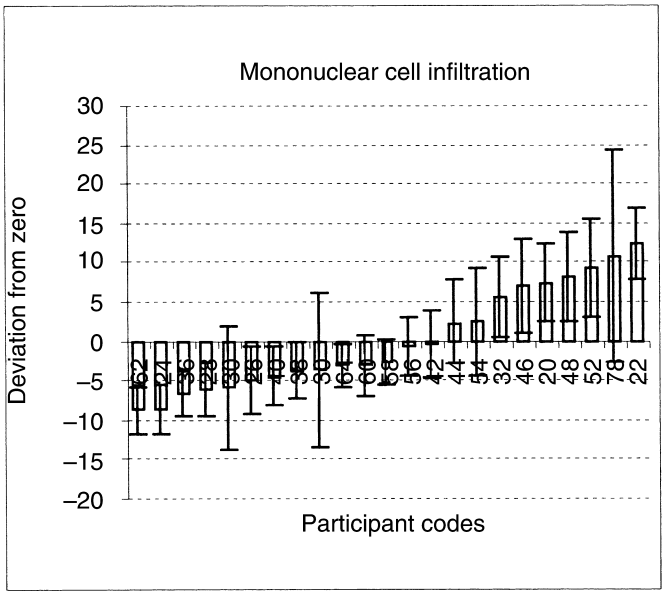
All of the responses were entered into a purpose-written database in the co-ordinating center in Leicester. At the end of each circulation, the average grade for each histological feature was calculated for each case and a report was produced for circulation to participants. Since each participant was identified in the database by a code number, a printout was produced for each participant informing how his/her assessment compared with the whole group. For example, tubulitis is graded on a scale of 0 to 3. The average tubulitis grade offered by all of the participants for all five cases in the first set was 1.1. If a participant was in the habit of "over-grading" tubulitis, that participant's average score for these five cases might be 1.6. This discrepancy would immediately be evident in the personal report. Participants were reminded at intervals that they should use this feedback to adjust their criteria for grading in order to move towards a consensus.

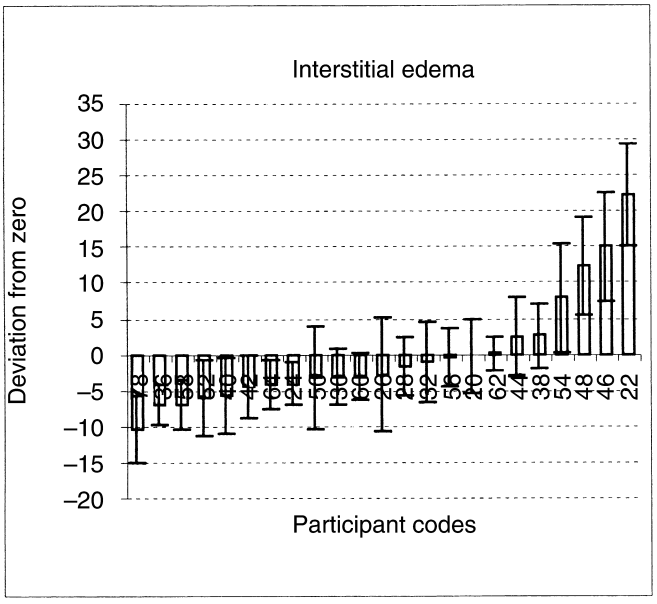
Statistical analysis

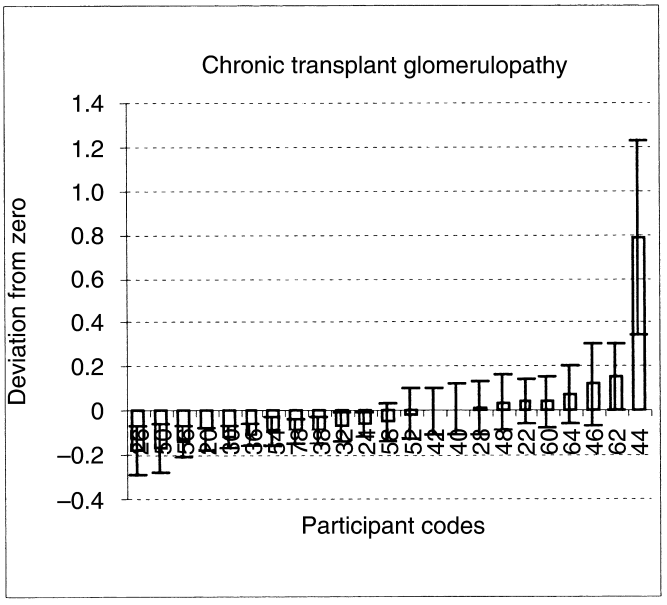
Do some pathologists consistently under-grade or over-grade? The average score for each feature was determined for each case. Then for each pathologist's grading for each feature for each case, the relevant average was subtracted from the individual pathologist's score. This

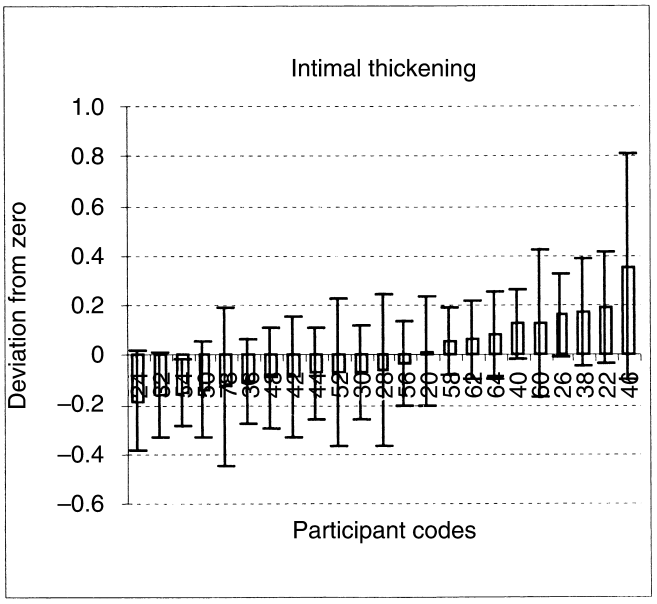
Fig. 1. Evidence of consistent under- or over-grading by individual pathologists, identified by code numbers (mathematical manipulations are explained in the text).

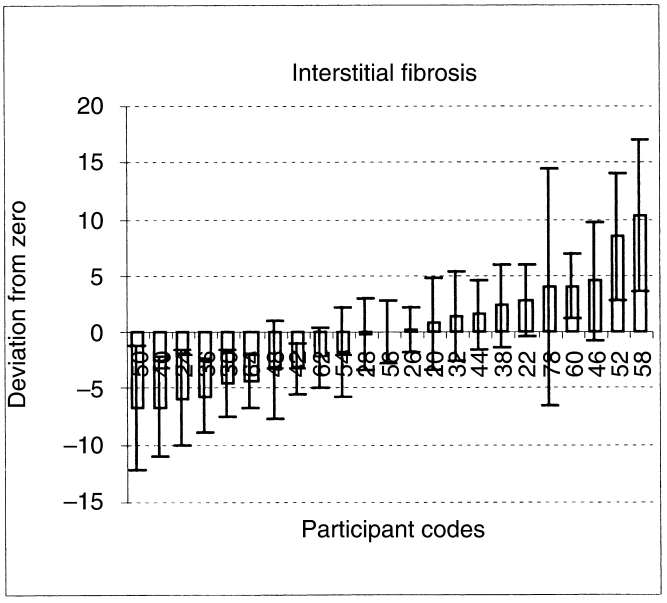

Tubulitis
Deviation from zero
1.5
1.0
0.5
0
−0.5
−1.0
Participant codes


Mononuclear cell infiltration
Deviation from zero
30
25
20
15
10
5
0
−5
−10
−15
−20
Participant codes


Interstitial edema
Deviation from zero
35
30
25
20
15
10
5
0
−5
−10
−15
−20
Participant codes


Chronic transplant glomerulopathy
Deviation from zero
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0
−0.2
−0.4
Participant codes


Intimal thickening
Deviation from zero
1.0
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
Participant codes


Interstitial fibrosis
Deviation from zero
20
15
10
5
0
−5
−10
−15
Participant codes

left a negative number if a pathologist under-scored this feature and a positive number if it had been over-scored. Next, the average of these "corrected to zero" figures was calculated for each pathologist for each feature, across all cases, together with 95% confidence intervals. The results were placed in rank order and plotted (Fig. 1). Random variation was expected, but where any pathologist had a 95% confidence limit that did not cross the zero axis, it represented evidence of systematic under- or over-scoring in comparison with the group.

Inter-observer variation. The analysis of inter-observer variation in a study of this type is mathematically complex. Cohen's kappa statistic is commonly employed, but this method was originally designed for ordinal data (for example, benign or malignant) rather than grades on a continuum, and it was originally designed for two observers rather than 22. To use kappa statistics when considering continuous variables, it is necessary to divide the variables into a number of defined ranges. Where available, the Banff definitions were used to do this on a four-point scale of 0 to 3 (for example, interstitial fibrosis, tubular atrophy). Where Banff definitions were not available [for example, numbers of cells per high-power field (HPF)], we examined the data and set levels that divided it into four groups of as equal a size as possible.

Most published methods for the calculation of kappa statistics also rely on a complete data set, which was not available to us because of intermittent failures of the postal system and loss of slides. To overcome the problem of the intermittently missing ratings, kappa statistics were calculated using the "one-way" multirater method described by Fleiss [11]. This does not use the information that it was the same set of observers who rated each of the slides.

To test whether there had been an improvement in interobserver reproducibility as a result of the ongoing feedback system, two approaches were used. First, the kappa statistic was calculated for each major feature within each of the sets of five cases. This produced a set of eleven kappa statistics for each histological feature. The numbers were plotted and subjected to linear regression to test the probability of the slope being greater than zero (Fig. 2).

This approach would be expected to reveal changes in interobserver reproducibility that were not uniform across the duration of the study, as such changes would produce a non-linearity of the plot. However, the relatively small numbers of observations in each group of five cases led to the production of kappa values with unacceptably large confidence limits; the majority crossed zero.

To resolve this dilemma, a second approach was used, in which observations from the first 25 cases and the last 30 cases were compared. Kappa values for the first and the second groups of slides were compared using the permutation test [12]. Each permutation consisted of randomly dividing the 55 slides into groups of 25 and 30 slides, respectively, and calculating the difference in kappa values between them. The resulting *P* value for the test was the proportion of these permutations where the absolute value of the difference in kappa was greater than the absolute value of the difference in kappa between the original groups of 25 and 30 slides. First, 1000 permutations were used, and then any features approaching statistical significance at the 5% level were retested with 10,000 permutations. The two-sided test was used in all cases.

To assess how individual pathologists were changing their practice over the course of the study, the "corrected to zero" figures described above were taken and divided into the first 25 and the last 30 cases. The average "score" for each pathologist for feature was calculated for the first and plotted to demonstrate graphically whether individual pathologists were using the feedback to "converge" toward the average scoring criteria (Fig. 3).

Correction for high-power field size. Several of the histological features depend upon an assessment of the frequency of a given event (for example, tubulitis, presence of a specific cell type) in a given area of the section, based on the area of the observer's high power field. As this area can be expected to differ between microscopes, participants were all asked to measure the diameter of their high power field using a stage micrometer. The area was calculated from the diameter. For those features that rely on this measurement, the numbers reported by the pathologists were corrected to a nominal area of 1 mm^2 and the above measures of inter-observer variation were repeated.

Does seeing different sections influence the results? In order to allow all the pathologists in this study to examine the same cases, it was not practical for everyone to examine exactly the same sections and six sets of slides were used. These were cut as serial sections, but it is nevertheless possible that differences between the sections could account for some of the interobserver variation. To test this hypothesis, the results for each histological feature were arranged into six groups, representing the six groups of pathologists; hence, within one group, the pathologists had actually seen the same sections. We then sought evidence of differences between the groups that could not have arisen by chance, by using the permutation test as described previously in this article [12].

Correlation with clinical outcome

Acute rejection. Of the 55 cases, 41 represented biopsies taken for the investigation of acute graft dysfunction and 14 were protocol biopsies with long follow-up. Retrospective study of changes in serum creatinine and response to therapy showed 26 to be cases of acute rejection, all of these obviously being within the "acute

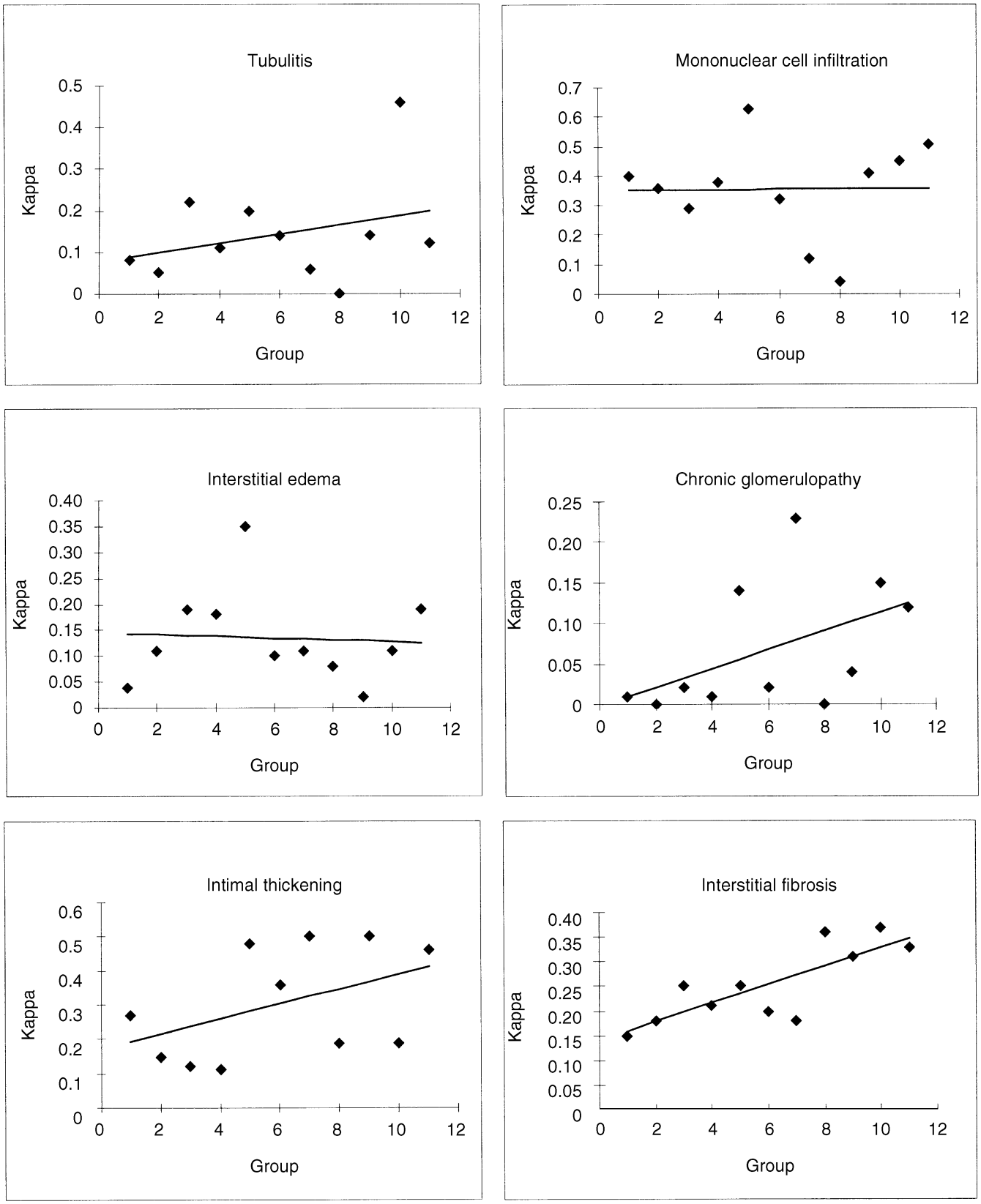


Fig. 2. Kappa values calculated for sequential groups of cases. Improving inter-observer agreement should be manifest as a line sloping upward to the right, but was statistically significant only for interstitial fibrosis ($P = 0.05$).

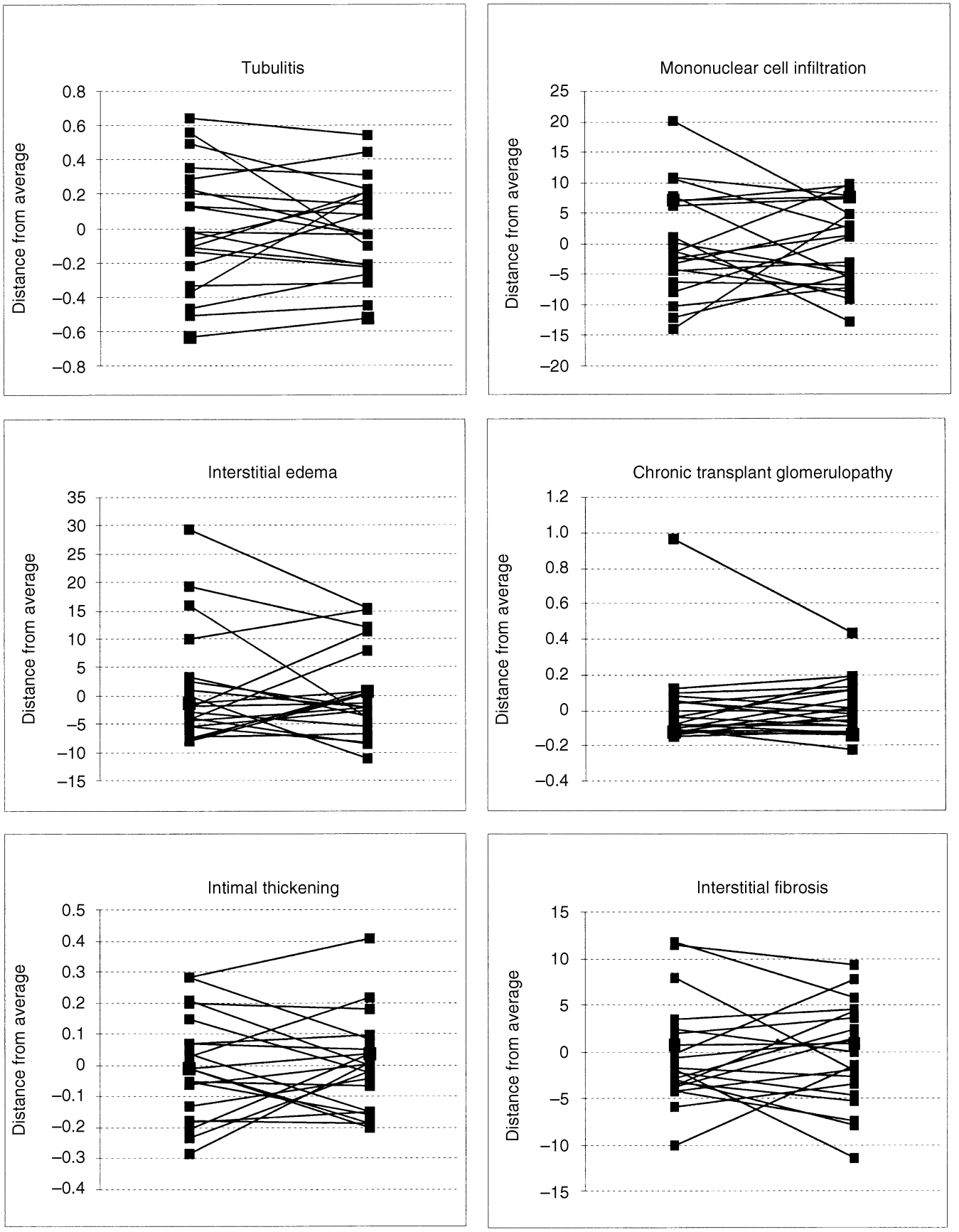


Fig. 3. Inter-observer variation as assessed in the first half versus the last half of the study. Improved agreement with the consensus of the group would be evident by a line that moves closer to zero from left to right. For most parameters it appears that some participants converge, but others do not adjust their criteria.

dysfunction" group. The scores for each histological feature were tested for association with this diagnosis by Mann-Whitney U test, using the MiniTab statistics program.

Testing for acute rejection. Various manipulations of the data were performed to assess the accuracy of different approaches to making a diagnosis of acute rejection, in the absence of any clinical data. Each pathologist's response to each of the 41 "acute" cases was analyzed to test whether the grades given would result in a diagnosis of acute rejection:

1. If Banff "suspicious" [10] or above was taken as indicating acute rejection;
2. If Banff "acute rejection type 1a" or above was required before diagnosis acute rejection;
3. Using the CCTT criteria [9].

It is arguable that the average score of such a large number of pathologists represents the best possible estimate of the severity of a histological feature within a given biopsy. The diagnostic value of these "consensus scores" was tested in the same way.

We have previously demonstrated the value of using computer-based expert support systems in integrating complex data sets to make a diagnosis of acute rejection [13, 14]. Hence, we designed an inference network using the commercially available "Netica" software package (Norsys Corporation) incorporating the 12 histological features with the highest *P* values when tested for association with acute rejection. The network was trained by entering the probability of acute rejection associated with each grade of these 12 variables, obtained using the data from all the "acute dysfunction" cases. It was then tested for its accuracy in diagnosing acute rejection using the same cases. We recognized that testing such a network with cases that have been used in its training is less than ideal, but the limited number of cases available within this study precluded a more thorough test.

Chronic decline in function. Data were used only from the 14 protocol biopsies, taken during stable graft function. These biopsies were all taken from 5 to 10 years ago; thus, the decline in graft function was assessed simply by taking the most recent available creatinine level, subtracting from this the creatinine level at the time of biopsy, and dividing the result by the time between the two measurements. This resulted in a split between six "stable" grafts (creatinine improved, unchanged or increased by less than 50 μmol/L over the period of follow-up) and eight "declining" grafts (graft failure or creatinine increased by more than 150 μmol/L over the period of follow-up). Significant differences between these groups were sought by Mann-Whitney U test, as described previously in this article for acute rejection.

Table 3. Reproducibility of assessment of the histological features

Feature	Kappa
Tubulitis—Banff grade	0.17
Tubulitis per 10 HPF	0.16
Luminal neutrophils	0.33
Isometric vacuolation	0.19
Anisometric vacuolation	0.12
Other forms of acute tubular damage	0.14
Tubular atrophy	0.29
Glomeruli: number present	0.53
Early type of allograft glomerulitis	0.21
Number of glomeruli completely sclerosed	0.47
Mesangial matrix increase	0.12
Glomeruli with segmental sclerosis	0.13
Chronic allograft glomerulopathy	0.11
Interstitial edema	0.17
Mononuclear cell interstitial infiltration	0.34
Interstitial fibrosis	0.30
Large lymphocytes	0.13
Plasma cells	0.13
Eosinophils	0.17
Neutrophils	0.05
Interstitial hemorrhage	0.32
Infarction	0.12
Number of arterial cross sections	0.19
Arteriolar hyaline thickening	0.11
Endothelial cell activation—arterial	0.21
Endothelial cell activation—venous	0.10
Neutrophils in peritubular capillaries	0.13
Intimal arteritis	0.35
Arteriolitis	0.24
Fibrous intimal thickening	0.36
Breaks in arterial elastica	0.22
Inflammatory cells in intima in chronic fibrosis	0.34

Kappa = 1 indicates perfect agreement; kappa = 0 indicates completely random results.

RESULTS

Do pathologists consistently under-grade or over-grade?

Similar results were found for every histological feature, apart from those that were so infrequent as to preclude analysis (for example, necrosis, interstitial hemorrhage). A selection of representative graphs of the more important histological features is shown in Figure 1. Examination of the error bars (representing 95% confidence limits) demonstrates that for every feature, some pathologists consistently over-grade and some under-grade. Examination of the participant codes (along the horizontal axis) shows a different sequence of codes for each case, with no discernable pattern; hence, a pathologist who consistently over-grades one feature does not seem to be more likely to over-grade another.

Inter-observer variation

Measurement of inter-observer variation is difficult and contentious, as discussed above. The usual practice

of defining ranges of kappa values as "good," "acceptable," and "poor" reproducibility ignores the fact that when a continuous variable is split into many grades (for example, tubulitis) the "agreement" between observers is inevitably less likely than if the same feature was merely recorded as "present" or "absent." Furthermore, the kappa value is likely to be estimated very inaccurately for those features that were present very infrequently (for example, intimal arteritis). Subject to these problems in interpretation, the overall kappa statistics for each feature are shown in Table 3. We did not formally test whether reproducibility varied significantly between biopsies of different type (for example, acute rejection versus not acute rejection). However, from visual examination of the data it was clear that for several variables (notably interstitial fibrosis), agreement was much better when the abnormality was absent. Whenever it was present, agreement on the severity was poor.

There was only limited evidence of improvement in inter-observer variation over the course of this study. When the kappa statistics for each group of five cases were calculated and plotted, the slope of the resultant regression line was upward (indicating improving agreement) for most variables, but with the exception of interstitial fibrosis ($P = 0.003$) this slope was never significantly greater than zero at the 5% level of significance (Fig. 2). Analysis the first 25 and last 30 cases also failed to show convincing improvements in reproducibility (Table 4).

A possible explanation for this disappointing result was evident when individual variation from the mean was plotted for the first 25 and last 30 cases (Fig. 3). Visual inspection of these graphs suggests that there are many lines which converge towards the zero axis from left to right, suggesting that these pathologists have achieved "convergence" toward the average criteria for the group, indeed, a few "overshoot." This effect is perhaps most prominent for tubulitis. However, there are also many participants where the line is remarkably horizontal, indicating that the feedback provided has had no impact on that pathologist's systematic over-grading or under-grading.

It must also be accepted that, due to the limited amount of tissue available, we did not include sections stained to highlight connective tissue or elastin and this may have contributed to the variation seen in some histological features.

Correction for high-power field size

The existence of a large variation in the area of the "high-power field" was confirmed. Two extreme measurements were eliminated as measurement errors and one pathologist could not give a figure as he had used several different microscopes in the course of the study. Even without these measurements the variation was very large (Fig. 4).

Correcting the relevant observations for the area of the high power field of the observer produced no improvement in the inter-observer variation. Such "corrected" data were not used further in the analysis.

Does seeing different sections influence the results?

For the histological features listed in Table 5, there was evidence of significant differences between the responses of pathologists who saw different sections from the same biopsy. For all other histological features, no differences were detectable, suggesting that the compromise of using a set of serial sections rather than everyone viewing the same sections had not introduced spurious inter-observer variation.

Correlation with clinical outcome

Acute rejection. The features that correlated with a subsequent clinical diagnosis of acute rejection are listed in Table 6. Despite the highly significant differences between the groups, plots of the mean values for each case show considerable overlap with every histological feature (Fig. 5). The values plotted are not individual opinions; they are the average scores from the values given by all the pathologists in the study. Therefore, it can be argued that they represent the best available estimate of the score of each feature in each available biopsy. The overlap confirms that no one histological feature is sufficient to make a reliable diagnosis of acute rejection in isolation.

Testing for acute rejection. The Banff or CCTT criteria were applied to the histological grades given for all the biopsies. The results were compared with the retrospective clinical diagnosis (acute rejection or not acute rejection) as defined during the case selection process. The numbers of correct diagnoses produced are shown in Table 7. This approach is not relevant to the accuracy of diagnosis in routine practice, because of the limited material available and the absence of any clinical information. It was used to make a comparison between they way in which two classification systems are used to produce a diagnosis, and we argue that this comparison is valid because exactly the same limitations apply to both. Not surprisingly, if Banff "suspicious" is accepted as indicating acute rejection, there is over-diagnosis of acute rejection, whereas if Banff "suspicious" is excluded, acute rejection is under-diagnosed. The CCTT criteria produce results that are intermediate between these two, but do not give an increase in the total number of correct diagnoses.

The computer-based inference network, which uses a much larger number of relevant histological features than either of the published schemas, produced correct diagnoses in 87% of the cases, using morphological data

Table 4. Evidence of decreased inter-observer variation over the course of the study

Feature	*P* value	
	Analysis in 11 groups of 5 cases	Analysis of first 25 vs. last 30 cases
Tubulitis—Banff grade	0.337	0.614
Tubulitis per 10 HPF	0.653	0.850
Luminal neutrophils	0.322	0.550
Isometric vacuolation	0.466	0.168
Anisometric vacuolation	0.519	0.812
Other forms of acute tubular damage	0.162	0.021
Tubular atrophy	0.576	0.866 (Decreased)
Glomeruli: number present	0.017	0.115
Early type of allograft glomerulitis	0.284 (Decreased)	0.356 (Decreased)
Number of glomeruli completely sclerosed	0.505	0.114
Mesangial matrix increase	0.147	0.602
Glomeruli with segmental sclerosis	0.221	0.472
Chronic allograft glomerulopathy	0.143	0.358
Interstitial edema	0.852 (Decreased)	0.134 (Decreased)
Mononuclear cell interstitial infiltration	0.818	0.166 (Decreased)
Interstitial fibrosis	0.05	0.030
Large lymphocytes	0.116	0.352
Plasma cells	0.274 (Decreased)	0.667 (Decreased)
Eosinophils	0.595	0.608
Neutrophils	0.031	0.225
Interstitial hemorrhage	0.635	0.540
Infarction	0.582	0.046
Number of arterial cross sections	0.276	0.228
Arteriolar hyaline thickening	0.533	0.618
Endothelial cell activation—arterial	0.864 (Decreased)	0.754
Endothelial cell activation—venous	0.774 (Decreased)	0.806 (Decreased)
Neutrophils in peritubular capillaries	0.006	0.031
Intimal arteritis	0.989 (Decreased)	0.844 (Decreased)
Arteriolitis	0.637	0.836
Fibrous intimal thickening	0.167	0.164
Breaks in arterial elastica	0.637	0.268
Inflammatory cells in intima in chronic fibrosis	0.537	0.816

"Decreased" indicates that the inter-observer agreement appeared to decrease slightly, though in no case was this statistically significant.

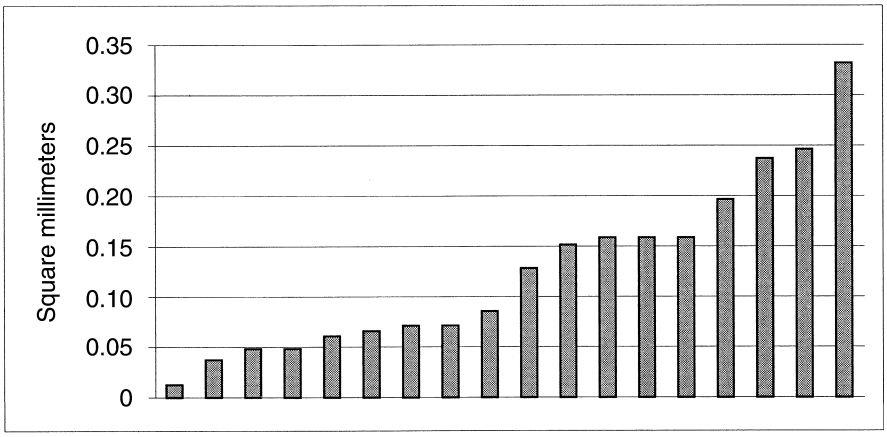


Fig. 4. Variation in the size of participants' "high-power field."

Table 5. Histological features showing evidence of differences between different sets of serial sections

Histological feature	*P* value
Sclerosed glomeruli	0.003
Number of glomeruli	0.006
Neutrophils in tubules	0.028
Elastic breaks	0.032
Activated lymphocytes/HPF	0.032

only. However, it must again be stressed that testing a network with the cases that have been used to train it may overestimate its performance.

Chronic decline in function. The features in the 14 protocol biopsies that correlated with a more rapid decline in renal function are listed in Table 8. It is notable that the most significant features are not those showing histological evidence of chronic damage, such as interstitial fibrosis or tubular atrophy, but are features that are

Table 6. Histological features which, when found in a biopsy taken to investigate acute graft dysfunction, showed a positive association with a retrospective clinical diagnosis of acute rejection

Histological feature	*P* value
Mononuclear cell infiltration	0.0008
Tubulitis (Banff grade)	0.0015
Tubulitis (count)	0.003
Interstitial edema	0.0036
Intimal arteritis	0.012
Large mononuclear cells	0.0013
Acute glomerulitis	0.027
Venous endothelial changes	0.047

normally considered to be evidence of acute damage. "Chronic" features such as interstitial fibrosis, tubular atrophy, intimal fibrosis, and chronic transplant glomerulopathy did not reach statistical significance, although it must be admitted that in this part of the study the sample size is relatively small.

DISCUSSION

This study has revealed large interobserver variation in the assessment of renal transplant biopsies, considerably larger than has been reported previously [7, 8]. To some extent, this is not surprising when the design of the study is considered. The participants had never worked together before. They had mostly trained in different countries, under different regimes, and before this study there had been no way other than verbal descriptions and published photographs to compare diagnostic criteria with pathologists elsewhere in the world.

Schemes for histological grading such as the Banff classification are intended to have worldwide application, so it can be argued that the measurement of interobserver variation in this study is considerably more relevant to the "real world" than studies involving small groups of colleagues. It is therefore appropriate to take the two stated aims of the Banff classification, and consider the implications of these results for each.

Implications for management of individual patients

The first stated aim of the Banff classification is "*to guide therapy in transplant patients*" [1]. Looking at the data presented here one might be tempted to wonder why transplant biopsies are considered to be useful in the management of individual patients. The proportions of correct diagnoses in Table 7 are alarmingly low, and there is no clear advantage between the Banff and CCTT approaches. Yet biopsies are widely believed to be useful in the diagnosis of acute graft dysfunction. Our results appear irreconcilable with previous reports of excellent concordance in the histological diagnosis of acute rejection, which are reported to be as high as 97.7% [15].

There are three explanations for this apparent discrepancy.

The first is the very severe and artificial limitations under which the pathologists were working. All clinical information was withheld. To provide enough sections from small needle biopsies each pathologist was provided with only two sections per case. Stains to highlight elastin and connective tissue were not included. As noted previously in this article, some sections did not fulfill the Banff definition of technical adequacy and interinstitutional differences in processing and staining caused problems for some pathologists. The accurate clinical diagnosis or exclusion of acute rejection may be difficult, even in retrospect, and may have been inaccurate in some cases. These problems do not detract from a study of reproducibility—indeed, with smaller samples one might expect less scope for disagreement—but they do make it much more difficult to arrive at a correct clinical diagnosis.

The second explanation is that pathologists do not, in practice, base diagnoses of acute rejection solely on the degree of tubulitis, interstitial mononuclear cell infiltration, and intimal arteritis. Whether consciously or not, the pattern recognition skills of an experienced histopathologist will include consideration of other features, such as the extent and nature of the infiltrate, the presence of edema, and the extent as well as the severity of the tubulitis. We have demonstrated, informally in this study and more rigorously elsewhere [13, 14] that computer-based decision support systems can integrate such complex datasets and come to a much more reliable diagnosis of acute rejection than can be achieved by considering only the Banff grades for tubulitis, mononuclear infiltration and intimal arteritis. Such computer-based data integration presumably mimics more closely the processes in the brain of an experienced pathologist. The Banff classification's approach to acute rejection is principally directed toward grading acute rejection, and in this it has been shown to be successful [2–6]. In excluding all but a small subset of the histological features it is not designed to provide ideal diagnostic precision for individual patient care.

The third point is that individual management decisions are not made in the absence of clinical information. In the first description of the Banff classification, it was stressed that decisions of clinical management should be made in the light of the clinical setting, and "individual centers will develop their own strategies for dealing with various biopsy findings." It is likely that in some centers a finding of mild tubulitis will usually trigger treatment of acute rejection, whereas in others mild tubulitis will usually be ignored. This study demonstrates that such differences may reflect differences between pathologists rather than between patient populations. Hence, as long as close liaison between the clinical team and the pathologist permits ongoing feedback about the accuracy of

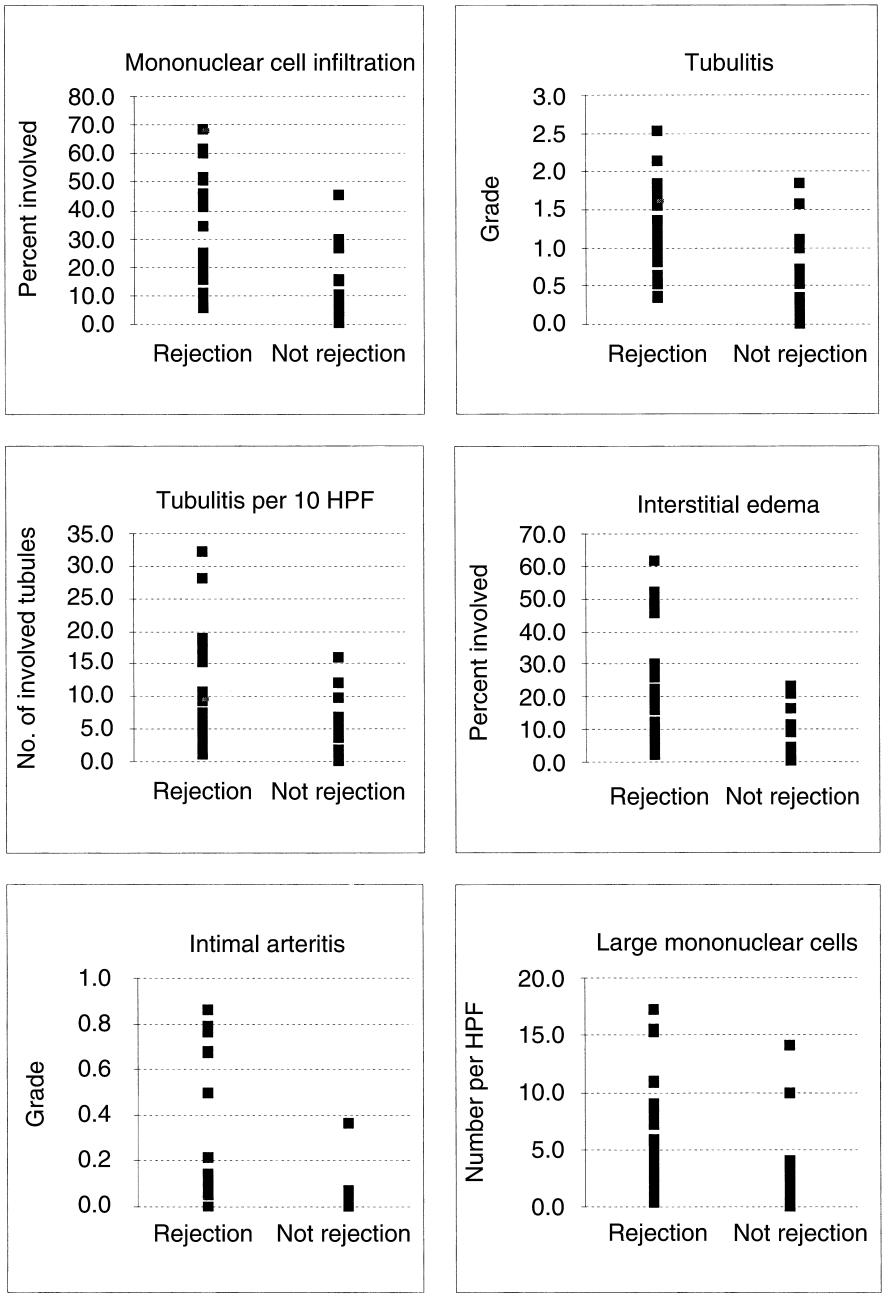


Fig. 5. Distribution of average grades for histological features associated with acute rejection. No single feature permits reliable discrimination.

diagnoses, such variations in histological grading need not be reflected in differences in patient management.

In addition to the diagnosis of acute rejection, it would be of great value if one could predict graft survival at an early stage of engraftment. Most biopsy-based studies that have addressed this problem have considered histological evidence of chronic damage such as interstitial fibrosis and tubular atrophy [16–20]. More recently, there has been emphasis on the concept of "subclinical acute rejection" as a cause of chronic graft failure—one that is potentially reversible [21]. The present study shows a striking correlation between "acute" features such as tubulitis and lymphocytic infiltration, rather than "chronic" features. The number of protocol biopsies in this study is small, but the result provides support for the importance of "subclinical acute rejection" in protocol biopsies as an important prognostic feature.

Implications for clinical trials and interpretation of published studies

The second aim of the Banff classification was "*to help establish an objective rejection end point in clinical trials*" [1]. Here interinstitutional variation in grading can be a major problem. Our results confirm that when biopsy

Table 7. Performance of published schemas in identifying cases of acute rejection from morphological data only, applying cut-off for diagnosis at two different points in the Banff schema

Correct diagnoses using pathologists' individual responses					
Cases of acute rejection (417 responses)			Cases not acute rejection (519 responses)		
Banff "suspicious"	Banff "acute rejection 1a"	CCTT	Banff "suspicious"	Banff "acute rejection 1a"	CCTT
307 (74%)	192 (47%)	253 (61%)	269 (52%)	407 (79%)	305 (59%)
Correct diagnoses using average grades given by the whole group					
Cases of acute rejection (26 cases)			Cases not acute rejection (29 cases)		
Banff "suspicious"	Banff "acute rejection 1a"	CCTT	Banff "suspicious"	Banff "acute rejection 1a"	CCTT
21 (81%)	12 (46%)	20 (77%)	13 (45%)	23 (79%)	14 (48%)

Table 8. Histological features which, when found in a protocol biopsy, showed a positive correlation with a subsequent more rapid decline in renal function

Histological feature	*P* value
Edema	0.006
Tubulitis grade	0.006
Eosinophil infiltration	0.013
Tubulitis per 10 HPF	0.023
Mononuclear cell infiltration	0.061

assessments are part of a clinical trial, these assessments must be carried out in a single center, preferably by a single pathologist. One cannot rely on results generated from histological grading if the grading is done in the various laboratories associated with a multicenter trial, as any useful signal is likely to be lost in the "noise" of interinstitutional variation.

Our results also show that great caution must be exercised when comparing biopsy results between institutions. We have never really known why some institutions report levels of acute rejection, even in protocol biopsies, which are much higher than others, despite apparently similar immunosuppressive regimens [22]. There is evidence that some of this variation is genuine, particularly where the patient population has different ethnic origins [23], but it seems very likely that part of this variation can be explained by variation in grading criteria between different pathologists.

This does not mean that published histological grades are meaningless. For example, it has recently been confirmed that within an episode of acute rejection, the grade of tubulitis correlates with the prognosis [5]. A correlation between more severe types of rejection and a poor prognosis has been reported by several groups [3, 4, 6, 24]. Such conclusions can be expected to remain true in other institutions, even if a biopsy that is called "Suspicious" in one institution is called "type 1b acute rejection" in another. Nevertheless, our results do mean that comparisons of biopsy findings between institutions require considerable care if the biopsies are reported by different pathologists. They also explain the anecdotal observation of transplant units that when a new pathologist is appointed, a period of "settling in" is required before mutual trust is established.

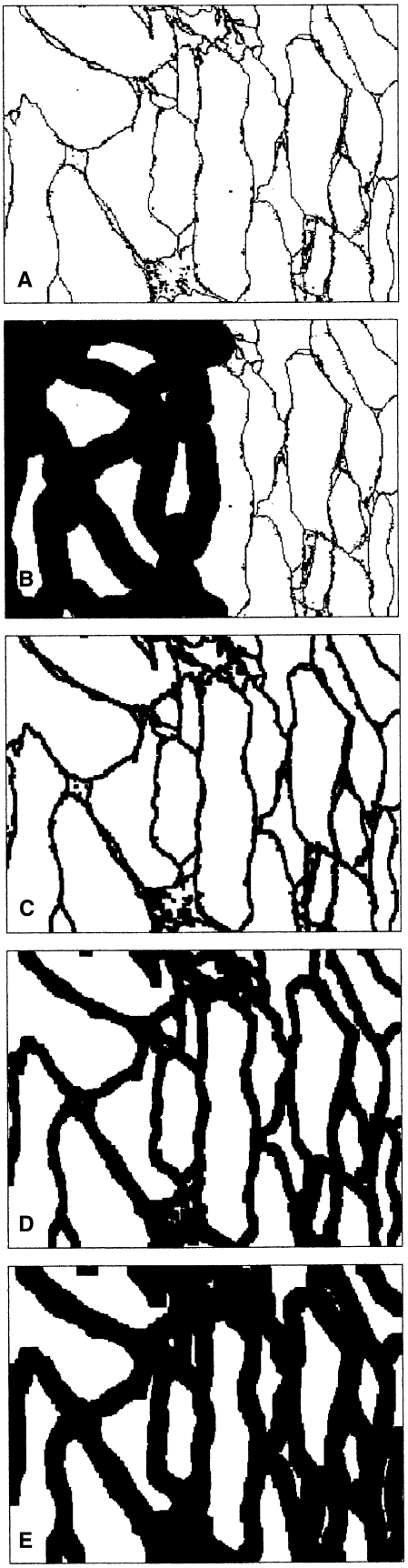
Implications for the future interpretation of transplant biopsies

Within one institution, the large variation reported in this study may have no impact at all on routine care. If the pathologist reports changes consistently and if these changes are interpreted in the light of local experience, then correct clinical decisions will be made even if the pathologist's grading is away from any international consensus. "Borderline" may always mean acute rejection in one institution and never in another. If all concerned are aware of this, patient care will not necessarily suffer.

Despite this caveat, a reduction in interinstitutional variation is obviously desirable for many reasons. In the short term, this must take the form of improving our existing approaches. In the long term, perhaps we should look harder for alternatives to conventional histological assessment of transplant biopsies.

To improve our existing approach there are two options: improved training and improved criteria.

International co-ordination of training will be difficult, but perhaps not impossible. International travel to training sessions might improve consistency, but this is too expensive for most laboratories, and which laboratory should offer the definition of what is correct? It is unlikely that any one laboratory could offer the resources to train all of the transplant pathologists in the world. For this reason we had hoped that the slide circulation and feedback system developed for this study could be extended to other laboratories, following the pattern of histopathology external quality assessment schemes in the United Kingdom [25, 26]. Unfortunately, the level of convergence we have produced does not justify this approach. Participants in this study offered several expla-



nations, including "information overload" as a result of using 32 histological features. The results in Figure 3 also suggest that some pathologists were reluctant to adjust their long-held opinions as to "correct" practice, even when shown to be in a minority. We are currently evaluating a slide circulation scheme that is limited to a smaller number of "critical" variables, and a scheme that involves the circulation of images rather than slides.

Improved definitions also are likely to help. Ongoing development of the Banff definitions has always been envisaged, and some major improvements were introduced in 1997 [10]. However, there are still ambiguities. Several of the definitions rely on an assessment of the "area affected." This is easy when a change is present or absent, but much harder (as with most biological processes) when the alteration develops gradually (Fig. 6). Discussion of such questions will be a feature of forthcoming Banff conferences. It is clear from our results that any definition based on the size of a microscope's high power field should be avoided.

There will, however, be a limit to the reproducibility of grading by histopathologists. The long-term solution is likely to involve replacement of such subjective grading by more objective measurements. There have already been reasonably successful attempts to diagnose rejection by measurement of gene expression [27, 28]. Several groups have shown that chronic damage can be measured more reproducibly using computerized image analysis [16–20]. However, these methods have been subjected to very little clinical evaluation when compared with years of experience of conventional biopsy assessment. New methods will carry their own, as yet unmeasured, interinstitutional variation. Implementation in routine practice will require prolonged, detailed, multicenter valuation. If the "gold standard" of the biopsy is somewhat tarnished, against what can the new methods be evaluated? This will require close collaboration between clinicians, histopathologists and molecular biologists. It may be difficult to obtain resources to support such work.

ACKNOWLEDGMENTS

This project is supported by a grant from the European Union: Standards Measurement and Testing Programme, Contract no. SMT4-CT98-7514. We are grateful to Dr. K. Solez for constructive discussions throughout the course of this project. We have benefited from discussion of this project with members of staff of the following pharmaceutical companies: Bristol Myers Squibb, Fujisawa, IMTIX-Sangstat, No-

Fig. 6. The problem of assessing area affected by aprogressive process. (*A*) 1-bit digital image of renal interstitium (Sirius red staining through crossed polaroids, digital negative). (*B-E*) digital manipulations of (A). If involvement is heavy and patchy (B) assessment of 'area affected' may be reliable, but if there is diffuse gradual expansion (C-E) how can one decide consistently when 0% involvement suddenly becomes 100% involvement?

vartis, and Roche. These companies have not provided any financial or material support.

Reprint requests to Dr. P.N. Furness, Department of Pathology, Leicester General Hospital, Gwendolen Road, Leicester LE5 4PW, England, United Kingdom.
E-mail: pnf1@le.ac.uk

REFERENCES

1. Solez K, Axelsen RA, Benediktsson H, *et al*: International standardization of criteria for the histologic diagnosis of renal allograft rejection: The Banff working classification of kidney transplant pathology. *Kidney Int* 44:411–422, 1993
2. Corey HE, Greenstein SM, Tellis V, *et al*: Renal allograft rejection in children and young adults: The Banff classification. *Pediatr Nephrol* 9:309–312, 1995
3. Gaber L, Schroeder T, Moore L, *et al*: The correlation of Banff scoring with reversibility of first and recurrent rejection episodes. *Transplantation* 61:1711–1715, 1996
4. Gaber LW, Moore LW, Gaber AO, *et al*: Correlation of histology to clinical rejection reversal: A thymoglobulin multicenter trial report. *Kidney Int* 55:2415–2422, 1999
5. Minervini MI, Torbenson M, Scantlebury V, *et al*: Acute renal allograft rejection with severe tubulitis (Banff 1997 grade IB). *Am J Surg Pathol* 24:553–558, 2000
6. Waiser J, Schreiber M, Budde K, *et al*: Prognostic value of the Banff classification. *Transplant Int* 13(Suppl 1):S106–S111, 2000
7. Solez K, Hansen HE, Kornerup HJ, *et al*: Clinical validation and reproducibility of the Banff schema for renal allograft pathology. *Transplant Proc* 27:1009–1011, 1995
8. Marcussen N, Olsen TS, Benediktsson H, *et al*: Reproducibility of the Banff classification of renal allograft pathology: Inter- and intra-observer variation. *Transplantation* 60:1083–1089, 1995
9. Colvin R, Cohen A, Saiontz C, *et al*: Evaluation of pathologic criteria for acute renal allograft rejection: Reproducibility, sensitivity and clinical correlation. *J Am Soc Nephrol* 8:1930–1941, 1997
10. Racusen LC, Solez K, Colvin RB, *et al*: The Banff 97 working classification of renal allograft pathology. *Kidney Int* 55:713–723, 1999
11. Fleiss JL: *Statistical Methods for Rates and Proportions* (2nd ed). Chichester, Wiley, 1981, pp 225–232
12. Efron B, Tibshirani RJ: *An Introduction to the Bootstrap*. London, Chapman & Hall, 1993, pp 170–174
13. Furness PN, Levesley J, Luo Z, *et al*: A neural network approach to the biopsy diagnosis of early acute renal transplant rejection. *Histopathology* 35:461–467, 1999
14. Kazi JI, Furness PN, Nicholson M: Diagnosis of early acute renal allograft rejection by evaluation of multiple histological features using a Bayesian belief network. *J Clin Pathol* 51:108–113, 1998
15. Pirsch JD, Miller J, Deierhoi MH, *et al*: A comparison of tacrolimus (FK506) and cyclosporine for immunosuppression after cadaveric renal transplantation. FK506 Kidney Transplant Study Group. *Transplantation* 63:977–983, 1997
16. Nicholson ML, Bailey E, Williams S, *et al*: Computerized histomorphometric assessment of protocol renal transplant biopsy specimens for surrogate markers of chronic rejection. *Transplantation* 68:236–241, 1999
17. Seron D, Moreso F, Bover J, *et al*: Early protocol renal allograft biopsies and graft outcome. *Kidney Int* 51:310–316, 1997
18. Dimeny E, Wahlberg J, Larsson E, Fellstrom B: Can histopathological findings in early renal allograft biopsies identify patients at risk for chronic vascular rejection? *Clin Transplant* 9:79–84, 1995
19. Isoniemi H, Taskinen E, Hayry P: Histological chronic allograft damage index accurately predicts chronic renal allograft rejection. *Transplantation* 58:1195–1198, 1994
20. Isoniemi HM, Krogerus L, von Willebrand E, *et al*: Histopathological findings in well-functioning, long-term renal allografts. *Kidney Int* 41:155–160, 1992
21. Rush D, Nickerson P, Gough J, *et al*: Beneficial effects of treatment of early subclinical rejection: A randomized study. *J Am Soc Nephrol* 9:2129–2134, 1998
22. Jain S, Curwood V, Kazi J, *et al*: Acute rejection in protocol renal transplant biopsies-institutional variations. *Transplant Proc* 32:616, 2000
23. Kazi JI, Furness PN, Nicholson M, *et al*: Interinstitutional variation in the performance of Baysian Belief Network for the diagnosis of acute renal graft rejection. *Transplant Proc* 31:3152, 1999
24. Mueller A, Schnuelle P, Waldherr R, van der Woude FJ: Impact of the Banff '97 classification for histological diagnosis of rejection on clinical outcome and renal function parameters after kidney transplantation. *Transplantation* 69:1123–1127, 2000
25. Furness P: *The U.K. National Renal Pathology External Quality Assessment Scheme Home Page.* http://www.le.ac.uk/pa/pnf1/eqa/
26. Furness P, Codling B, Goldie D, *et al*: *Recommendations for the Development of Histopathology/Cytopathology External Quality Assessment Schemes.* London, Royal College of Pathologists, 1998
27. Suthanthiran M: Molecular analyses of human renal allografts: Differential intragraft gene expression during rejection. *Kidney Int* 51(Suppl 58):S15–S21, 1997
28. Suthanthiran M: Clinical application of molecular biology: A study of allograft rejection with polymerase chain reaction. *Am J Med Sci* 313:264–267, 1997