J Clin Pathol 1998;51:108–113

Diagnosis of early acute renal allograft rejection by evaluation of multiple histological features using a Bayesian belief network

J I Kazi, P N Furness, M Nicholson

Abstract

***Background and aims*—The development of the Banff classification of renal transplant pathology has allowed the standardisation of approaches to transplant biopsy histology and reduced interobserver and interdepartmental variation. The usefulness of the Banff classification in the diagnosis of acute rejection has previously been tested by sending sections from 21 "difficult" biopsies to almost all of the renal transplant pathologists in the UK. Although the Banff classification improved reproducibility, the accuracy of diagnosis of early acute rejection was unchanged from the "conventional" approach. Perhaps this is because in making a diagnosis of acute rejection, the Banff classification uses only two features: tubulitis and intimal arteritis. To include more features on a systematic basis would be laborious for a human observer. Therefore, a Bayesian belief network was developed for this task.**

***Methods*—The network was initialised with observations from 110 transplant biopsies. Its performance was then tested on 21 biopsies that had been seen by 37 different renal transplant pathologists in an earlier study. These biopsies had been selected to represent histologically difficult problems but, in retrospect, they all had clear diagnoses of rejection or non-rejection on clinical grounds.**

***Results*—Using the Bayesian belief network, a relatively inexperienced pathologist made 19 of 21 correct diagnoses, better than had been achieved by any of the pathologists who had seen the same sections previously (17 of 21), and considerably better than the average proportion of correct diagnoses provided by all 37 renal transplant pathologists (65%). Application of the system by a second pathologist produced a tendency to overdiagnosis of acute rejection, illustrating the consequences of interobserver variation.**

***Conclusions*—In the diagnosis of acute rejection, further useful information can be extracted from features that are currently not considered in the Banff classification. Integration of data by a computer can give a more reliable diagnosis of early acute rejection, but routine application will require the development of a more sophisticated system that can also accommodate clinical data, perhaps one that can continue to "learn" as more data are entered.**

(J Clin Pathol 1998;51:108–113)

Keywords: Bayesian belief network; renal allograft rejection; Banff classification

Department of Pathology, Leicester General Hospital, Gwendolen Road, Leicester LE5 4PW, UK
P N Furness

Department of Surgery, Leicester General Hospital
M Nicholson

Department of Pathology, Sindh Institute of Urology and Transplantation, Dow Medical College and Civil Hospital, Karachi—74400, Pakistan
J I Kazi

Correspondence to:
Dr Furness.
email: pnf1@le.ac.uk

Accepted for publication
9 December 1997

A histological diagnosis of acute renal allograft rejection is rarely questioned. However, the morphological appearances of acute rejection develop gradually, so it is not surprising that in centres where allograft biopsy is used early in the investigation of graft dysfunction, pathologists often have difficulty in making this gold standard diagnosis with certainty. The problem has been highlighted by the finding that many of the supposedly specific features of acute rejection, such as tubulitis, can be found quite often in protocol based biopsies of stable grafts with good function.[1 2]

The Banff classification of renal transplant pathology[3] provides a rational basis by which the severity of a variety of histological features, including acute rejection, can be graded. The reproducibility of this system has been widely tested.[4–8] There is good evidence that its clear definitions reduce interobserver and interinstitution variation. It is therefore of great value in research into clinical graft rejection. However, a benefit has not been proven in terms of improved accuracy of the diagnosis or exclusion of an acute rejection episode for the individual patient.

Recently, we completed a study in which renal transplant biopsies were selected on the basis that they had caused difficulty in the diagnosis or exclusion of acute rejection. Cases were only used if a retrospective review of clinical data provided a clear diagnosis, using strict criteria. Sections were circulated to almost all the pathologists in the UK who routinely report such biopsies. We found the expected improvement in reproducibility of diagnosis if the Banff system had been used, when compared with an informal, "conventional" approach.[9] Unfortunately, when the subsequent clinical course was used to define the presence or absence of acute rejection, using the Banff classification resulted in no improvement in the number of correct diagnoses.[9]

We suspected that this was because the Banff classification concentrates on only two of the many features that have conventionally been used by transplant pathologists to assess acute rejection: tubulitis and intimal arteritis. Features judged to be less important or unproved, such as eosinophilic infiltrates or

Table 1 Histological features considered for incorporation into the Bayesian network

Histological feature	*Scoring system*
Tubulitis	0–3, as in Banff classification
Intimal arteritis	0–3, as in Banff classification
Interstitial haemorrhage	0–3: absent, 1–25%, 26–50%, more than 50% of area
Acute glomerulitis	0–3, as in Banff classification
Activated lymphocytes	0–3: absent; 1–25% of lymphocytes; 26–50% of lymphocytes; over 50% of lymphocytes
Venulitis	0–3: absent; scanty lymphocytes, few venules; abundant lymphocytes, <50% of venules; abundant lymphocytes, >50% of venules; abundant lymphocytes, all venules.
Tubular epithelial damage	0–3: (feature subsequently dropped as found to be too difficult to assess consistently)
Oedema	Percentage area of biopsy showing obvious interstitial oedema (viewed at low magnification)
Interstitial infiltrates	Percentage area of biopsy showing obvious interstitial infiltration by lymphocytes (viewed at low magnification)
Eosinophils	Number per single high power field in most heavily infiltrated area
Plasma cells	Number per single high power field in most heavily infiltrated area
Neutrophils	Number per single high power field in most heavily infiltrated area (feature subsequently dropped as neutrophils were found as commonly in both groups)
Arterial endothelial mononuclear cell adherence	Present or absent. (Cells adherent to luminal surface of endothelium; c/f intimal arteritis)
Venous endothelial mononuclear cell adherence	Present or absent. (Cells adherent to luminal surface of endothelium; c/f venulitis)

evidence of lymphocyte activation, are ignored. A structured system such as the Banff classification would become unwieldy if many minor features were included. We argued that if a computer, rather than a human brain, was used to relate a larger number of features in a systematic and reproducible way, the effort required would be reduced to an acceptable level, and reproducibility would be enhanced.

There has been recent interest in the literature in the use of Bayesian belief networks to address this type of problem.[10] A computer program is commercially available that has been designed to accept such "fuzzy" variables as a pathologist's impression of nuclear variation and architectural distortion, and to incorporate them in a systematic way to arrive at a diagnosis with a defined degree of confidence. The mathematical basis of these calculations has recently been described in this journal by Montironi *et al.*[10] We adapted this system to the evaluation of renal allograft biopsies, in an attempt to improve the accuracy of diagnosis of acute rejection in biopsies from "early" cases. Having developed the system, we tested it using our collection of "difficult" biopsies, which had all been seen previously by most renal transplant pathologists in the UK.[9]

Methods and materials

DEFINING THE CONDITIONAL PROBABILITY MATRIX

The computer program "Bayes for Win" was purchased from Diagsoft (Munich, Germany) and installed on an IBM PS2 microcomputer. The histological features that we considered initially are listed in table 1. For each of the features, a variety of possible grades must first be defined. This can be very flexible; a mixture of 0 to +++, percentage involvement, cell counts, etc, can be used for different variables. This program requires initialisation by the provision of a "conditional probability matrix" (CPM). To define the CPM for each possible result for each feature one must answer the question: "if this feature is present at this intensity, what is the probability that the specified outcome (acute rejection) is actually present?". The quality of these data obviously

Table 2 Conditional probability matrix for acute rejection. Figures for not acute rejection are given in parentheses

Feature							
	Absent	*1*	*2*	*3*			
Tubulitis	0.049 (046)	0.098 (0.41)	0.39 (0.1)	0.463 (0.03)			
Intimal arteritis	0.146 (0.862)	0.278 (0.086)	0.286 (0.034)	0.29 (0.017)			
Interstitial haemorrhage	0.159 (0.588)	0.254 (0.235)	0.286 (0.118)	0.301 (0.059)			
Acute glomerulitis	0.172 (0.455)	0.241 (0.273)	0.276 (0.182)	0.311 (0.091)			
Activated lymphocytes	0.04 (0.6)	0.28 (0.2)	0.32 (0.133)	0.36 (0.067)			
Venulitis	0.143 (0.5)	0.25 (0.25)	0.286 (0.167)	0.321 (0.083)			
	0%	*1–10%*	*11–20%*	*21–30%*	*31–50%*	*51–75%*	*76–100%*
Interstitial oedema	0.025 (0.3)	0.1 (0.2)	0.125 (0.167)	0.15 (0.133)	0.175 (0.1)	0.2 (0.067)	0.225 (0.033)
Interstitial infiltrates	0.024 (0.031)	0.1 (0.207)	0.13 (0.172)	0.17 (0.103)	0.191 (0.07)	0.192 (0.069)	0.193 (0.069)
	None	*1–5/HPF*	*6–10/HPF*	*11–20/HPF*	*21–30/HPF*	*31–40/HPF*	*> 40/HPF*
Eosinophils	0.038 (0.444)	0.135 (0.167)	0.154 (0.111)	0.163 (0.083)	0.164 (0.083)	0.173 (0.056)	0.173 (0.056)
Plasma cells	0.059 (0.368)	0.118 (0.211)	0.137 (0.158)	0.157 (0.105)	0.167 (0.079)	0.176 (0.053)	0.186 (0.026)
	Present	*Absent*					
Mononuclear adherence: venules	0.364 (0.667)	0.636 (0.333)					
Mononuclear adherence: arterioles	0.273 (0.778)	0.727 (0.222)					

HPF, high power field.

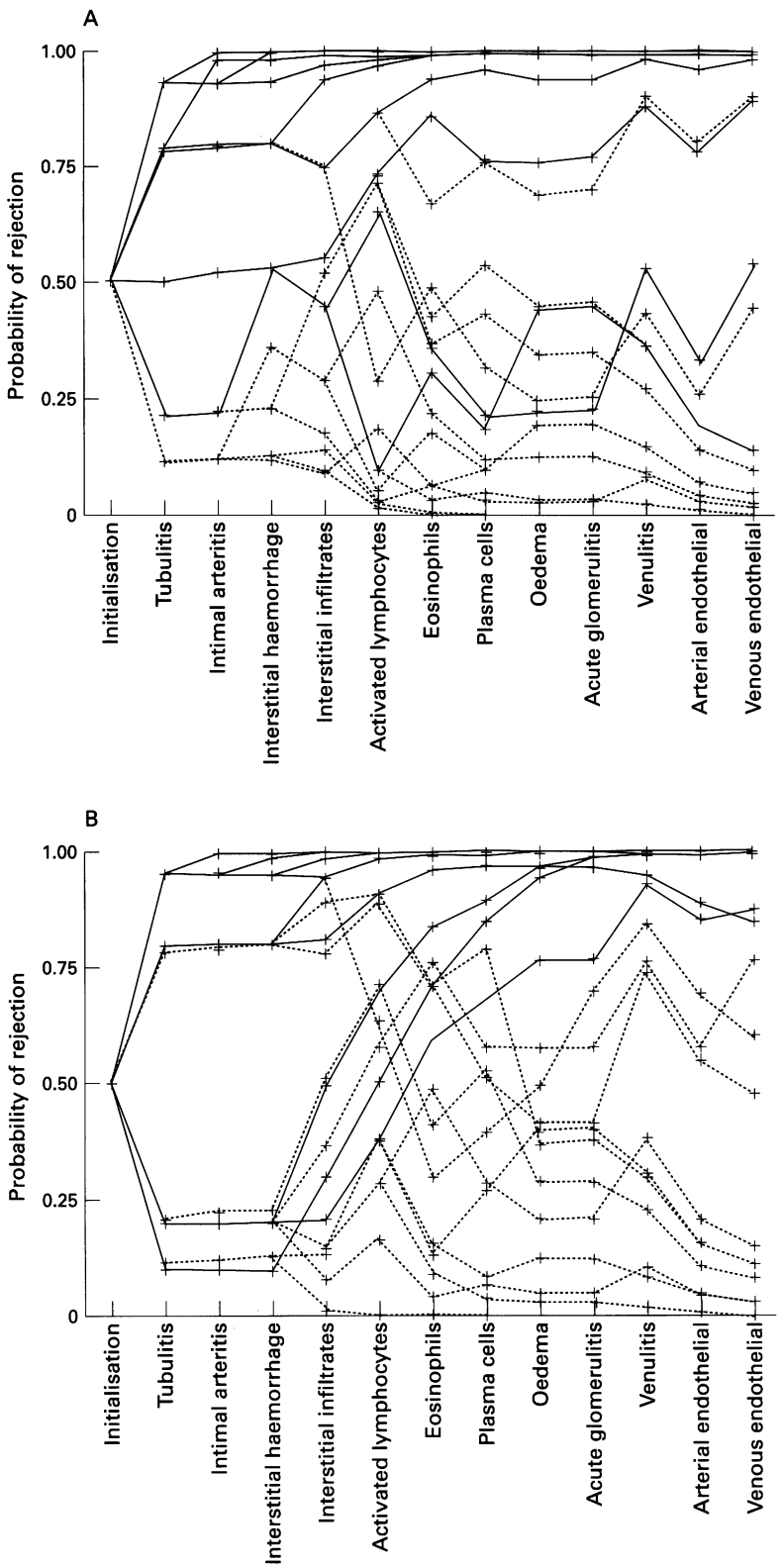

Figure 1 Progress of belief in a diagnosis of rejection (= 1) or not rejection (= 0) as information is entered into the Bayesian belief network for each of the 21 test biopsies. Solid lines are cases that, on retrospective study, represent acute rejection; dotted lines are not acute rejection. (A) Results achieved by the individual whose observations defined the conditional probability matrix. Note the even spread of results towards the top and bottom of the graph as data entry proceeds. (B) Data entry by a different individual; separation is less clear and the results are tending towards the top of the graph, illustrating the adverse effect of interobserver variation.

influences the quality of the output. It can successfully be provided merely as expert opinion,[11] but the results are likely to be better if they are based on actual observations.[12] Therefore, we took 110 transplant biopsies that all had reasonably confident diagnoses of acute rejection or non-rejection, based on a retrospective review of the casenotes by a senior clinical member of the transplant team (MN). These "training biopsies" were graded blindly by one observer (JIK) for each of the variables.

It was evident from this preliminary analysis that neutrophilic infiltration would not be contributory, as it was seen to a very similar intensity in rejection and non-rejection cases. An impression was gained that tubular epithelial damage could not be assessed consistently. This was confirmed by one observer who repeated the grading on separate days; very little correlation was found, so these two variables were dropped.

It was then necessary to adjust the probabilities slightly to allow for the possibility of rare events, which were not represented in our training series of 110 biopsies. Some of the figures in table 2, especially the boxes representing more severe changes, are derived from observation of as little as one event. A probability of zero or 100% at any point in the CPM would lead the network to give a diagnosis with a spurious absolute certainty and invalidate consideration of any other features. Therefore, the figures were manually "smoothed" to remove such extremes. The resultant figures for the CPM for all the variables are shown in table 2, as they were entered into the Bayes for Win program.

Table 3 Percentage of correct diagnoses by 37 UK renal transplant pathologists when considering the 21 test biopsies. As the Banff classification does not rigidly define which category should imply a diagnosis of acute rejection, three different cut off points are given

Method	*Correct**
"Conventional" unstructured diagnosis	64.8%
Banff category 3 or 4 ("borderline" or "acute rejection") considered to be a diagnosis of acute rejection	59.3%
Any grade of Banff category 4 ("acute rejection") considered to be a diagnosis of acute rejection	62.7%
Banff category 4 grade II or III only considered to be a diagnosis of rejection	63.2%

*50% is random.

TESTING THE SYSTEM

To assess the performance of this system in the diagnosis of acute rejection, we took 21 carefully selected biopsies that had initially caused one of us (PNF) some difficulty in the diagnosis or exclusion of acute rejection, but that all had carefully validated diagnoses from the subsequent clinical course. These biopsies had all previously been circulated to most renal pathologists in the UK, who had been asked to decide on the presence or absence of acute rejection by a conventional approach, and to apply the Banff classification of renal allograft pathology.[9] The proportion of correct diagnoses obtained by each approach is given in table 3. The fact that these proportions are only slightly above that expected from random guessing (50%) attests to the difficulty of making a diagnosis with these biopsies.

Sections from these specimens were all graded blindly for each of the variables in table 1 by the same single observer who had made the observations that defined the CPM (JIK). This was achieved by assigning a single grade for

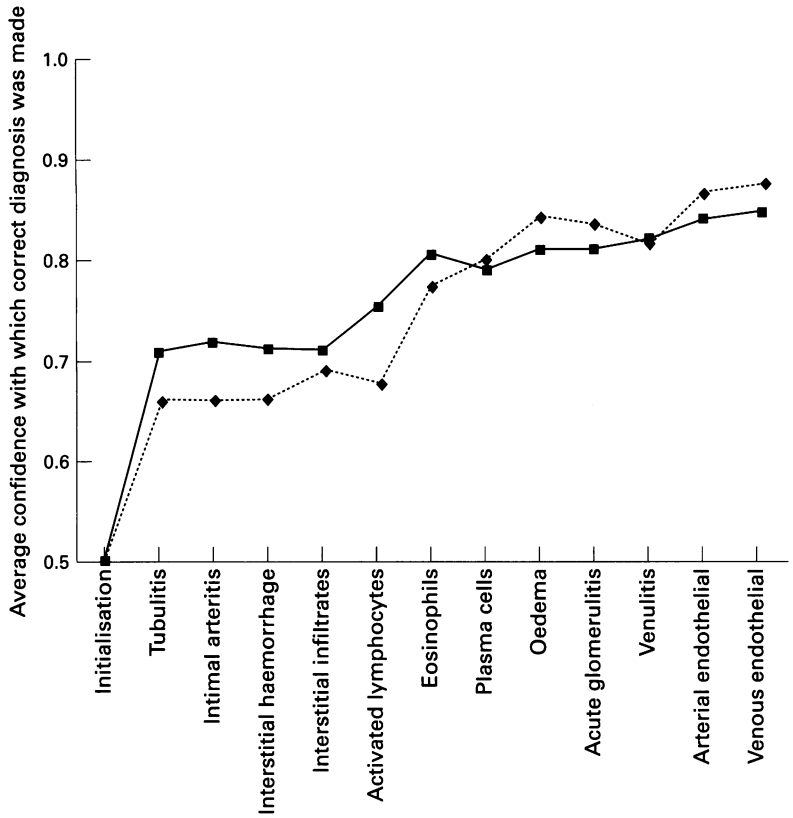

Figure 2 The value of belief in the correct diagnosis (as defined by retrospective clinical review), averaged over the 21 test cases, as data entry proceeds. The trend towards improved accuracy of diagnosis is shown by the upward slope of the lines. Solid line, original observer; broken line, second observer, illustrating interobserver variation. The feature that provided most difficulty in consistent evaluation was the proportion of activated lymphocytes. This feature was useful to the first observer, but interobserver error made it reduce the accuracy of diagnosis for the second observer.

each feature to each biopsy, although the program does have the facility to accept "fuzzy logic"—for example, a single biopsy could be recorded as grade 1 tubulitis 40%, grade 2 tubulitis 60%, if the observer is not convinced that the features completely justify classification as grade 2.

The results were entered into the Bayes for Win program and the level of belief of a diagnosis of acute rejection that the program generated was recorded at each step. To test the influence of interobserver variation on the system, the same 21 "difficult" biopsies were graded independently by a second observer (PNF), and entered into the program in the same way, without altering the CPM that had been derived from the first observer's findings.

The test group and the training group both had approximately equal numbers of biopsies representing rejection and non-rejection, so the a priori probability of a case representing rejection was set at 0.5. The cut off for concluding whether a diagnosis of rejection had been made after the data had been analysed was similarly defined as a probability of 0.5.

Results

The level of certainty of the diagnosis for each biopsy (the "belief") at each stage of data entry is shown in fig 1. The level of belief with which the correct diagnosis was made at each stage of data entry was then calculated. The mean value of this figure is shown in fig 2. Each figure shows the results of the first observer, using his own CPM (obviously, with no interobserver error), and the results of the second observer using the same CPM, therefore illustrating the effects of interobserver variation. The confidence with which the correct diagnosis was made for each case by each observer is shown in table 4. Direct comparisons are not valid, but as a guide to the difficulty of each case, the percentage of correct diagnoses received from the UK's transplant pathologists is also provided.

The Banff classification considers only acute tubulitis and intimal arteritis in the diagnosis of acute rejection. When only tubulitis had been considered by the network, the first observer produced 15 out of 21 correct diagnoses with an average belief in the correct diagnosis of 0.71; the second observer had only 14 of 21 correct, with an average belief of 0.66. This is comparable to the average proportion of correct diagnoses produced by the UK's renal transplant pathologists in the previous study.[9] In that study, the highest number of correct diagnoses made by any single pathologist was

Table 4 Level of belief with which the network provided the correct diagnosis (as defined by retrospective review), and proportion of correct diagnoses given by pathologists for each case

	Confidence of correct diagnosis using network		*% of pathologists correct in UK Banff study*[9]	
Case (clinical diagnosis)	*Observations by JIK*	*Observations by PNF*	*Conventional diagnosis*	*Banff 4 = rejection*
1 (Not rejection)	0.9	0.92	41*	46*
2 (Not rejection)	0.999	0.89	56	75
3 (Not rejection)	0.97	0.97	56	54
4 (Not rejection)	0.99	0.99	95	96
5 (Rejection)	0.99	0.99	82	82
6 (Rejection)	0.98	0.99	33*	18*
7 (Rejection)	0.54	0.84	58	29*
8 (Rejection)	0.14*	0.99	26*	19*
9 (Not rejection)	0.98	0.55	86	89
10 (Rejection)	0.99	0.99	100	100
11 (Rejection)	0.99	0.99	86	59
12 (Not rejection)	0.85	0.96	71	89
13 (Rejection)	0.89	0.87	50	19*
14 (Not rejection)	0.1*	0.39*	20*	29*
15 (Rejection)	0.99	0.99	83	50
16 (Not rejection)	0.95	0.99	56	62
17 (Rejection)	0.99	0.99	89	83
18 (Not rejection)	0.95	0.84	72	91
19 (Not rejection)	0.56	0.23*	50	71
20 (Not rejection)	0.99	0.88	50	71
21 (Rejection)	0.99	0.99	94	92

*Belief < 0.5 or < 50%, therefore regarded as an incorrect diagnosis.

17 of 21. In the present study, intimal arteritis was seen very rarely in the training set and not at all in the test set, so its contribution to the diagnosis was minimal. However, when all the other features had been entered into the network, the number of correct diagnoses improved to 19 of 21 for both observers (average belief of correct diagnosis 0.847 and 0.875).

At the end of the exercise, both observers had only two wrong diagnoses. For the first observer, these were one case of rejection and one case of non-rejection. When a second observer used the CPM defined by the first (fig 1B), there was a tendency towards overdiagnosis of rejection; the second observer overdiagnosed acute rejection in two cases.

Discussion

The rapid and accurate diagnosis or exclusion of episodes of acute rejection is fundamental to good management of renal allograft patients, and in most centres such diagnoses are heavily dependent on the results of transplant biopsies. However, our previous study[9] demonstrated that if a biopsy is taken early in the rejection process, there can be considerable variation in the pathologist's opinion of the presence or absence of acute rejection. This has obvious and important therapeutic implications.

Figure 2 illustrates how the different variables that we studied contribute to the certainty with which the diagnosis is made. The upward slopes of the lines beyond the point at which tubulitis and intimal arteritis has been entered demonstrates that the inclusion of "minor" features of acute rejection, such as interstitial oedema and eosinophilic infiltration, can improve the accuracy of the diagnosis in these early cases beyond the level which is possible using the Banff classification alone.

With such a large number of variables, all of different significance, it is unrealistic to expect the human brain to correlate the data in a consistent manner. Indeed, to attempt to do so would be to return to the traditional approach to transplant biopsy reporting, where the pathologist evaluated the features listed in an informal manner then synthesised a diagnosis on the basis of a "feeling" for the case based on years of experience. We suggest that the present data show that it is possible to have the best of both approaches. The subtlety of considering many variables can be combined with the systematic reproducibility of the Banff classification by using a computer to carry out the analysis in a predefined manner.

Both figures illustrate the effects of interobserver variation. Figure 2 indicates part of the reason; the first observer found that including observation of activated lymphocytes improved the quality of diagnosis, but the second observer did not. Review of the raw data showed that the second observer consistently overestimated the numbers of large lymphocytes compared with the first observer. The observation that some variables, such as numbers of eosinophils, performed better for the second observer is harder to explain, particularly as one would expect the maximum number of eosinophils to be subject to comparatively little interobserver variation.

Thus, interobserver variation has a significant effect in this system. We have not assessed interinstitution variation. The CPM was derived using cases from one institution, and it would be naive to expect it to function as well in another institution, where typical biopsy appearances might be modified by many feature, such as differences in immunosuppressive regimen and biopsy policy. Most of these problems might be overcome by using a more sophisticated neural network, in which the pathologist enters the "correct" diagnosis for each case whenever this becomes obvious from the subsequent clinical course, and the network then "learns" by adjusting its internal architecture. Such a system would, over time, be trained to the characteristics of the institute and the pathologist supplying the data, and would therefore be independent of interobserver variation. If the pathologist's criteria for grading a particular feature changed over time, or if the immunosuppressive regimen changed, the system would adapt to it. Furthermore, examination of such an evolving network could provide a dynamic measurement of which features were proving to be of most value in the assessment of the biopsies.

This system appears to be a promising method for the diagnosis of difficult cases of early acute rejection but, of course, the approach has limitations. Unlike the Banff classification, it does not classify the severity of rejection, nor does it provide the associated prognostic data. This could perhaps be overcome by having the output as belief in each of a number of grades of rejection, but obtaining a gold standard with which to grade the training cases would then be much more difficult. We have not tested the network in routine practice, nor have we assessed interobserver variation adequately; although this should be irrelevant with a more sophisticated network that learns as cases are added.

So far, we have not considered the input of clinical data, but it would not be difficult to add further parameters reflecting the degree of clinical suspicion of acute rejection, elevation of serum creatinine, time after transplantation, serum concentrations of immunosuppressive drugs, increase in graft size, etc. The result would be a diagnosis of the presence or absence of acute rejection for each biopsy that carried with it a comparatively objective measurement of the confidence with which the diagnosis had been made. It might prove possible to extend this approach to the diagnosis of other causes of early graft dysfunction; a comparable network could perhaps facilitate the further development of a chronic allograft damage index.[13]

With appropriate further development, we suggest that the potential benefit for patient care provided by this systematic approach to renal transplant histology is self evident.

We are grateful to all the renal transplant pathologists in the UK for expressing opinions on the sections that we used to test this system, and to Dr Kim Solez and his staff for independent review of the biopsies used in this study and for constructive discussion of this project.

1 Rush DN, Henry SF, Jeffery JR, *et al.* Histological findings in early routine biopsies of stable renal allograft recipients. *Transplantation* 1994;**57**:208–11.
2 Rush DN, Jeffery JR, Gough J. Sequential protocol biopsies in renal transplant patients. Clinico-pathological correlations using the Banff schema. *Transplantation* 1995;**59**:511–4.
3 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* 1993;**44**:411–22.
4 Dooper MM, Hoitsma AJ, Koene RA, *et al.* Evaluation of the Banff criteria for the histological diagnosis of rejection in renal allograft biopsies. *Transplant Proc* 1995;**27**:1005–6.
5 Gaber L, Schroeder T, Moore L, *et al.* The correlation of Banff scoring with reversibillity of first and recurrent rejection episodes. *Transplantation* 1996;**61**:1711–15.
6 Laine J, Krogerus L, Jalanko H, *et al.* Renal allograft histology and correlation with function in children on triple therapy. *Nephrol Dial Transplant* 1995;**10**:95–102.
7 Marcussen N, Olsen TS, Benediktsson H, *et al.* Reproducibility of the Banff classification of renal allograft pathology: inter- and intra-observer variation. *Transplantation* 1995;**60**:1083–9.
8 Solez K, Hansen HE, Kornerup HJ, *et al.* Clinical validation and reproducibility of the Banff schema for renal allograft pathology. *Transplant Proc* 1995;**27**:1009–11.
9 Furness P, Kirkpatrick U, Taub N, *et al.* A UK-wide trial of the Banff classification of renal transplant pathology in routine diagnostic practice. *Nephrol Dial Transplant* 1997; **12**:995–1000.
10 Montironi R, Whimster WF, Collan Y, *et al.* How to develop and use a Bayesian belief network. *J Clin Pathol* 1996;**49**:194–201.
11 Montironi R, Bartels PH, Thompson D, *et al.* Prostatic intraepithelial neoplasia (PIN). Performance of a Bayesian belief network for diagnosis and grading. *J Pathol* 1995;**177**:153–62.
12 Durrant N. Bayesian statistics is valuable provided it is based on data [letter]. *BMJ* 1997;**314**:74.
13 Isoniemi H, Taskinen E, Hayry P. Histological chronic allograft damage index accurately predicts chronic renal allograft rejection. *Transplantation* 1994;**58**:1195–8.