Transplant International

ORIGINAL ARTICLE

A method to reduce variability in scoring antibody-mediated rejection in renal allografts: implications for clinical trials – a retrospective study

Byron Smith[1], Lynn D. Cornell[2], Maxwell Smith[3], Cherise Cortese[4], Xochiquetzal Geiger[4], Mariam P. Alexander[2], Margaret Ryan[3], Walter Park[5], Martha Catalina Morales Alvarez[5], Carrie Schinstock[5,6], Walter Kremers[1,5] & Mark Stegall[5,7]

1 Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
2 Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA
3 Department of Laboratory Medicine and Pathology, Mayo Clinic, Scottsdale, AZ, USA
4 Department of Laboratory Medicine and Pathology, Mayo Clinic, Jacksonville, FL, USA
5 The William J von Liebig Center for Transplantation and Clinical Regeneration, Mayo Clinic, Rochester, MN, USA
6 Department of Internal Medicine, Division of Nephrology and Hypertension, Mayo Clinic, Rochester, MN, USA
7 Department of Medicine, Division of Transplantation Surgery, Mayo Clinic, Rochester, MN, USA

Correspondence
Mark Stegall, MD, Mayo Clinic,
Rochester, MN 55905, USA
Tel.: 507-266-2812
fax: 507-266-2810
e-mail: Stegall.mark@mayo.edu

SUMMARY

Poor reproducibility in scoring antibody-mediated rejection (ABMR) using the Banff criteria might limit the use of histology in clinical trials. We evaluated the reproducibility of Banff scoring of 67 biopsies by six renal pathologists at three institutions. Agreement by any two pathologists was poor: 44.8–65.7% for glomerulitis, 44.8–67.2% for peritubular capillaritis, and 53.7–80.6% for chronic glomerulopathy (cg). All pathologists agreed on cg0 (n = 20) and cg3 (n = 9) cases, however, many disagreed on scores of cg1 or cg2. The range for the incidence of composite diagnoses by individual pathologists was: 16.4–22.4% for no ABMR; 17.9–47.8% for active ABMR; and 35.8–59.7% for chronic, active antibody-mediated rejection (cABMR). A "majority rules" approach was then tested in which the scores of three pathologists were used to reach an agreement. This increased consensus both for individual scores (ex. 67.2–77.6% for cg) and for composite diagnoses (ex. 74.6–86.6% cABMR). Modeling using these results showed that differences in individual scoring could affect the outcome assessment in a mock study of cABMR. We conclude that the Banff schema has high variability and a majority rules approach could be used to adjudicate differences between pathologists and reduce variability in scoring in clinical trials.

Transplant International 2019; 32: 173–183

Key words
rejection, kidney clinical

Received: 23 July 2018; Revision requested: 14 August 2018; Accepted: 30 August 2018;
Published online: 2 October 2018

Introduction

Improving renal allograft survival will almost certainly require the development of novel therapeutic agents that prevent chronic injury. To demonstrate the efficacy of these therapies, new surrogate endpoints for clinical trials are needed [1]. Demonstrating a differential impact on renal allograft histology could be a very effective means of showing the efficacy of a drug. However, the poor reproducibility of the Banff scoring system among pathologists is a major drawback to its use as an end point in clinical trials [2,3].

The Banff 2017 Kidney Meeting Report recognized this problem and made recommendations on best

doi:10.1111/tri.13340

practice for pathology endpoints in clinical trials including the development of an adjudication mechanism to address discordance between pathologists [4]. The report also presented revised criteria for the diagnosis of both acute, active antibody-mediated rejection (aABMR) and chronic, active antibody-mediated rejection (cABMR) and removed the terminology of “acute, active ABMR”. However, the diagnoses of aABMR and cABMR still rely heavily on the ability of a pathologist to identify and grade glomerulitis (g), peritubular capillaritis (ptc), and chronic glomerulopathy (cg). Thus, reproducibility remains an issue.

The goal of the current study was twofold. First, we aimed to compare the reproducibility of the main individual histologic scores (ptc, g, and cg) and the composite of these scores used for the aABMR and cABMR diagnoses based on the Banff classification. Second, we sought to test a process by which scores from multiple pathologists could be combined to yield a more consistent score, i.e., to provide an adjudication mechanism to address discordance between pathologists as directed by the Banff report. Our hypothesis for these studies was that the scores between the individual pathologists will not show close agreement, but that using a “majority rules” diagnosis from a group of three pathologists will improve the consistency of diagnoses and thus might provide a superior methodology to score biopsies in future clinical trials.

Materials and methods

This study was approved by the Mayo Clinic Institutional Review Board (IRB 14-006319). The transplant database at Mayo Clinic in Rochester, MN was searched by the initial reviewing pathologist (LDC) to identify a subset of biopsies to adequately represent the different grades of peritubular capillaritis, glomerulitis, and cg (Banff ptc, g, and cg scores 0–3) for a total of 12 graded categories. A total of 67 biopsies obtained from years 2005–2013 were identified with at least six biopsies in each graded category. Forty-eight biopsies had an original diagnosis of antibody-mediated rejection. Of those, two were mixed (one grade 1B and one grade 2A), two had concomitant IgA nephropathy, and one had BK nephropathy.

Slide selection

Each biopsy case contained 10 slides with 3–4 histologic sections per slide, cut at 3–4 microns per section. These were stained for hematoxylin and eosin (H&E), periodic acid-Schiff (PAS), Masson trichrome, and Jones-methenamine silver, as per routine histologic preparation for clinical cases. From these 10 slides, three slides from each case (one each H&E, PAS, and silver stained) with the most representative tissue were chosen by the initial reviewing pathologist (LDC) and coded with a case number. These coded slides were shipped to each participating renal pathologist for interpretation**.** Biopsies were all scored without knowledge of preformed donor-specific antibody (DSA) or C4d staining status, indication for biopsy, time post-transplant, or other clinical or laboratory parameters. In addition to the test cases, example cases with each of the graded categories as assessed by the initial pathologist (LDC) were reviewed prior to the test cases. Each pathologist was given written guidelines for scoring. At Mayo Clinic, C4d is performed by immunofluorescence on fresh-frozen tissue and therefore could not be repeated for this study (no frozen tissue existed). Therefore, the original C4d reading was used for this study.

Active and chronic ABMR diagnosis

In addition to assigning individual Banff scores, a diagnosis of no rejection, aABMR, or cABMR was given based on the aggregate of scores, purely based on strict Banff 2017 criteria [4]. Note that the diagnosis was not made by the pathologist, but only derived from the Banff scores. For these rereads, pathologists were blinded to the C4d and DSA results and thus were not biased by these data. However, since one of our goals was to mimic central pathology rereads in a clinical trial, the lack of central reread for C4d actually mimics what happens in most trials. Thus we determined agreement on ABMR using the original C4d scoring done on 67 biopsies. Of these 50 (74.6%) were negative and 17 (25.4%) were positive. Active ABMR was assigned using the Banff 2017 criteria [4] as: either (i) C4d ≤ 1 and ptc + g score ≥ 2 or (ii) C4d > 1 and ptc > 0 or g > 0. A concomitant cg score >0 resulted in the cABMR diagnosis. Electron microscopy and genomic testing was not routinely done and not used as criteria for aABMR or cABMR. The reproducibility of acute tubular injury was not assessed and therefore not considered when assigning the diagnosis.

Majority rules approach

For the majority rules approach we used a consensus score based on various combinations of three pathologist’s scores from a pool of six pathologists. Specifically, the six pathologists' scores were combined in different ways resulting in 20 distinct groups (1 + 2 + 3, 1 + 3 + 4, etc.) of three pathologists. We then compared variability among the 20 groups by considering exclusive groups (10 total comparisons) to overestimation of agreement due to pathologist overlap. For individual Banff scores, the majority rules score was given by rounding the average Banff scores of the three pathologists that comprised the group. For the diagnosis, the majority rules diagnosis is simply a majority rules vote (equivalent to taking the rounded average of binary variables).

Mock trial simulation

We examined how a majority rules approach using three pathologists increases the probability that an assigned diagnosis actually approaches the "true diagnosis" and the correct patients would be enrolled in the trial. We assumed that as we increase the number of pathologists reading a biopsy, at some point a true consensus would emerge. First we considered the diagnosis of a consensus of six pathologists (4/6) as the "true diagnosis." We then evaluated the frequency of a single pathologist versus three pathologists (using majority rules) reaching the "true diagnosis." We used the diagnosis of cABMR versus no cABMR with the same reproducibility and frequency as described above. In order to demonstrate the impact of a lack of reliability, the mock trial process was repeated over a grid of effect sizes as measured by the odds ratio from 0.15 to 0.5 by increments of 0.05.

Statistical analysis

Accuracy was assessed through the pairwise percent agreement. Reliability was assessed through Cohen's Kappa [5]. All comparisons carried out between groups of three pathologists only used exclusive groups of pathologists, (i.e., 1 + 2 + 3 vs. 4 + 5+6, but not 1 + 2 + 3 vs. 1 + 5 + 6). Kappa statistics were calculated using the "psych" package and heatmaps were generated using the "gplots" in R (Vienna, Austria).

In order to form a diagnosis tree, we hypothesized a 150 patient sample (to directly convert from percentages). Of this 150 patient sample, the "true" diagnosis rate of cABMR versus no cABMR was established by using the consensus of all six pathologists. The reclassification using a three pathologist combination or a single pathologist can then be calculated and directly converted to sensitivity, specificity, positive predictive value, and negative predictive value.

Subsequent results from a hypothetical clinical trial displayed as contingency tables with enrollment and end point classification variability were constructed with conditional probabilities of diagnosis based on latent classes [6]. Equivalently, the disagreement between any two pathologists (or two sets of three pathologists) means that the counts in a table could change – a patient may be recorded as no cABMR by one pathologist and cABMR by another pathologist. How often this occurs depends on the agreement between the pathologists. For example, consider a trial with 90% agreement and 75 patients on a treatment arm that resulted in 70 ABMR cases and five cases with no rejection. Then, a rescoring which results in 70*0.1 = 7 patients switching from the "ABMR" category to the "no rejection" category would be contained within the variability of that diagnosis. Similarly, 5*0.1 = 0.5 implies that one patient switching from the "no rejection" category to the "ABMR" category would also be contained within the variability.

Results

Patient characteristics

Sixty-seven biopsies from 44 patients were studied (Table 1). 83.6% (56/67) of the biopsies were done for surveillance purposes at a median (IQR) time of 1.0 (0.6–2.1) years post-transplant. Of the patients biopsied, the median (IQR) age at the time of transplantation was 47 (35–54) years, and the majority was female 68.2% (30/44), Caucasian 90.9% (40/44), and received their kidney from a living donor 95.5% (42/44). The leading cause of end stage renal disease (ESRD) in this group was glomerulonephritis [43.2% (19/44)]. The majority of patients [81.8% (36/44)] had a positive B-flow cytometric crossmatch at the time of transplantation. The majority of patients were on tacrolimus, mycophenolate mofetil, and prednisone for maintenance immunosuppression [95.4% (42/44)]. The majority of the allografts were still functioning at 2 and 5 years post-transplant (97.7% and 81.8%, respectively).

Reproducibility: scoring agreement among individual pathologists

Agreement on individual Banff scores

We first examined the specific agreement for individual g, ptc, and cg scores by performing pairwise

Table 1. Patient characteristics.

	All patients (*N* = 44)
Number of biopsies reviewed/patient *n* (%)	
1	28 (63.6)
2	11 (25.0)
3	3 (6.8)
4	2 (4.5)
Age at transplantation years median (IQR)	47 (35–54)
Years from transplantation to biopsy median (IQR)	1.0 (0.6–2.1)
Female gender *n* (%)	30 (68.2)
Race *n* (%)	
Caucasian	40 (90.9)
Hispanic	2 (4.5)
African American	1 (2.3)
Pacific Islander	1 (2.3)
Donor Type *n* (%)	
Living unrelated donor	25 (56.8)
Living related donor	17 (38.6)
Deceased donor	2 (4.5)
Etiology of end stage renal disease *n* (%)	
Glomerulonephritis	19 (43.2)
Cystic Kidney Disease	5 (11.4)
Diabetes Mellitus	5 (11.4)
Hypertension	4 (9.1)
Congenital Renal Disease	2 (4.5)
Other	7 (15.9)
Unknown	2 (4.5)
Positive B-flow crossmatch *n* (%)	36 (81.8)
Maintenance immunosuppression *n* (%)	
Tacrolimus, Mycophenolate mofetil, Prednisone	42 (95.4)
Other	2 (4.5)
Allograft status 2 years post-transplant *n* (%)	
Active	43 (97.7)
Failure	1 (2.3)
Allograft status 5 years post-transplant *n* (%)	
Active	36 (81.8)
Failure	8 (18.2)

comparisons among individual pathologists for each biopsy (ex. pathologist #1 vs. #2, #1 vs. #3, etc.) resulting in six pairwise comparisons. The specific agreement on scoring was relatively low. For the g score (glomerulitis), the exact agreement on the score among any two pathologists ranged from 44.8 to 65.7%; for ptc score (peritubular capillaritis), the range was 44.8 to 67.2%; and for cg the range was 53.7 to 80.6%. However, the vast majority of the g, ptc, or cg scores either matched or were ± 1 between pathologists (Table 2, Fig. 1). For example, the number of cg scores with differences ≥2 ranged from 0 to 7 (0 to 10.5%). Thus, there was general agreement, but not exact agreement g, ptc, and cg scoring. The average kappa statistics were 0.39, 0.38, and 0.48 for g, ptc, and cg scores, respectively.

Despite many discrepancies there was complete agreement regarding cg score among the six pathologists in 29 cases – 20 scored a cg0 and nine scored a cg3. In contrast, there was never complete agreement regarding scores of cg1 or cg2 suggesting that assessing intermediate levels of cg had higher variability.

The effect of pathologists working at the same site was also investigated for possible confounding. By separating comparisons between pathologists across sites and within sites, we find that the percent agreement and kappa statistic within site is contained within the range of those across sites for all individual scores (data not shown). Therefore, pathologists working closely at the same site did not bias the mean and range of reported statistics in this study.

Agreement on the diagnosis of ABMR versus no ABMR

The agreement between single pathologists on a diagnosis of ABMR (either aABMR or cABMR) versus no rejection ranged from 86.6 to 97.0%. However, it was not always the same biopsies that led to disagreements in these pairwise comparisons. In the worst case scenario, 13.4% of biopsies would be classified differently just by being read by a different pathologist. In 77.6% of biopsies (52/67), all six pathologists agreed a diagnosis of either ABMR (aABMR or cABMR)/no ABMR including: Of the 52 with complete agreement, six were scored as no ABMR and 46 met critieria for ABMR (when combined with C4d data). In 15 biopsies (22%), at least one pathologist's composite score led to a different diagnosis (i.e., in 22% of cases, reading the same slide by a different pathologist would result in a different diagnosis).

Agreement on no ABMR, active ABMR, and chronic active ABMR

When the cg score is added to the ABMR composite score to determine presence or absence of cABMR, there are three possible agreement/disagreement groups: cABMR, aABMR, and no ABMR. More choices resulted in less agreement. There was complete agreement among all six pathologists about the diagnoses (cABMR, aABMR and no ABMR) in only 43.3% of biopsies (29/67). All six pathologists agreed upon six cases of no ABMR, six cases of aABMR, and 17 cases of cABMR. However, in 56.7% (38/67) at least one pathologist's scores lead to a different composite diagnosis. The incidence of the three diagnoses as designated by individual pathologists ranged from 16.4

Table 2. Diagnostic agreement among pathologists based on individual Banff scores. The cell values correspond to the count and percentage of the 67 total cases with no difference (zero), one, two, or three in Banff scores given by the two pathologists.

			Magnitude of difference between first and second pathologist			
	First pathologist	Second pathologist	Zero (%)	One (%)	Two (%)	Three (%)
Glomerulitis	Path 1	Path 2	39 (58.21)	22 (32.84)	6 (8.96)	0 (0)
	Path 1	Path 3	30 (44.78)	33 (49.25)	4 (5.97)	0 (0)
	Path 1	Path 4	39 (58.21)	27 (40.3)	1 (1.49)	0 (0)
	Path 1	Path 5	41 (61.19)	22 (32.84)	3 (4.48)	1 (1.49)
	Path 1	Path 6	32 (47.76)	27 (40.3)	8 (11.94)	0 (0)
	Path 2	Path 3	33 (49.25)	29 (43.28)	5 (7.46)	0 (0)
	Path 2	Path 4	41 (61.19)	23 (34.33)	3 (4.48)	0 (0)
	Path 2	Path 5	44 (65.67)	17 (25.37)	6 (8.96)	0 (0)
	Path 2	Path 6	36 (53.73)	18 (26.87)	12 (17.91)	1 (1.49)
	Path 3	Path 4	37 (55.22)	26 (38.81)	4 (5.97)	0 (0)
	Path 3	Path 5	32 (47.76)	27 (40.3)	7 (10.45)	1 (1.49)
	Path 3	Path 6	34 (50.75)	22 (32.84)	9 (13.43)	2 (2.99)
	Path 4	Path 5	44 (65.67)	20 (29.85)	3 (4.48)	0 (0)
	Path 4	Path 6	38 (56.72)	26 (38.81)	3 (4.48)	0 (0)
	Path 5	Path 6	35 (52.24)	20 (29.85)	12 (17.91)	0 (0)
Chronic glomerulopathy	Path 1	Path 2	42 (62.69)	21 (31.34)	3 (4.48)	1 (1.49)
	Path 1	Path 3	54 (80.6)	12 (17.91)	1 (1.49)	0 (0)
	Path 1	Path 4	52 (77.61)	12 (17.91)	3 (4.48)	0 (0)
	Path 1	Path 5	37 (55.22)	22 (32.84)	6 (8.96)	1 (1.49)
	Path 1	Path 6	44 (65.67)	14 (20.9)	6 (8.96)	3 (4.48)
	Path 2	Path 3	46 (68.66)	17 (25.37)	3 (4.48)	1 (1.49)
	Path 2	Path 4	49 (73.13)	13 (19.4)	4 (5.97)	1 (1.49)
	Path 2	Path 5	39 (58.21)	19 (28.36)	5 (7.46)	3 (4.48)
	Path 2	Path 6	46 (68.66)	10 (14.93)	8 (11.94)	3 (4.48)
	Path 3	Path 4	53 (79.1)	14 (20.9)	0 (0)	0 (0)
	Path 3	Path 5	35 (52.24)	22 (32.84)	8 (11.94)	1 (1.49)
	Path 3	Path 6	46 (68.66)	13 (19.4)	6 (8.96)	2 (2.99)
	Path 4	Path 5	39 (58.21)	19 (28.36)	7 (10.45)	1 (1.49)
	Path 4	Path 6	48 (71.64)	11 (16.42)	6 (8.96)	2 (2.99)
	Path 5	Path 6	46 (68.66)	16 (23.88)	1 (1.49)	3 (4.48)
Peritubular capillaritis	Path 1	Path 2	34 (50.75)	27 (40.3)	5 (7.46)	1 (1.49)
	Path 1	Path 3	37 (55.22)	28 (41.79)	2 (2.99)	0 (0)
	Path 1	Path 4	43 (64.18)	21 (31.34)	3 (4.48)	0 (0)
	Path 1	Path 5	30 (44.78)	31 (46.27)	5 (7.46)	1 (1.49)
	Path 1	Path 6	36 (53.73)	25 (37.31)	6 (8.96)	0 (0)
	Path 2	Path 3	43 (64.18)	19 (28.36)	4 (5.97)	1 (1.49)
	Path 2	Path 4	36 (53.73)	29 (43.28)	2 (2.99)	0 (0)
	Path 2	Path 5	34 (50.75)	26 (38.81)	7 (10.45)	0 (0)
	Path 2	Path 6	35 (52.24)	27 (40.3)	5 (7.46)	0 (0)
	Path 3	Path 4	31 (46.27)	33 (49.25)	3 (4.48)	0 (0)
	Path 3	Path 5	34 (50.75)	27 (40.3)	5 (7.46)	1 (1.49)
	Path 3	Path 6	38 (56.72)	25 (37.31)	4 (5.97)	0 (0)
	Path 4	Path 5	34 (50.75)	30 (44.78)	2 (2.99)	1 (1.49)
	Path 4	Path 6	45 (67.16)	20 (29.85)	2 (2.99)	0 (0)
	Path 5	Path 6	38 (56.72)	29 (43.28)	0 (0)	0 (0)

to 22.4% for no ABMR; 17.9 to 47.8% for aABMR; and 35.8 to 59.7% for cAMBR. (Table 3). The kappa statistic for cABMR versus no cABMR (no ABMR or aABMR) ranged from 0.43 to 0.74.

Majority rules approach

Table 3 shows the ranges of individual scores and diagnoses among individual pathologists and between the

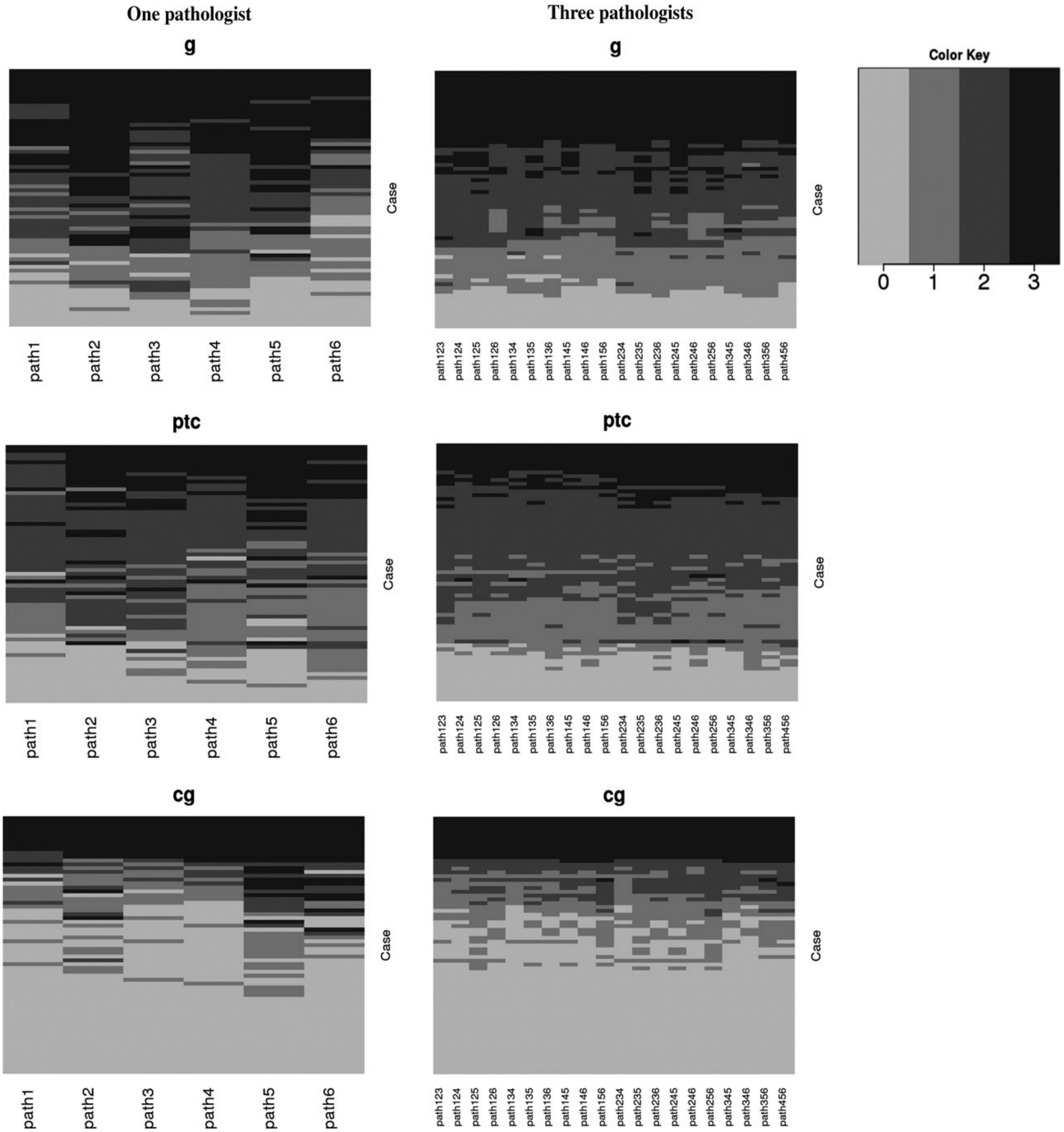


Figure 1 Heatmaps for Banff scores using one pathologist versus a 3 pathologist "consensus" score. The darker the gray, the higher the case was scored by that pathologist.

various majority rules groups of three pathologists. Using the majority rules approach, the agreement for g, ptc, and cg increased markedly to 72.2%, 70.4%, and 72.2%, respectively (Table 4). The kappa statistic also increased to 0.62, 0.60, and 0.57, for g, ptc, and cg, respectively.

Table 4 shows the mean and range percent agreement and kappa statistics for Banff scores and diagnoses using a single pathologist or a "majority rules" approach. As expected, the kappa statistic improved for all categories using majority rules. For the diagnosis of ABMR versus No Rejection, agreement improved from a range of 86.6–97.0% to a range of 91.0–97.0%. The "majority rules" kappa ranged from 0.71 to 0.91.

Similar to the diagnosis of aABMR, a majority rules approach involving three pathologist's scores led to

Table 3. Range of diagnosis incidences using a single pathologist or a "majority rules" approach.

Diagnosis	One pathologist, %	Three pathologists, %
No rejection	16.4–22.4	16.4–23.9
aABMR	17.9–47.8	28.4–50.7
cABMR	35.8–59.7	32.8–49.3

aABMR, acute active antibody-mediated rejection; cABMR, chronic, active antibody-mediated rejection.

improved consistency in the various diagnoses. The various combinations of three pathologist's scores led to the same diagnosis (cABMR vs. aABMR or no rejection) in an average of 85.4% of cases and ranged from 80.6 to 89.6%. When considering the three-level diagnosis as no rejection versus aABMR versus cABMR, there was average agreement of 72.3% (range = 59.7% to 80.6%) using one pathologist and 80.0% (range = 74.6–86.6%) using a majority rules method. The incidence of the three diagnoses as designated by three pathologist majority rules ranged from 16.4 to 23.9% for no ABMR; 28.4 to 50.7% for aABMR; and 32.8 to 49.3% for cAMBR.

Implications of using a majority rules approach in clinical trial design: an example

Inclusion criteria

Table 5 and Fig. 2 show the mock trial and statistical variability of using one pathologist versus three in a majority rules approach. Calculations are based on the six pathologist consensus diagnosis. Note that two cases had a tie (three vs. three) and these cases were excluded from subsequent calculations. When only one pathologist determined a positive cABMR diagnosis, a mean of 25.8% (100% minus 74.2%) of cases were classified differently than the consensus of six pathologists (true diagnosis). In contrast, when the majority rules approach was applied using three pathologists; on average, only 17.1% of cases were classified differently than the "true diagnosis."

Clinical trial result

Clinical trial results can also be influenced by diagnostic reliability. To illustrate this, a mock clinical trial was designed with the following assumptions: (i) All patients enrolled had biopsy-proven cABMR [prevalence of the underlying disease was estimated using the average positive predictive value (number of true cABMR cases predicted correctly divided by the total predicted number of cABMR cases)], (ii) The rate of treatment being successful was 45% in the control group and 68% in the treatment group consistent with reported treatment efficacies [7,8], and (iii) the range of cABMR variability used corresponds to 79.5% for a single pathologist and 85.4% for a "majority rules" group based on known variability (Table 4). Using these assumptions we calculated the impact of diagnostic variability on the trial results (Fig. 3). In an ideal situation the reported odds ratio would be 0.390 ($P = 0.0049$), i.e., drug is effective in reversing ABMR Fig. 3a. However, the results from a second pathologist reading the same slides could be an odds ratio of 0.530 ($P = 0.0757$) Fig. 3b. If the majority rules approach (three pathologists) applied, the clinical trial result would more likely match the result as that from the ideal situation [odds ratio of 0.482 ($P = 0.0370$)] and deem the drug effective as before Fig. 3c. The impact of diagnostic variability can be

Table 4. Mean and range percent agreement and kappa statistics for Banff scores and diagnoses using a single pathologist or a "majority rules" approach.

	One pathologist		Three pathologist 'majority rules'	
Score	% Agreement [mean (range)]	Kappa [mean (range)]	Agreement [mean (range)]	Kappa [mean (range)]
Glomerulitis	55.2 (44.8, 65.7)	0.39 (0.24, 0.53)	72.2 (67.2, 79.1)	0.62 (0.55, 0.72)
Peritubular capillaritis	54.5 (44.8, 67.2)	0.38 (0.27, 0.54)	70.4 (62.7, 80.6)	0.60 (0.49, 0.73)
Chronic glomerulopathy	67.5 (53.7, 80.6)	0.48 (0.31, 0.65)	72.2 (67.2, 77.6)	0.57 (0.50, 0.64)
Acute, active antibody-mediated rejection (Y/N)	90.2 (86.6, 97.0)	0.70 (0.53, 0.91)	94.0 (91.0, 97.0)	0.82 (0.71, 0.91)
Chronic, active antibody-mediated rejection (Y/N)	79.5 (70.1, 88.1)	0.59 (0.43, 0.74)	85.4 (80.6, 89.6)	0.70 (0.60, 0.78)
Diagnosis	72.3 (59.7, 80.6)	0.57 (0.39, 0.70)	80.0 (74.6, 86.6)	0.69 (0.61, 0.79)

Table 5. Sensitivity, specificity, positive predictive value, and negative predictive value each single pathologist. Average scores are given for single pathologists and three pathologist "majority rules" for comparison as well.

	Sensitivity	Specificity	PPV	NPV
Three path average	100.00	89.4	82.9	100.0
One path average	94.8	83.0	74.2	96.9
Path 1	92.2	90.9	83.9	95.7
Path 2	96.1	76.8	68.1	97.4
Path 3	92.2	92.9	87.0	95.8
Path 4	92.2	92.9	87.0	95.8
Path 5	100.0	60.6	56.7	100.0
Path 6	96.1	83.8	75.4	97.6

generalized to other effect and sample sizes as shown in Fig. 4.

Discussion

We found that the reproducibility of the individual Banff scores needed to diagnose ABMR (glomerulitis, peritubular capillaritis, and cg) was poor when evaluating the histologic interpretation of 67 biopsies by six pathologists from three transplant centers. The diagnosis of ABMR based on a composite of individual scores, had reduced variability but remained inadequate considering that reliable and accurate histologic diagnosis is essential for the design of effective clinical trials, as illustrated in our mock clinical trial design. In only 43.3% of the cases did the composite of individual Banff scores lead to the same diagnosis for all pathologists. We devised a majority rules approach using three pathologists for ABMR diagnostic classification and found that it led to a more reliable and reproducible results than the use of a single pathologist. This alternative approach to histologic interpretation is consistent with the best practice recommendations for histologic diagnosis outlined in the Banff 2017 meeting report and would likely improve our ability to perform effective clinical trials in ABMR.

A potential limitation of this study was that most cases were purer forms of ABMR that lacked concurrent signs of cellular rejection. Thus, the reproducibility of the lesions related to ABMR in other types of patients (i.e., those more conventional patients with nonadherence) might be limited.

It is well known that histologic interpretation is complex and the reproducibility of most Banff lesions is low [9]. Many studies have looked at the acute Banff scores with kappa values ranging from 0.19 to 0.50 for glomerulitis [2,10–13]. Another study showed kappa values using two pathologists to score peritubular capillaritis were as low as 0.28–0.53 [14,15]. The interobserver variability of transplant glomerulopathy was found to 0.14 [2] although this was improved to values as high as 0.47 with later diagnosis criteria [16]. Similar poor interobserver variability has been observed with other Banff lesions including: interstitial fibrosis (linearly weighted kappa = 0.04–0.15), peritubular capillary basement membrane multilaminations (kappa = 0.66–

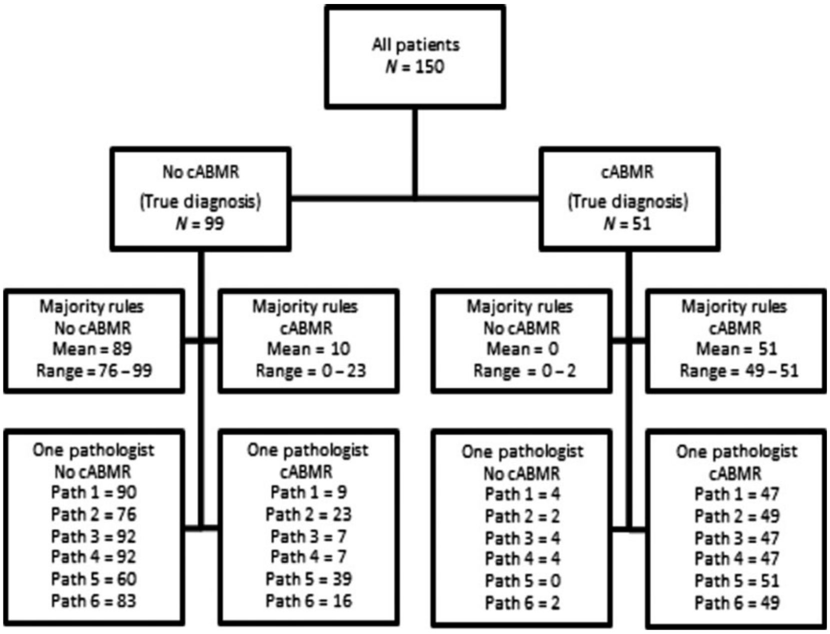
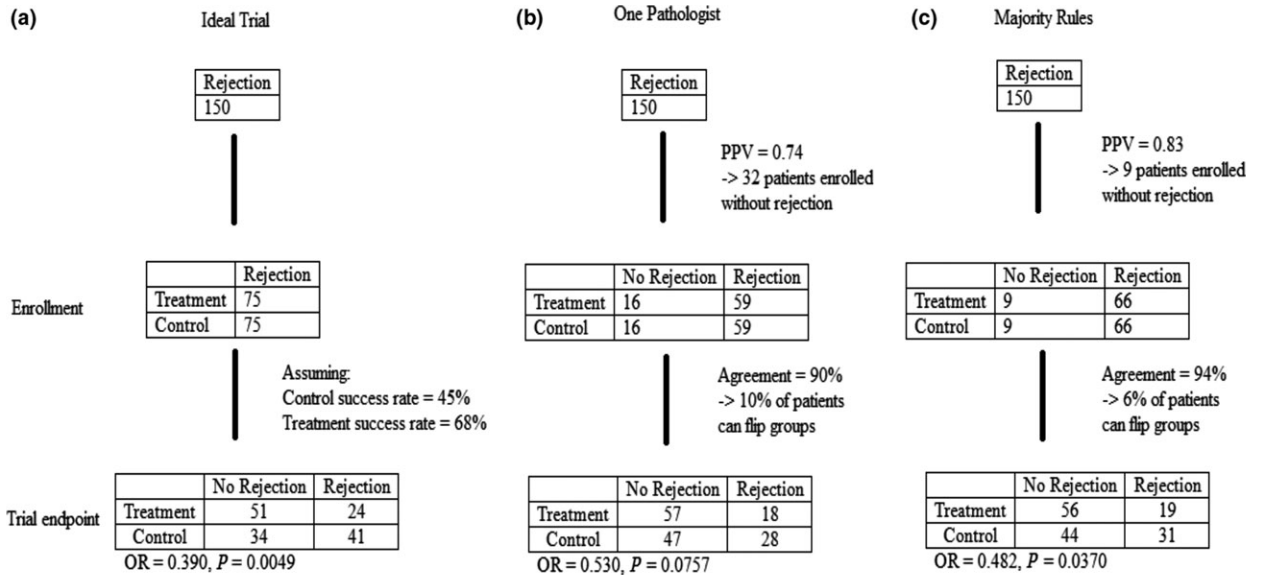


Figure 2 A flow diagram for diagnosis in a hypothetical clinical trial for chronic, active antibody-mediated rejection (cABMR) involving 150 patients. The ground truth is given by the six pathologist consensus diagnosis and the range of diagnoses either in agreement or not is given for a three pathologist consensus or a single pathologist.


(a)
Ideal Trial
(b)
One Pathologist
(c)
Majority Rules
Rejection
150
PPV = 0.74
-> 32 patients enrolled without rejection
PPV = 0.83
-> 9 patients enrolled without rejection
Enrollment
Assuming:
Control success rate = 45%
Treatment success rate = 68%
Agreement = 90%
-> 10% of patients can flip groups
Agreement = 94%
-> 6% of patients can flip groups
Trial endpoint

(a) Enrollment

	Rejection
Treatment	75
Control	75

(a) Trial endpoint

	No Rejection	Rejection
Treatment	51	24
Control	34	41

OR = 0.390, $P = 0.0049$

(b) Enrollment

	No Rejection	Rejection
Treatment	16	59
Control	16	59

(b) Trial endpoint

	No Rejection	Rejection
Treatment	57	18
Control	47	28

OR = 0.530, $P = 0.0757$

(c) Enrollment

	No Rejection	Rejection
Treatment	9	66
Control	9	66

(c) Trial endpoint

	No Rejection	Rejection
Treatment	56	19
Control	44	31

OR = 0.482, $P = 0.0370$

Figure 3 A mock trial involving 150 patients. In the ideal situation a pathologist would always and only identify positive cases as positive for chronic, active antibody-mediated rejection (cABMR). Given the PPV and diagnostic agreement for one pathologist and a three pathologist "majority rules", the conclusion of the trial could be affected.

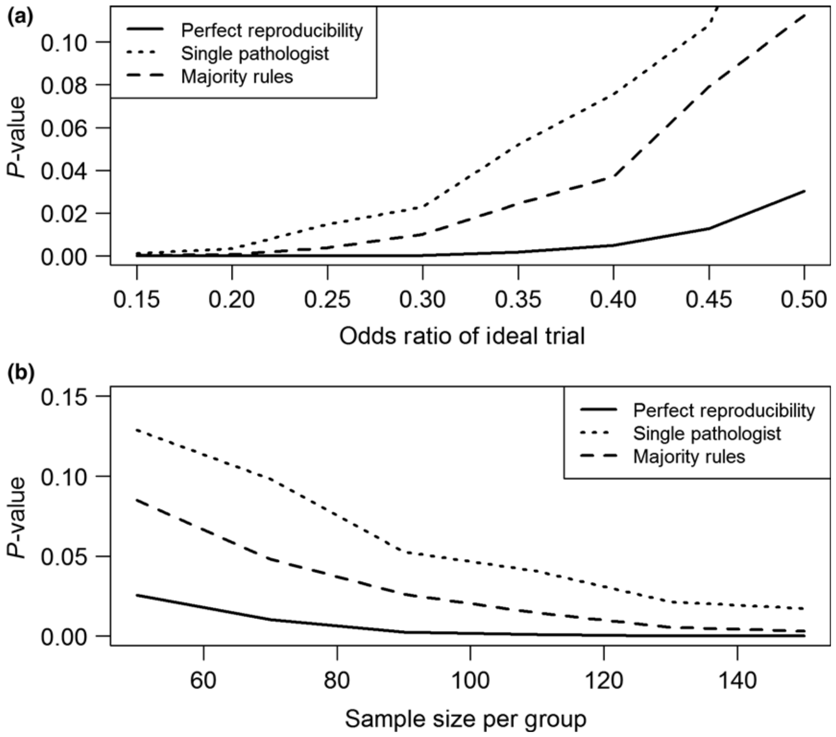


Figure 4 The effect of a lack of reproducibility versus (a) effect size and (b) sample size based on the above agreement and PPV. In order to produce lines, the mock trial was repeated for a variety of different effect sizes (odds ratios) and sample sizes.

0.73), arteriolar hyalinosis (ICC = 0.06–0.38), and acute cellular rejection (kappa = 0.47–0.72) [3,17–20].

Emphasizing the importance of reproducibility of the histologic scores comprising the diagnosis of ABMR is not merely an academic argument. Currently, there are two Phase III clinical trials enrolling patients with biopsy-proven ABMR with the goal of either preventing eGFR decline/graft loss [13] or preventing the progression of cg [14]. The distribution of diagnoses in the two trials with the same biopsies, but a different central pathologist might be: 16.4–22.4% for no ABMR; 17.9–47.8% for aABMR; and 35.8–59.7% for cAMBR. This variability might have a major impact on determining the efficacy of a potential therapeutic agent, whether the drug effect is underestimated (false negative), or the effect cannot be reproduced in subsequent trials or clinical settings (false positive). Furthermore, our data show the progression of cg is subject to major variability when the damage is mild to moderate (Banff cg score 1/Banff cg 2). We suggest that a “majority rules” approach using the scores of three pathologists may be an appropriate compromise when using the Banff ABMR schema for clinical trials. In practice, a study could have a local read and a central read by a study pathologist. If there is disagreement regarding either criteria for inclusion or the primary endpoint, a third pathologist reread would be needed to determine a “majority rules” score.

Although we found that the reproducibility of individual Banff scores was relatively low, we do not conclude that certain pathologists are performing poorly. On the contrary, we believe that the lack variability in pathologic interpretation results from the fact that currently pathologists are forced to provide a discrete score to a complex continuous process. Not only does the pathologist have to determine the presence or absence of a specific lesion (i.e., glomerulitis), the pathologist must determine the severity of that score (1–3). For cg, the pathologist must also determine the presence and score in the “most affected glomerulus not globally sclerosed.” Our data revealed that despite the lack of *exact* agreement, there was actually good *general* agreement among pathologists.

Going forward, reproducibility might be increased in other ways. First, use of images has been shown to improve inter-rater reproducibility [12,21]. Next, a morphometric scoring system might decrease scoring variability. In this approach, a pathologist or a trained technician would annotate computer-scanned biopsies for each lesion (i.e., determine presence or absence). A computer program then would determine the extent to which the lesion is present and thus limit variability in this aspect of scoring. This approach might also allow for a continuous scoring system obtained through morphometric analysis of biopsies which should decrease problems associated with discretizing [22,23]. For example, an interstitial fibrosis morphometric score correlations have been reported as high as 0.9636 although reproducibility was lower [17]. For cg, the fraction of the glomerular basement membrane with contour doubling may be a useful alternative. Such a system might be able to assess the progression of cg (Banff cg lesions) from mild (Banff cg score 1) to severe (Banff cg score 3) in serial biopsies.

The ideal ABMR histologic classification system should not only be highly reproducible, but also should have a high correlation with clinically important outcomes such as graft function and graft survival. While validity, or the correlation of Banff scores with outcomes, is not guaranteed by a more reliable measure, any relationship with this measure is limited by the reliability. In clinical trials, histology might be used to assess response to therapy similar to oncology studies in which a histologic response is used. Essentially, this means validation of histology as a biomarker for ABMR. Whether or not the current Banff system is sufficient or requires further modification is unclear at the present. However, we believe that a major goal of the Banff working groups going forward should be to correlate histologic data with graft outcomes in order to validate future scoring systems. The use of gene expression has been suggested to be a more accurate biomarker of ABMR than histology [24]. Modeling histology, gene expression, and clinical data might lead to a model that is truly predictive of outcomes and provides useful end points for future clinical trials.

Authorship

LC: designed this study. BS and WK: collected and analyzed data. BS, WP, MCMA, CS, and M Stegall: contributed to writing. LC, M Smith, CC, XG, MPA, and MR: participated in scoring biopsies.

Funding

CS is the partially funded by CTSA Grant Number KL2 TR002379 from the National Center for Advancing Translational Sciences (NCATS), a component of the National Institutes of Health (NIH). Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NIH.

Conflict of interest

The authors have declared no conflicts of interest.

REFERENCES

1. Stegall MD, Gaston RS, Cosio FG, Matas A. Through a glass darkly: seeking clarity in preventing late kidney transplant failure. *J Am Soc Nephrol* 2015; **26**: 20.
2. Gough J, Rush D, Jeffery J, *et al.* Reproducibility of the Banff schema in reporting protocol biopsies of stable renal allografts. *Nephrol Dial Transplant* 2002; **17**: 1081.
3. Veronese FV, Manfro RC, Roman FR, *et al.* Reproducibility of the Banff classification in subclinical kidney transplant rejection. *Clin Transplant* 2005; **19**: 518.
4. Haas M, Loupy A, Lefaucheur C, *et al.* The Banff 2017 Kidney Meeting Report: revised diagnostic criteria for chronic active T cell-mediated rejection, antibody-mediated rejection, and prospects for integrative endpoints for next-generation clinical trials. *Am J Transplant* 2018; **18**: 293.
5. Cohen J. A Coefficient of Agreement for Nominal Scales. *Educ Psychol Measur* 1960; **20**: 37.
6. Lyles RH, Lin J. Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting. *Stat Med* 2010; **29**: 2297.
7. Kim MG, Kim YJ, Kwon HY, *et al.* Outcomes of combination therapy for chronic antibody-mediated rejection in renal transplantation. *Nephrology* 2013; **18**: 820.
8. Billing H, Rieger S, Susal C, *et al.* IVIG and rituximab for treatment of chronic antibody-mediated rejection: a prospective study in paediatric renal transplantation with a 2-year follow-up. *Transpl Int* 2012; **25**: 1165.
9. Mengel M, Sis B, Halloran PF. SWOT analysis of Banff: strengths, weaknesses, opportunities and threats of the international Banff consensus process and classification system for renal allograft pathology. *Am J Transplant* 2007; **7**: 2221.
10. 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. Furness PN, Taub N, Convergence of European Renal Transplant Pathology Assessment Procedures Project. International variation in the interpretation of renal transplant biopsies: report of the CERTPAP Project. *Kidney Int* 2001; **60**: 1998.
12. Furness PN, Taub N, Assmann KJ, *et al.* International variation in histologic grading is large, and persistent feedback does not improve reproducibility. *Am J Surg Pathol* 2003; **27**: 805.
13. Marcussen N, Olsen TS, Benediktsson H, Racusen L, Solez K. Reproducibility of the Banff classification of renal allograft pathology. Inter- and intraobserver variation. *Transplantation* 1995; **60**: 1083.
14. Gibson IW, Gwinner W, Brocker V, *et al.* Peritubular capillaritis in renal allografts: prevalence, scoring system, reproducibility and clinicopathological correlates. *Am J Transplant* 2008; **8**: 819.
15. Batal I, De Serres SA, Mfarrej BG, *et al.* Glomerular inflammation correlates with endothelial injury and with IL-6 and IL-1beta secretion in the peripheral blood. *Transplantation* 2014; **97**: 1034.
16. Haas M, Sis B, Racusen LC, *et al.* Banff 2013 meeting report: inclusion of c4d-negative antibody-mediated rejection and antibody-associated arterial lesions. *Am J Transplant* 2014; **14**: 272.
17. Farris AB, Chan S, Climenhaga J, *et al.* Banff fibrosis study: multicenter visual assessment and computerized analysis of interstitial fibrosis in kidney biopsies. *Am J Transplant* 2014; **14**: 897.
18. Seron D, Moreso F, Fulladosa X, Hueso M, Carrera M, Grinyo JM. Reliability of chronic allograft nephropathy diagnosis in sequential protocol biopsies. *Kidney Int* 2002; **61**: 727.
19. Liapis H, Gaut JP, Klein C, *et al.* Banff histopathological consensus criteria for preimplantation kidney biopsies. *Am J Transplant* 2017; **17**: 140.
20. Liapis G, Singh HK, Derebail VK, Gasim AM, Kozlowski T, Nickeleit V. Diagnostic significance of peritubular capillary basement membrane multilaminations in kidney allografts: old concepts revisited. *Transplantation* 2012; **94**: 620.
21. Ozluk Y, Blanco PL, Mengel M, Solez K, Halloran PF, Sis B. Superiority of virtual microscopy versus light microscopy in transplantation pathology. *Clin Transplant* 2012; **26**: 336.
22. Cox DR. Note on grouping. *J Am Stat Assoc* 1957; **52**: 543.
23. Fedorov V, Mannino F, Zhang R. Consequences of dichotomization. *Pharm Stat* 2009; **8**: 50.
24. Halloran PF, Reeve J, Akalin E, *et al.* Real time central assessment of kidney transplant indication biopsies by microarrays: the INTERCOMEX study. *Am J Transplant* 2017; **17**: 2851.