American Journal of Transplantation 2013; 13: 1235–1245
Wiley Periodicals Inc.

© Copyright 2013 The American Society of Transplantation and the American Society of Transplant Surgeons

doi: 10.1111/ajt.12193

Banff Initiative for Quality Assurance in Transplantation (BIFQUIT): Reproducibility of C4d Immunohistochemistry in Kidney Allografts

M. Mengel[a,*], S. Chan[b], J. Climenhaga[b], Y. B Kushner[c], H. Regele[d], R. B Colvin[c] and P. Randhawa[e]

[a] *Department of Laboratory Medicine and Pathology*
[b] *Alberta Transplant Applied Genomics Centre, Division of Nephrology and ImmunologyDepartment of Medicine, University of Alberta, Edmonton, Canada*
[c] *Department of Pathology, Harvard Medical School and Massachusetts General Hospital, Boston, MA*
[d] *Insitute for Pathology, Innsbruck Medical University, Innsbruck, Austria*
[e] *Division of Transplantation PathologyDepartment of Pathology, University of Pittsburgh, Pittsburgh, PA*
* *Corresponding author: Michael Mengel, mmengel@ualberta.ca*

Detection of C4d is crucial for diagnosing antibody-mediated-rejection. We conducted a multicenter trial to assess the reproducibility for C4d immunohistochemistry on paraffin tissue. Unstained slides from a tissue microarray (TMA) comprising 44 kidney allograft specimens representing a full analytical spectrum for C4d were distributed to 73 institutions. Participants stained TMA slides using local protocols and evaluated their slides following the Banff C4d schema. Local staining details and evaluation scores were collected online. Stained slides were returned for centralized panel re-evaluation. Kappa statistics were used to determine reproducibility. Poor interinstitutional reproducibility was observed (kappa 0.17), which was equally due to limitations in interobserver (kappa 0.44) and interlaboratory reproducibility (kappa 0.46). Depending on the cut-off, reproducibility could be improved by omitting C4d grading and only considering ± calls. Heat-induced epitope recovery (pH 6–7, 20–30 min, citrate buffer) with polyclonal antibody incubation (<1:80, >40 min) appeared as best practice. The BIFQUIT trial results indicated that C4d staining on paraffin sections varies considerably between laboratories. Refinement of the current Banff C4d scoring schema and harmonization of tissue processing and staining protocols is necessary to achieve acceptable reproducibility.

Key words: Banff classification, C4d, immunohistochemistry, transplantation

Abbreviations: AMR, antibody-mediated rejection; BIFQUIT, Banff Initiative for Quality Assurance in Transplantation; EDTA, ethylenediaminetetraacetic acid; IHC, immunohistochemistry; TMA, tissue microarray.

Received 31 July 2012, revised 31 December 2012 and accepted for publication 03 January 2013

Introduction

Detection of C4d in the microcirculation of allografts is crucial for diagnosing antibody-mediated rejection (AMR; Refs. [1–3]). Detailed criteria for the evaluation of C4d staining in kidney transplants are described in the 2007 Banff update (4). At the 2009 and 2011 Banff consensus meetings, refinement of criteria for evaluating C4d staining in cardiac and pancreatic allografts was accomplished (2,5,6). In routine practice, C4d is evaluated by either immunohistochemistry (IHC) on formalin-fixed, paraffin-embedded sections or by immunofluorescence on frozen sections. Despite its central importance in clinical decision making, little has been done to evaluate its reproducibility across laboratories (7–10).

Tissue microarrays (TMAs) have become a valuable tool in conducting external quality assessment trials in the area of IHC (11). TMAs allow multiple cores of paraffin-embedded tissue to be placed on a single glass slide. This slide can be tested in laboratories under identical staining conditions, thus eliminating confounders of interlaboratory variability assessments, while enabling a comprehensive assessment of numerous cases in a highly consistent, cost and time efficient manner (12).

In 2009, the Banff initiative for quality assurance in transplantation (BIFQUIT) was launched (6) and organized a multicenter BIFQUIT trial with the aim to assess and improve the standards for a reproducible C4d IHC assessment in human kidney transplant tissue. Here we present the results from the first such BIFQUIT trial.

Materials and Methods

Tissue selection and processing

Nineteen paraffin blocks from human kidney transplant nephrectomies and one native nephrectomy (negative control) were collected from eight different institutions (see Acknowledgments section). The transplant nephrec-

Table 1: Protocols for assessing the influence of tissue fixation on the C4d IHC result

Protocol	Description
Standard	Immediate onset of fixation for 24 h in buffered, standard Formalin (4%)
Delay	Delayed onset of fixation for 12 h (storing the tissue at 4°C) and then fixation for 24 h in buffered, standard Formalin (4%)
Frozen	Snap freezing of tissue (simulating a frozen section procedure) and then immediate fixation for 24 h in buffered, standard Formalin (4%)
Overfixation	Immediate onset of fixation for 5 days (simulation of shipment over long weekend) in buffered, standard Formalin (4%)
Underfixation	Immediate onset of fixation for only 1 h (including formalin time in the tissue processor, i.e. simulation of very rush processing) in buffered, standard Formalin (4%)
Ethanol	Immediate onset of fixation for 24 h in ethanol (100%) instead of formalin

After fixation, all specimens were processed and embedded in paraffin like standard kidney transplant biopsies.

tomies included one specimen from a strongly C4d positive (highly sensitized) patient which was received fresh and processed according to six different fixation protocols (described in detail in Table 1). For assembling the TMA, all blocks were first stained for C4d using the standard IHC protocol at the organizing laboratory (University of Alberta, Edmonton). Areas were then pre-selected to represent a broad analytical spectrum (i.e. from C4d negative, mild/focal C4d, to strongly/diffusely C4d positive). With the exception of the nephrectomy used for the fixation experiment, two areas per block representative for the C4d staining pattern of this case, were identified, and tissue cores (n = 44) 1.3 mm in diameter were punched out and used to construct a single TMA as described previously (12). From this TMA, standard paraffin sections were cut and mounted on coated slides suitable for IHC staining.

Trial design

The BIFQUIT trial is outlined in Figure 1. Participants from 73 institutions in 19 countries were sent unstained, paraffin embedded TMA slides. Using an online survey monkey, detailed information was requested regarding the locally applied staining protocol, the locally generated C4d scores and the general trial design.

Each single participant (multiple participants were permitted per institution) evaluated the locally stained slides following the Banff criteria for C4d scoring (4), i.e. Banff score 0, negative; 1, <10%; 2, 10% – 50%; 3, >50%. Tissue cores that were inadequate for evaluation for technical reasons (e.g. tissue core floated off, no tissue present on tissue level) were indicated as "99" in the score sheet. Following evaluation by participating institutions, locally stained slides were returned to the organizers and re-evaluated by a panel (RBC, YK, MM, PR and HR) using identical Banff C4d scoring criteria. The central panel convened in Boston, reviewed all slides at a multiheaded microscope simultaneously, and generated consensus calls for each tissue core on the returned TMA slides (44 returned TMA slides × 44 tissue cores = 1936 panel reads). From all slides returned to the panel, a "reference slide" for further comparison was identified per consensus by the panel as the "best C4d stain" out of all reviewed, locally stained TMA slides. The "best C4d stain" was identified per consensus, i.e. all panel members agreed on that this stain was one of the best ("within the top five") taking into consideration staining intensity, lack of background and specificity in staining pattern.

Statistical Analysis

All statistical calculations were done in R as the analysis platform. For an explorative data analysis we used Bland–Altman plots which show on the *x*-axis the mean of the two scores (i.e. panel + participants C4d score/2), which essentially represents the best guess as to the "correct" result and on the *y*-axis the difference between the two scores (panel C4d scores – participants C4d scores; Ref. [13]).

Weighted kappa statistics were calculated for the interinstitutional, interobserver and interlaboratory agreement, as defined below, measuring the degree of agreement between reads and taking into account the degree of agreement that is expected due to chance alone (14). The general guidelines for interpreting the significance of kappa values are: <0 as indicating no agreement (or less agreement than expected by chance), 0–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial and 0.81–1 as near perfect agreement. Weighted kappa statistics, in which disagreements between observers are rated differently (bigger discrepancies carry greater weights), were calculated unless the variance in one respective reading was below 0.25 (i.e. 25%). For these, a prevalence-adjusted kappa was calculated, which adjusts the kappa statistic to take into account the higher expected agreement between observers associated with low variance.

Assessing the reproducibility of C4d IHC on paraffin sections

The overall reproducibility of C4d staining between institutions results from two major sources: the interlaboratory (influence of technical differences in the staining protocol applied at different laboratories) and interobserver reproducibility (influence of the subjective interpretation of the stained slide). The interlaboratory reproducibility is further influenced by pre-analytical (tissue processing) and analytical (staining protocol) variables.

Interinstitutional reproducibility: In the real world diagnostic setting the overall variation for C4d assessment between institutions (i.e. the same patient tested at different transplantation centers) comprises variance originating from using different processing and staining protocols plus that from different observers applying the semi-quantitative Banff C4d scoring system. To this end, we calculated the kappa values by comparing the C4d scores provided by each participant form their locally stained slides to the corresponding C4d scores provided by the local participant from the reference case (i.e. the scores provided by the participant at the center that was identified by the panel as the "best C4d stain"). Mean kappa values were calculated for each tissue core and for all 78 participants.

Interlaboratory reproducibility: To assess the influence of the technical component on the C4d results independent from the subjectivity of individual observers, we had the central panel evaluate all locally stained TMAs. Thus, interlaboratory agreement was calculated by comparing the panel consensus read for each tissue core to the panel read of the corresponding tissue core on the reference slide. Considering that the reference stain potentially is extraordinary sensitive, i.e. turns all cases included in TMA into diffusely C4d positive, we also calculated the interlaboratory reproducibility by comparing the panel read for each tissue core to the read of the majority of participants for this particular tissue core. By this approach we aimed to assess the reproducibility across a broad analytical spectrum for the test cases. In addition, We calculated the interlaboratory reproducibility using

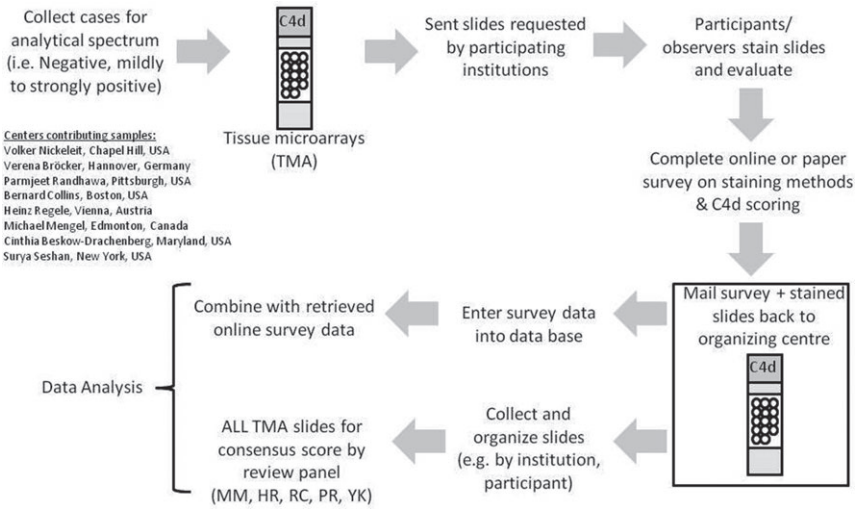

Figure 1: Design of the C4d BIFQUIT trial.

the full scale of C4d scores (C4d0-3) and using binary C4d positive/negative calls at various cut-offs (C4d0 vs. C4d1, 2, 3; Cd40, 1 vs. 2, 3; C4d0, 1, 2 vs. C4d3 (i.e. a focal vs. diffuse).

Interobserver reproducibility: This component examined the impact on reproducibility resulting from the subjective application of the Banff C4d scoring system by different observers. The interobserver reproducibility was calculated by comparing the reads of the local participants to the consensus reads of the panel recorded on the same TMA slide. Mean kappa values were calculated from the 1936 tissue cores using the two different approaches as previously stated: using the full scale of C4d scores (C4d 0–3) and using binary C4d positive/negative calls (C4d0 vs. C4d1, 2, 3; Cd40, 1 vs. 2, 3; C4d0, 1, 2 vs. C4d3 at various cut-offs (i.e. focal vs. diffuse).

Assessing the influence of pre-analytical and analytical components on the C4d result

In order to assess the influence of tissue fixation and processing (pre-analytical component) on the C4d staining result, we compared the inter-laboratory kappa values for those six tissue cores in the TMA obtained from the nephrectomy specimen with variable processing as described in Table 1. To assess the impact of different staining protocols and reagents (analytical component), the 15 laboratories with the best reproducibility, (i.e. the highest kappa-values for interlaboratory reproducibility) were compared to those fifteen with the worst in terms of the methods used for epitope recovery (enzyme vs. heat, high (EDTA buffer) versus low (citrate buffer) pH), antibody type, antibody dilution, incubation time and type of detection system. This class comparison approach was chosen over an unsupervised exploration of the raw data returned by the participants through the methods questionnaire, based on prior knowledge that IHC results in paraffin sections are primarily influenced by, tissue fixation, epitope recovery and antibody incubation/detection (8).

Results

Trial demographics

A total of 78 participants submitted their C4d scores; 44 returned stained slides to the organizing center of the 71 originally sent out. A total of 50 completed surveys for C4d staining methods and general trial questions were returned. Ninety-eight percent of the participants felt that a regular participation in external proficiency testing for C4d staining is important. Eighty percent of the participants indicated that the staining results on the BIFQUIT TMA slide were comparable to those observed in their routine practice.

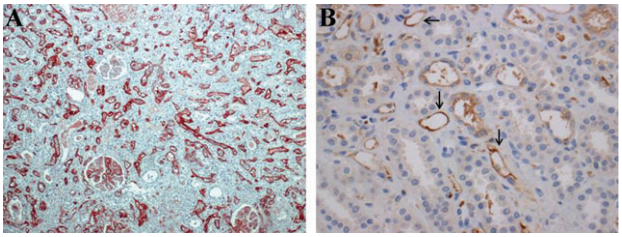

Figure 2: Diffuse C4d positive (C4d 3) tissue core on the reference slide (A, 100× magnification). A different laboratory generated linear C4d staining in peritubular capillaries (arrows in B, 400× magnification) in a native kidney specimen (negative control in the trial TMA).

The panel identified 178 of the 1936 tissue cores (9.2%) as inadequate for evaluation for technical reasons. The reference case was generated by a very sensitive laboratory with 35/44 (79.5%) of the tissue cores staining diffusely C4d positive (C4d 3, Figure 2A) as read by the panel (2/44 were read by the panel as C4d0, 2/44 as C4d1, 2/44 as C4d2, 3/44 as 99). Seven laboratories returned stains for the native kidney specimens (i.e. the supposed negative controls), which were read as C4d positive by the panel, five as C4d1 (including the reference case) and two as C4d3 (Figure 2B).

Table 2 compares the panel read on the "best C4d stained slide" with the frequencies of the corresponding participants reads done on the slides stained in their own laboratories. Strong concordance is seen for clearly positive (C4d 3) and negative (C4d 0) cases, while more variation is

observed with intermediated cases (C4d 1 and C4d 2). For some cases the "best" laboratory was able to produce a diffuse positivity for C4d while the participants laboratories produced a broad spectrum of C4d results from negative over focal to diffuse positive (e.g. case #2, indicated by only light colors in the respective row of Table 2). With those few cases in which the "best" laboratory produced a focal (C4d 1 or C4d 2) C4d result, the majority of the participants was less sensitive and reported negative C4d stains (case #9 and #10). Figure 3 shows a Bland–Altman plot revealing that the panel assigned consistently higher scores than the local participants did, resulting in a bias-index of 0.48 (bias-index = the average difference between the panel reads and the participants reads = [sum of (panel read – participant read)]/n). Both, the panel and the participants are in relatively close agreement at diffuse C4d positive and C4d negative cases (smaller bias-index at the far right and far left of the *x*-axis in Figure 3).

Reproducibility of C4d IHC on paraffin sections

Interinstitutional reproducibility: Interinstitutional reproducibility was poor when comparing all participants to the reference participant with a mean kappa-value of 0.17 (Table 3). Not for a single tissue core on the TMA, was there universal agreement on a C4d score amongst all participants. Altering the comparison benchmark (i.e. using the mean of all participants instead of that of the best performing institution), the interinstitutional reproducibility improved marginally to kappa 0.26. By simplifying the scoring schema to a positive/negative call the interinstitutional reproducibility improved markedly to kappa 0.63 if a C4d0 versus C4d $\geq$ 1 cut-off was applied and to kappa 0.54 with a C4d0, 1 versus C4d2, 3, and kappa 0.53 with a C4d $\leq$ 2 versus C4d3 cut-off.

Interlaboratory reproducibility: The average weighted kappa value for assessing the influence of the laboratory component on the reproducibility, i.e. comparing only the panel reads (= one observer) derived from all TMAs stained at various laboratories, was moderate with kappa 0.46 (Table 3). Considering only positive/negative calls for the analysis resulted in considerable improvement of the interlaboratory reproducibility with a mean kappa value of 0.77 for a C4d0 versus C4d $\geq$ 1 cut-off (kappa 0.74 with a C4d0, 1 vs. C4d2, 3, and kappa 0.63 with a C4d $\leq$ 2 vs. C4d3 cut-off). Adjusting the benchmark from the best stain to the majority call (i.e. using the most prevalent score assigned in each tissue core by the participants and thus eliminating the scoring bias imposed by the panel) improved the reproducibility (kappa 0.60).

Interobserver reproducibility: Interobserver agreement in stain interpretation was assessed by comparing the C4d scores for each tissue core assigned by the consensus panel to that assigned by the individual participant. The average interobserver kappa value was found to be moderate with 0.44 (Table 3). Comparing only positive/negative calls between observers caused the mean kappa value to improve to 0.77 for a C4d0 versus C4d $\geq$ 1 cut-off. But for the other, intermediate cut-offs, in particular for the focal versus diffuse cut-off no relevant improvement of the interobserver reproducibility was observed (kappa 0.59 with a C4d0, 1 vs. C4d2,3, and kappa 0.43 with a C4d $\leq$ 2 vs. C4d3 cut-off).

Figure 4 summarizes the reproducibility results for the C4d IHC on paraffin sections observed in this first BIFQUIT trial. Those participants performing with an at least fair or moderate overall reproducibility can be seen in the upper right corner of the graph (highlighted in gray). Sixty-nine percent (69%) of the participants performed the stain with an at least fair interobserver and interlaboratory reproducibility, but only 18% with at least moderate, if the current Banff scoring schema for C4d was applied (Figure 4 A). This markedly improved to 82% of the participants performing the stain with an at least fair interobserver and interlaboratory reproducibility (59% with at least moderate) if only a simple positive/negative call (C4d0 vs. C4d $\geq$ 1) without semiquantitative grading was applied (Figure 4B). Such scatter plots were sent out to each participant with the respective participant highlighted in the graph, thus providing individual feedback regarding the performance in the BIFQUIT trial compared to the other participants.

Significance of pre-analytical and analytical components

Insufficient fixation time (<1 h) or fixation in ethanol had a significant impact on C4d staining results between laboratories: substantial interlaboratory reproducibility after standard fixation (kappa 0.64) was reduced by under-fixation (kappa 0.23) and further after ethanol fixation to being essentially not reproducible at all (kappa –0.32, i.e. below chance).

Greater than 80% of the participating laboratories used automated stainers for the C4d IHC. Most applied a heat-based antigen recovery method, of these the majority at a pH7–10 using EDTA-based buffers (56%), followed by citrate-based buffers at pH 5–7 (26%). A broad, continuous range of dilutions for the only available polyclonal antibody were used (1:10–1:2000) while 70% of the laboratories used polymer-based systems for detection. The majority of top 15 laboratories in terms of reproducing the "best C4d stain" applied heat-induced epitope recovery using a citrate buffer at pH 6–7 on Ventana machines with a primary antibody dilution of <1:80 incubated for 20–40 min. Higher dilutions of the primary antibody were associated with a worse reproducibility compared to the reference case. These observations suggest the following protocol as optimal for a C4d immunohistochemical stain on formalin-fixed paraffin-embedded sections (Table 4): heat induced epitope recovery at pH 6–7 using citrate buffer, followed by incubation of the polyclonal antibody at <1:80 for >40 min with a polymer detection system. The laboratory which generated the "best" C4d stain (which

Table 2: Distribution of staining results provided by participants for each individual specimen included in the TMA

Number of tissue core on TMA (type of fixation as described in Table 1)	Panel consensus on "best C4d stain"	Number of participant C4d calls on corresponding slides stained in their own laboratory				
		C4d 0	C4d 1	C4d 2	C4d 3	99[1]
1	C4d 3	10	11	8	48	1
2	C4d 3	20	13	25	20	0
3	C4d 3	0	2	5	70	1
4	C4d 3	2	8	8	60	0
5	C4d 3	35	27	13	3	0
6	C4d 3	1	3	17	56	1
7	C4d 1	35	32	7	2	2
8	C4d 0	51	9	4	2	12
9	C4d 2	43	21	9	3	2
10	C4d 2	44	21	10	2	1
11	C4d 3	2	4	8	50	14
12	C4d 3	5	2	4	64	3
13	C4d 3	24	8	6	6	34
14	C4d 99	4	3	1	1	69
15	C4d 99	4	8	4	2	59
16	C4d 99	1	1	6	2	68
17	C4d 3	2	2	16	58	0
18	C4d 3	2	2	11	62	1
19	C4d 3	8	12	15	43	0
20	C4d 3	2	3	13	53	7
21	C4d 3	1	2	3	70	2
22	C4d 3	7	4	23	43	1
23	C4d 0	67	7	3	0	0
24	C4d 1	66	6	0	1	5
25	C4d 3	16	25	27	6	1
26	C4d 3	10	15	22	28	3
27	C4d 3	3	4	6	63	2
28	C4d 3	9	12	26	21	10
29	C4d 3	1	2	5	20	49
30	C4d 3	4	2	4	66	2
31	C4d 3	17	37	20	3	1
32	C4d 3	0	6	24	44	4
33	C4d 3	5	26	27	10	10
34	C4d 3	4	1	6	65	2
35	C4d 3	2	1	1	70	4
36	C4d 3	9	27	30	5	7
37 (Standard)	C4d 3	12	25	23	17	1
38 (Delay)	C4d 3	16	27	25	9	1
39 (frozen)	C4d 3	5	8	15	47	3
40 (Overfixation)	C4d 3	11	6	15	44	2
41 (Underfixation)	C4d 3	25	34	10	8	1
42 (Ethanol)	C4d 3	35	24	10	3	6
43	C4d 3	1	3	27	45	1
44	C4d 3	0	1	6	70	1

Legend (% participants) The different shades represent:
0–20
>20–40
>40–60
>60–80
>80–100

[1]99 = tissue core cannot be evaluated for technical reasons.

was subsequently used by the panel as the reference slide) had a protocol as follows: autoclave with citrate buffer for 10 min, manual staining in a humid chamber with incubation of the polyclonal antibody at 1:10 overnight (16 h) and application of an avidin–biotin based detection system.

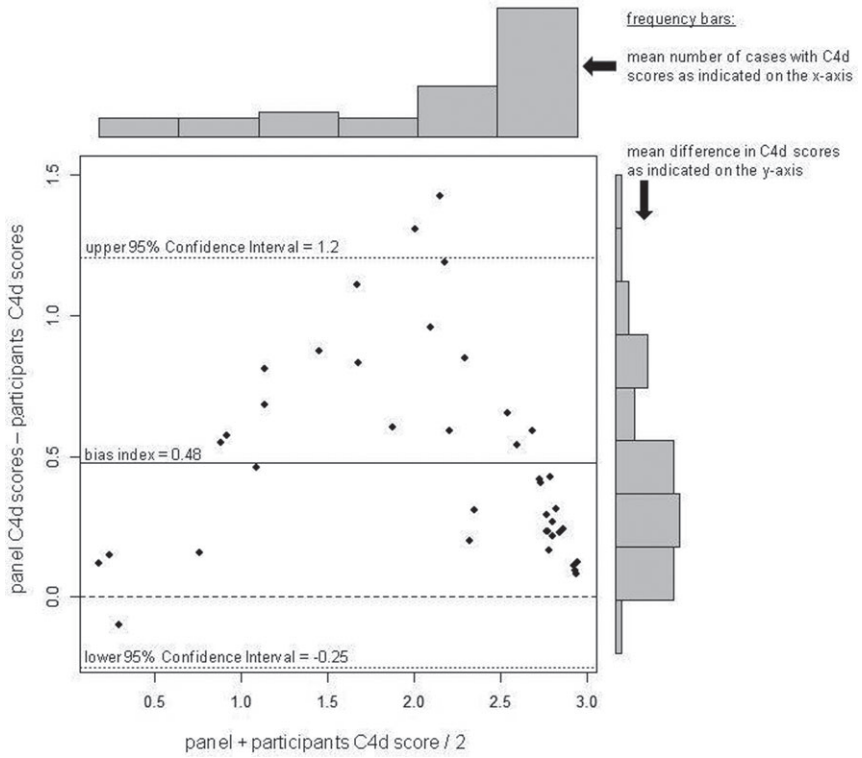

Figure 3: Bland–Altman plot: If a plot shows dots scattered all over the place, above and below zero, it suggests that there is no consistent bias of one observer (panel) versus the other (participants). The BIFQUIT Bland–Altman plot highlights that the panel provides consistently higher scores compared to the local participants, indicated by the fact that essentially all the points (= individual tissue cores on the TMA) are above the zero line. The plot also reveals, that the discrepancies between panel and observers is greatest with cases of focal C4d staining, i.e. a Banff C4d score of 2. However, most of the tested cases have a C4d score of 3 (indicated by the higher frequency bars at the top of the graph) and show only minor discrepant scoring between panel and participants (indicated by the higher frequency bars at the right-hand side of the graph).

Discussion

Current classification systems include C4d staining as an essential component for diagnosing AMR (1–3). In this multicenter, multinational Banff trial, we observed poor interinstitutional reproducibility for the C4d stain using IHC on paraffin section from renal allograft specimens. This was equally attributable to limitations in technically reproducing the stain as well as between observers using the current Banff C4d grading schema. The observations from this first BIFQUIT trial indicate that further refinement of the Banff C4d scoring schema and harmonization of tissue processing and staining protocols are necessary in order to achieve acceptable reproducibility for C4d staining on paraffin sections.

The need for ongoing proficiency testing and thorough standardization in diagnostic IHC is well recognized and part of national and international consensus guidelines (8–10,15). Class I IHC tests provide adjunctive diagnostic information to pathologists, while class II tests are so-called "stand alone" tests that are reported independently of other clinical or laboratory information. The results of these tests are frequently used as predictive biomarkers or companion diagnostics and clinicians rely on the results to stratify patients for therapies (7,8). In this regard, C4d IHC assumes an intermediate position. It is not a "stand alone" test for AMR, but the diagnosis of AMR cannot be rendered without C4d. It should be noted that alternative criteria for AMR are currently being evaluated in light of the recent recognition of C4d-negative AMR (5,16). At this time, the result of the C4d stain carries significant weight in the clinical decision-making and the presence or absence of C4d may serve as a biomarker for monitoring response to treatment and disease intensity/grade (17–19). Therefore efforts should be made to make the test as reproducible as possible.

We observed limited reproducibility between institutions for C4d results, which was equally due to differences in how the stain was performed and how the stain was interpreted. The review panel consistently assigned higher C4d scores than the local participants, indicating that local pathologists may have adjusted their scoring to their local laboratory. Such local adjustments typically happen over time by correlating C4d results with in-house clinical feedback, morphology and antibody titers, or by taking into account prior knowledge (e.g. paraffin C4d IHC is less sensitive than immunofluorescence on frozen tissue (20–22). To address this bias introduced by the panel read, which may be due to the fact that the panel comprised individuals potentially more familiar with paraffin stains than the average participant, we also assessed the reproducibility for a majority call of all participants. This only marginally improved the kappa value for interinstitutional reproducibility from 0.17 (slight) to 0.26 (fair) and the interlaboratory reproducibility from 0.46 (moderate) to 0.6 (moderate). In this BIFQUIT trial all laboratories stained the same tissue. There is the possibility that participating laboratories would have generated 'better' stains if the tissue was processed

Table 3: Summary of mean kappa-values for the different components influencing the reproducibility of a C4d stain

	Mean kappa-value comparing Banff C4d scores to reference case	Mean kappa-value comparing positive/negative calls (C4d0 vs. C4d1, 2, 3) to reference case	Mean kappa-value comparing positive/negative calls (C4d0, 1 vs. C4d2, 3) to reference case	Mean kappa-value comparing positive/negative calls (C4d0, 1, 2 vs. C4d3) to reference case	Mean kappa-value comparing Banff C4d scores to majority calls by all participants
Interinstitutional reproducibility = variation between institutions resulting from differences in the staining/laboratory procedure + subjectivity/interpretation by different observersMean kappa values were calculated by comparing the C4d scores provided by each participant from their locally stained slides to the corresponding C4d scores provided by the local participant from the reference case	0.17 ± 0.3(–0.68 to 0.65)	0.63 ± 0.2(0.23–0.92)	0.54 ± 0.2(–0.3 to 0.93)	0.53 ± 0.2(0–0.85)	0.26 ± 0.2(–0.5 to 0.64)
Interlaboratory reproducibility = fraction of the total variation resulting from differences in the staining/laboratory procedureMean kappa values were calculated by comparing the panel consensus read for each tissue core to the panel read of the corresponding tissue core on the reference slide	0.46 ± 0.6 (–0.77 to 1.0)	0.77 ± 0.1 (0.35 to 1.0)	0.74 ± 0.4 (0.17 to 1.0)	0.63 ± 0.4 (–0.22 to 1.0)	0.60 ± 0.4 (–0.36 to 1.0)
Interobserver reproducibility = fraction of the total variation resulting from subjectivity/interpretation by different observersMean kappa values were calculated by comparing the reads of the local participants to the consensus reads of the panel recorded on the same TMA slide	0.45 ± 0.2 (0.11–0.9)	0.77 ± 0.2 (0.03–1.0)	0.59 ± 0.2 (0.23–0.94)	0.43 ± 0.2 (–0.18 to 0.95)	NA

Significance of kappa values: <0 indicating no agreement (or less agreement than expected by chance), 0–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial and 0.81–1 almost perfect agreement.

NA, not applicable, because for the interobserver comparison per definition the comparator has to be the panel read and cannot be a majority call.

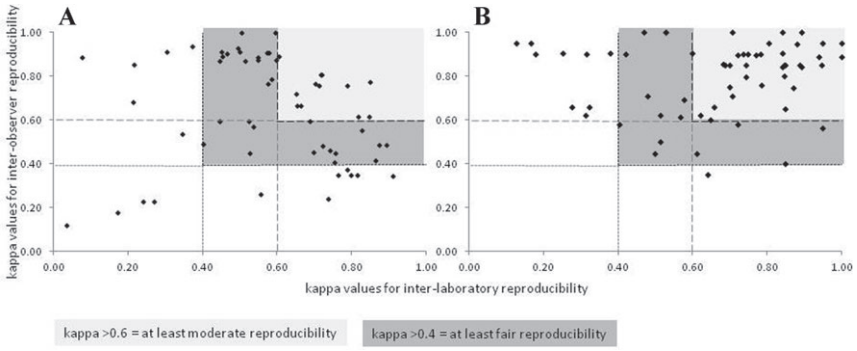

Figure 4: Scatter plots showing the interlaboratory and interobserver reproducibility for each individual participant/participating laboratory using the Banff C4d schema (A) or positive versus negative calls (B; C4d0 vs. C4d≥1).

Table 4: Recommendations for C4d immunohistochemistry on paraffin sections

Avoid under-fixation (<1 h) and/or ethanol fixation of tissue specimen
Use heat induced epitope recovery with citrate buffer at pH 6–7
At least incubate polyclonal anti-C4d antibody at a dilution of <1:80 for >40 min
Better results were observed with harsher pre-treatment (e.g. autoclave epitope recovery), higher polyclonal anti-C4d antibody concentrations (1:10), and longer incubation times for the polyclonal anti-C4d antibody (over night/12–16 h)
Refinement of the 2007 Banff C4d grading schema has potential to improve interobserver reproducibility

at their institution. However, minor changes in tissue processing, as included in the trial TMA (Table 1), did not significantly alter the results. Only severe mishandling of the tissue, i.e. severe underfixation (<1 h) and ethanol fixation had significant negative effects on C4d results. Most laboratories now use standardized formalin fixation and automated tissue processing. Thus, the pre-analytical components likely only have minor impact on the real world variance in C4d IHC.

It seems to be relatively safe to recommend a stringent and most sensitive staining protocol for C4d IHC with the aim of detecting as many AMR cases as possible. Few potentially false positive results were seen in this BIFQUIT trial (7/44 laboratories, mostly as C4d1, only two C4d3 false positives). We used histologically unremarkable tissue from tumour nephrectomies as negative controls, but the possibility of complement activation in the setting of ischemia cannot be entirely ruled out; this may have been detected by "sensitive" laboratories as positive C4d. However, the more likely explanation is that complement rich plasma stuck non-specifically on the endothelium, and was misinterpreted by the observers (including the panel) as specific C4d deposition in capillaries of native kidneys. Previous single center studies from Basel, Switzerland already described that C4d staining using immunofluorescence in frozen tissue is not only more sensitive but also more reproducible between observers (20). The interobserver kappa values between two observers from the same institution for frozen immunofluorescence stains were perfect ((1).0) and for paraffin sections only good (0.57–0.63). Better reproducibility on paraffin sections was observed with higher dilutions of the primary antibody at the expense of further decreased sensitivity. These data and our own observations indicate that with C4d IHC on paraffin section at high primary antibody concentrations, nonspecific pseudo-linear C4d stain in peritubular capillaries can occur as background staining and be misinterpreted as C4d positivity.

The motivation for introducing the Banff C4d scoring schema at the 2007 meeting (4) was to standardize C4d IHC and immunofluorescence reporting, thus making them more amenable for comparisons between studies to evaluate the clinical relevance of focal or negative C4d. A generally accepted notion is that detailed standard consensus criteria for histological assessment improve reproducibility and thus make data comparable (23). However, it stands to reason and previous studies have already validated that if a grading schema is overly complex, reproducibility is poor thus defeating the purpose of the entire system (24). As expected and previously shown in a smaller studies (20) and confirmed here, specimens showing intermediate C4d scores (i.e. C4d1 or C4d2) are less reproducible compared to those with negative or diffusely positive C4d. A simplified grading schema, i.e. a positive versus negative call (any C4d stain vs. no C4d stain) was associated with improved reproducibility. However, interobserver reproducibility remained weak if a positive/negative call was applied with cut-off at the focal/diffuse interface (C4d $\leq$ 2 vs. C4d3), i.e. the intermediate scores with greatest variation across all participants, potentially explaining the conflicting data in the literature regarding the significance of focal C4d positivity (18,25–28). Even using a highly simplified scoring schema (any C4d stain vs. no C4d stain, which essentially is a reduction ad absurdum), only 59% of the participants performed and evaluated the stain with at least a moderate interobserver and interlaboratory reproducibility. Altering the benchmark for C4d positivity, e.g. adjusting it to the majority and not the "best" participating laboratories improved reproducibility. However, this is at the cost of limited sensitivity of detecting C4d, since the "best"

laboratories stained considerably more cases as C4d 3, than the majority of participants did. Furthermore, defining C4d cut-offs at which clinical interventions get triggered, represents a completely different task requiring appropriate validation. But before this can be scope of prospective studies, standardization of C4d staining is paramount, otherwise the threshold for therapeutic interventions need to be established and validated at each transplant center independently.

Regular participation in IHC quality assurance trials is known to lead to improved reproducibility following individual participant feedback (15). We provided each participant with individual performance feedback. This information together with best practice recommendations for staining and tissue processing (see Table 4) can lay the foundation for subsequent BIFQUIT trials with the aim to monitor improvement of the reproducibility of C4d IHC staining. However, a limitation of this first BIFQUIT trial is that the technical protocol details were reported very heterogeneously preventing an unsupervised statistical exploration with the aim to identify the true best practice. But international standardization of C4d IHC seems feasible, since only one polyclonal antibody is currently commercially available and standardized processing of core biopsies is achievable using automated tissue processors and strainers. Combining standardized staining protocols with computer based image analysis algorithms might then be the next step towards acceptable reproducibility of C4d. A recent single center experience in a small series of C4d positive cases observed a superior reproducibility using digital pathology and image analysis for C4d assessment compared to pathologists visually scoring the slides (29). The Banff working group for digital pathology could take this on as a task to generate a consensus image analysis algorithm for C4d IHC. The ultimate aim of such concerted Banff efforts would be to assure that every patient receives an identical C4d result independent of where his or her biopsy is obtained and evaluated.

Acknowledgments

We thank Victoria Sheldon and Akshatha Raghuveer for an outstanding logistical support. This trial was supported by a research grant from Astellas Canada Inc. to M.M for facilitating Banff Working Group activities.

The following institutions contributed tissue for constructing the BIFQUIT C4d TMA:

Volker Nickeleit, Chapel Hill, USA; Verena Bröcker, Hannover, Germany; Parmjeet Randhawa, Pittsburgh, USA; A. Bernard Collins, Boston, USA; Heinz Regele, Vienna, Austria; Michael Mengel, Edmonton, Canada; Cinthia Beskow-Drachenberg, Maryland, USA; Surya Seshan, New York, USA.

The authors would like to thank all centers participating in the BIFQUIT trial for contributing their time and resources, and valuable feedback during and after the trial. We apologize to those participants to whom we were unable to deliver the trials slides (usually due to customs regulations) and those participants from whom we received the slides too late to be reviewed by the panel and included into this analysis.

Centers who registered for the Banff C4d BIFQUIT trial:
Department of Pathology, The Methodist Hospital, Houston, TX, USA
The University of North Carolina, Department of Pathology and Laboratory Medicine, Division of Nephropathology, Chapel Hill, NC, USA
The Ohio State University, Columbus, OH, USA
Oregon Health & Science University, Department of Pathology, Portland, OR, USA
Department of Pathology Dartmouth-Hitchcock Medical Center, Lebanon NH, USA
University of Illinois Medical Center, Department of Pathology, Chicago IL, USA
Anatomic Pathology, Laboratory Service, North Florida/South Georgia Veterans Health System, Gainesville, Fl, USA
Huntsman Cancer Hospital University of Utah Department of Pathology, Salt Lake City, UT, USA
Department of Pathology, Cedars-Sinai Medical Center, Los Angeles, CA, USA
University of Virginia Health Sciences Center, Department of Pathology, Charlottesville, VA, USA
Department of Pathology University of Chicago Medical Center, Chicago, IL, USA
Pathology Department London Health Sciences centre, London, Ontario Canada
Department of Laboratory Medicine St. Michael's Hospital, Toronto, Ontario, Canada
Washington University School of Medicine – Department of Pathology and Immunology, Division of Anatomic Pathology, St Louis, MO, USA
Department of Pathology University of Maryland Hospital, Baltimore MD, USA
Dept. of Laboratory Medicine & Pathology Memorial Medical Center, Springfield, IL, USA
Department of Pathology University of Texas Medical Branch Department of Pathology, Galveston, TX, USA
HOSPITAL DAS CLÍNICAS- Prédios dos Ambulatórios Divisão de Anatomia Patológica, São Paulo,-Brazil
Hospital Infantil de Mexico "Federico Gomez" Departamento de Patologia Calle, Mexico City, MEXICO
Department of Cellular & Anatomical Pathology Derriford Hospital, Plymouth, UK
Department of Cellular Pathology, John Radcliffe Hospital,, Oxford, UK
Department of Clinical and Transplant Pathology Institute for Clinical and Experimental Medicine, Prague, Czech Republic
Department of Pathology Health Sciences Centre, Winnipeg, MB, Canada
Dept of Pathology, University Health Network, University of Toronto, Toronto, ON, Canada
Department of Laboratories, Seattle Children's Hospital, Seattle, WA, USA
ProPath, Dallas, TX, USA
Dept Pathology, Foothills Medical Centre, Calgary AB, Canada
Pontificia Universidad Católica de Chile Escuela de Medicina Departamento de Anatomía Patológica, Santiago, Chile
Department of Pathology Oslo University Hospital, Oslo, Norway
Transplantation Laboratory- HUSLAB Helsinki University Central Hospital, Helsinki, Finland
Surgical Pathology Montefiore Medical Center, NY, USA
Division of Transplant Pathology, University of Pittsburgh, Department of Pathology, UPMC-Montefiore Hospital, Pittsburgh, PA, USA
Dept of Pathology – Presbyterian Hospital Weill Cornell Medical College, New York, NY, USA
Institut fuer Pathologie Medizinische Hochschule Hannover, Hannover, Germany
Klinisches Institut für Pathologie, Wien, Austria

Department of Pathology Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA
Department of Pathology Singapore General Hospital, Singapore
Service de Pathologie CHUQ, Hôtel-Dieu de Québec, Québec City, Québec, Canada
Department of Pathology, Princess Margaret Hospital, Hong Kong
RUA JOSÉ MARIA DE OLIVEIRA CASACA, BAIRRO JARDIM MARIA CÂNDIDA SÃO JOSÉ DO RIO PRETO, SP, Brazil
Dept. of Pathology HADASSAH MEDICAL ORGANIZATION, HADASSAH UNIVERSITY HOSPITAL KIRYAT HADASSAH, Jerusalem, Israel
Emory University Department of Pathology Emory University Hospital, Atlanta, GA, USA
Department of Cellular Pathology Barts and the London NHS Trust, London, UK
Department of Pathology LSU Health Sciences Center, LA, USA
Nephorology Department Hospital Vall d'Hebron, Barcelona, Spain
Department of Pathology St. John Hospital & Med. Ctr. Detroit, MI, USA
University of Arizona Department of Pathology, Tucson, AZ, USA
1st Department of Pathology Medical School National and Kapodistrian, Athens, Greece
Imperial College Healthcare NHS Trust Hammersmith Hospital Dept Histopathology, London, UK
Leiden University Medical Center Dept. of Pathology, Leiden, The Netherlands
Department of Pathology Massachusetts General Hospital, Boston MA, USA
Dept. of Pathology, University of Washington Medical Center, Seattle, WA, USA
Department of Pathology Albert Einstein Medical Center Philadelphia, PA, USA
Intermountain Central Laboratory, Salt Lake City, UT, USA
Dept. of Pathology, UMC Utrecht, Utrecht, The Netherlands
Pathology & Laboratory Medicine, St. Paul's Hospital, Vancouver, BC, Canada
Wake Forest Univesrsity School of Medicine Dept. of Pathology, Winston-Salem, NC, USA
PATHOLOGY Mayo Medical Laboratories, Rochester, MN, USA
Dept. of Pathology University of Iowa Hospital, Iowa City, IA, USA
Pathologische ontleedkunde UZ Leuven campus, Leuven, Belgium
Clinical Pathology and Cytology Gula Stråket, Göteborg, Sweden
Histopathology Department Mubarak Al Kabeer Hospital City of Jabriyah Governate of Hawally State of Kuwait
Surgical Pathology QA and Compliance Fletcher Allen Health Care/University of Vermont, Burlington, Vermont, USA
Department of Anatomical Pathology Austin Hospital, Heidelberg, Australia
Rhode Island Hospital, Providence, RI, USA
Institute of Pathology Faculty of Medicine University of Ljubljana, Ljubljana Slovenia
Cellular Pathology – University Hospitals Birmingham NHS Foundation Trust The Medical School, Birmingham, UK
Department of Histopathology, Postgraduate Institute of Medical Education and Research, Chandigarh, INDIA
Consultant Pathologist MBC, Riyadh, Kingdom of Saudi Arabia
Rua Candido Gaffree, Rio de Janeiro, Brazil
Section of Pathology/Anatomía Patológica Fundació Puigvert, Barcelona. Spain
Department of Pathology Baystate Medical Center, Tufts University School of Medicine, Springfield, MA, USA
Department of Pathology Medical University of South Carolina, Charleston, SC, USA
Servicio de Anatomia Patologica Hospital Universitario Miguel Servet, Zaragoza, Spain
Department of Pathology and Laboratory Medicine University of Wisconsin-Madison, Madison, WI, USA
Institute for Pathology University Clinic, Basel, Switzerland
Pathology Department, The Johns Hopkins Medical Institutions, Baltimore MD, USA
Department of Laboratory Medicine and Pathology, University of Alberta, Edmonton, Canada

Disclosure

The authors of this manuscript have conflicts of interest to disclose as described by the *American Journal of Transplantation*. Michael Mengel is co-owner of multiblock GmbH (Hannover, Germany), a company producing TMAs and providing services for quality assurance in diagnostic IHC. None of the other authors has to disclose any conflicts of interest as described by the American Journal of Transplantation.

Results of the BIFQUIT-Trial were presented in part at the 11th Banff Meeting held in 2011 and at the 2012 Annual Meetings of the United States Canadian Academy of Pathology as well as of the *American Society of Transplantation*.

References

1. Berry GJ, Angelini A, Burke MM, et al. The ISHLT working formulation for pathologic diagnosis of antibody-mediated rejection in heart transplantation: evolution and current status (2005–2011). J Heart Lung Transplant 2011; 30: 601–611.
2. Drachenberg CB, Torrealba JR, Nankivell BJ, et al. Guidelines for the diagnosis of antibody-mediated rejection in pancreas allografts-updated Banff grading schema. Am J Transplant 2011; 11: 1792–1802.
3. Racusen LC, Colvin RB, Solez K, et al. Antibody-mediated rejection criteria – an addition to the Banff 97 classification of renal allograft rejection. Am J Transplant 2003; 3: 708–714.
4. Solez K, Colvin RB, Racusen LC, et al. Banff 07 classification of renal allograft pathology: updates and future directions. Am J Transplant 2008; 8: 753–760.
5. Mengel M, Sis B, Haas M, et al. Banff 2011 Meeting Report: New Concepts in Antibody-Mediated Rejection. Am J Transplant 2012; 12: 563–570.
6. Sis B, Mengel M, Haas M, et al. Banff '09 Meeting Report: Antibody Mediated Graft Deterioration and Implementation of Banff Working Groups. Am J Transplant 2010; 10: 464–471.
7. Goldstein NS, Hewitt SM, Taylor CR, Yaziji H, Hicks DG. Recommendations for improved standardization of immunohistochemistry. Appl Immunohistochem Mol Morphol 2007; 15: 124–133.
8. Torlakovic EE, Riddell R, Banerjee D, et al. Canadian Association of Pathologists-Association canadienne des pathologistes National Standards Committee/Immunohistochemistry: best practice recommendations for standardization of immunohistochemistry tests. Am J Clin Pathol 2010; 133: 354–365.
9. Clinical and Laboratory Standards Institute. Quality Assurance for Design Control and Implementation of Immunohistochemistry Assays: Approved Guideline. 2 ed. Wayne, PA, USA: CLSI, 2011.

10. Taylor CR. New revised Clinical and Laboratory Standards Institute Guidelines for Immunohistochemistry and Immunocytochemistry. Appl Immunohistochem Mol Morphol 2011; 19: 289–290.
11. Mengel M, von Wasielewski R, Wiese B, Rudiger T, Muller-Hermelink HK, Kreipe H. Inter-laboratory and inter-observer reproducibility of immunohistochemical assessment of the Ki-67 labelling index in a large multi-centre trial. J Pathol 2002; 198: 292–299.
12. Mengel M, Kreipe H, von Wasielewski R. Rapid and large-scale transition of new tumor biomarkers to clinical biopsy material by innovative tissue microarray systems. Appl Immunohistochem Mol Morphol 2003; 11: 261–268.
13. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. Statistician 1983; 32: 307–317.
14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159–174.
15. Wasielewski R, Hasselmann S, Ruschoff J, Fisseler-Eckhoff A, Kreipe H. Proficiency testing of immunohistochemical biomarker assays in breast cancer. Virchows Arch 2008; 453: 537–543.
16. Sis B, Jhangri G, Bunnag S, Allanach K, Kaplan B, Halloran PF. Endothelial gene expression in kidney transplants with alloantibody indicates antibody-mediated damage despite lack of C4d staining. Am J Transplant 2009; 9: 2312–2323.
17. Archdeacon P, Chan M, Neuland C, et al. Summary of FDA antibody-mediated rejection workshop. Am J Transplant 2011; 11: 896–906.
18. Colvin RB. Antibody-mediated renal allograft rejection: diagnosis and pathogenesis. J Am Soc Nephrol 2007; 18: 1046–1056.
19. Wiebe C, Gibson IW, Blydt-Hansen TD, et al. Evolution and clinical pathologic correlations of de novo donor-specific HLA antibody post kidney transplant. Am J Transplant 2012; 12: 1157–1167.
20. Seemayer CA, Gaspert A, Nickeleit V, Mihatsch MJ. C4d staining of renal allograft biopsies: a comparative analysis of different staining techniques. Nephrol Dial Transplant 2007; 22: 568–576.
21. Chantranuwat C, Qiao JH, Kobashigawa J, Hong L, Shintaku P, Fishbein MC. Immunoperoxidase staining for C4d on paraffin-embedded tissue in cardiac allograft endomyocardial biopsies: comparison to frozen tissue immunofluorescence. Appl Immunohistochem Mol Morphol 2004; 12: 166–171.
22. Nadasdy GM, Bott C, Cowden D, Pelletier R, Ferguson R, Nadasdy T. Comparative study for the detection of peritubular capillary C4d deposition in human renal allografts using different methodologies. Human Pathol 2005; 36: 1178–1185.
23. 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: 2221–2226.
24. 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: 805–810.
25. Kedainis RL, Koch MJ, Brennan DC, Liapis H. Focal C4d+ in renal allografts is associated with the presence of donor-specific antibodies and decreased allograft survival. Am J Transplant 2009; 9: 812–819.
26. Kikic Z, Regele H, Nordmeyer V, et al. Significance of peritubular capillary, glomerular, and arteriolar C4d staining patterns in paraffin sections of early kidney transplant biopsies. Transplant 2011; 91: 440–446.
27. Magil AB, Tinckam KJ. Focal peritubular capillary C4d deposition in acute rejection. Nephrol Dial Transplant 2006; 21: 1382–1388.
28. Worthington JE, McEwen A, McWilliam LJ, Picton ML, Martin S. Association between C4d staining in renal transplant biopsies, production of donor-specific HLA antibodies, and graft outcome. Transplant 2007; 83: 398–403.
29. Brazdziute E, Laurinavicius A. Digital pathology evaluation of complement C4d component deposition in the kidney allograft biopsies is a useful tool to improve reproducibility of the scoring. Diagn Pathol 2011; 6(Suppl 1): S5.