Article

Int J Organ Transplant Med 2026; 33(2), 33–47

Reporting Sufficiency of Continuous Kidney Allograft Rejection Scores Across Cohort Workload and Lesion Composition

Clyde F. Barker[1], Aneesha Agarwal[2,*]

[1] *Hospital of the University of Pennsylvania, Philadelphia, United States*
[2] *Indian Institute of Science Education and Research Bhopal*
[*] *Correspondence: aneeshaagar@mbu.iisc.ernet.in*

ABSTRACT

Graded rejection scores can provide additional information quantitatively regarding kidney allograft biopsy reports, but only if the claim about reporting this score is in line with the conditions of diagnosis under which the score is reported. In this study we investigate which continuous kidney allograft rejection scores maintain the reporting function in validation cohorts, which involve local workload annotation, and which ones require additional etiological evidence for interpretation. Analysis is based on aggregated values obtained from 6272 derivation biopsies, 11,043 European validation biopsies, and 2185 United States validation biopsies. No individual patient records were reconstructed. Cohort diagnostic workload was calculated as cases per 100 biopsies, cohort displacement was measured using Jensen-Shannon divergence and log2 load shift, task stability was estimated through stable discrimination load and portability-floor score, residual diagnostic signal was estimated through ancillary signal deficit, lesion composition was estimated using entropy-derived effective lesion number and dominant lesion share. Workload of TCMR/TI including PVAN tasks increased from 19.32 cases per 100 biopsies in the derivation cohort to 55.47 cases per 100 biopsies in the United States validation cohort. Jensen-Shannon divergence from the derivation profile is equal to 0.0094 for the European validation cohort and 0.0997 for the United States validation cohort. Tasks of broad rejection separation, Borderline TCMR/TCMR, TCMR, and AMR/MVI have high values of portability-floor, while PVAN has the highest value of ancillary signal deficit – 18.37 per 100 phenotype cases. Lesion composition further differentiates balanced and lesion-dominant scores – AMR/MVI and activity have effective lesion numbers 3.98 and 5.95, while TCMR/TI and chronicity have dominant lesion shares 46.3% and 40.0%, respectively. It should be noted that the answer to the research question is task-dependent: AMR/MVI and activity allow for direct composite scoring; TCMR/TI, chronicity, mixed rejection, and Borderline TCMR/TCMR require local and component-based scoring; while PVAN requires additional viral findings besides the inflammation score.

KEYWORDS: kidney transplantation; Banff classification; allograft rejection; antibody-mediated rejection; T cell-mediated rejection; microvascular inflammation; continuous scoring; cohort validation; lesion composition; BK polyomavirus nephropathy.

1 Introduction

Interpretation of kidney allograft biopsy findings is an organized diagnostic procedure where histological lesions, immunologic findings, clinical context, and any other available information are integrated to produce a report that needs to guide treatment and follow-up. The Banff classification is the primary international diagnostic nomenclature for this activity [1, 2] due to its standardization of lesion scoring and categorization of antibody-mediated rejection (AMR), T cell-mediated rejection (TCMR), borderline inflammation, mixed rejection, chronic active processes, chronic injury, recurrent disease, infection, and biopsies without rejection [3]. Reports from the

last several Banff meeting [4, 5] have reinforced the idea that rejection is not an immutable condition but rather a spectrum of phenotypes characterized by microvascular inflammation, antibody findings, tubulointerstitial inflammation, chronic injury, transcript expression patterns, and non-rejection diagnoses [6].

Category-based diagnosis is critical, but the biological signal underlying the category is continuous. Two biopsies that received the same AMR diagnosis can differ with respect to the degree of glomerulitis, peritubular capillaritis, C4d positivity, chronic glomerulopathy, donor-specific antibody findings, and transcript activity. Two biopsies that received the TCMR diagnosis can differ in relation to the dominant lesion (tubulitis, interstitial inflammation, intimal arteritis, or inflammation in scarred cortex). Chronicity also varies by compartment since chronic glomerulopathy, interstitial fibrosis, tubular atrophy, and chronic vascular injury carry different diagnostic meaning. Continuous score can preserve the continuous level of injury, but the report must still describe the score value and the constituent lesions responsible for the result.

The recent introduction of continuous kidney transplant rejection indices has made the reporting issue from speculative to pragmatic. Vaulet and colleagues described continuous indices for AMR/MVI, TCMR/TI, activity, and chronicity for a large derivation cohort and two validation cohorts thus providing transplant pathology community with quantifiable scores that preserve Banff lesion structure [7]. The indices are promising since they recognize the continuous nature of the rejection severity. But, a strong discrimination in a derivation cohort does not automatically ensure the same reporting status of the index in every clinical service. A clinical service that gets many protocol biopsies without rejection will interpret the score differently from the service that is rich in cellular inflammation, mixed rejection, or viral nephropathy.

The problem, therefore, is not only in discrimination ability of the score. The problem is in the sufficiency of the clinical context for the reporting of the score. Discrimination, local diagnostic workload, and lesion composition should be taken into account simultaneously. High AUC for common tasks can provide enough reason for routine quantitative language. High AUC for rare or etiologically complex tasks may require more caution. Balanced composite score can be easily reported as a summary score while unbalanced score that contains one lesion of the maximum possible weight is harder to report.

Microvascular inflammation serves as an example. The 2022 Banff reconsidered MVI and biopsy-based transcript diagnosis while subsequent work highlighted that MVI can occur in contexts other than classic antibody-mediated rejection and should be interpreted taking into account antibody findings, transcript expression, chronicity, and clinical information [8]. Persistent inflammation after AMR therapy is also clinically important that promotes the reporting of the score while warns against treating the score as the full explanation of the condition [9]. Thus, the AMR/MVI score can be a useful tool for assessment of microvascular injury provided that the score stays grounded on Banff definitions and immunologic information.

The case of cellular rejection and tubulointerstitial inflammation is different. Banff TCMR categories rely on the following structures: tubulitis, interstitial inflammation, arteritis, and chronic active lesions [10]. Interpretation of inflammation in the scarred cortex is especially dependent on disease history and lesion composition [11, 12]. A TCMR/TI score can be discriminating enough, but still susceptible to the changes in the diagnostic workload of the receiving cohort due to the abundance of inflammation-related pathology. In this case, the numerical score requires a component note since the same sum of score can be achieved by a combination of tubulitis, interstitial inflammation, intimal arteritis, PVAN-related inflammation, or mixed immune injury.

PVAN is the clearest example of a score that requires information external to rejection-oriented histology score. BK polyomavirus nephropathy is an etiologic viral diagnosis, not an inflammation pattern. Consensus guidelines stress importance of screening for plasma BKPyV DNAemia, tissue assessment [13, 14], immunohistochemistry when needed, immunosuppression reduction, and differential diagnosis from rejection [15, 16]. A continuous inflammation score can be helpful for characterization of injury, but it can never replace viral load, SV40 large T antigen immunohistochemistry, and diagnosis of polyomavirus nephropathy.

Broader reporting context is also evolving. Antibody standardization becomes increasingly important for

transplant diagnostics [17]. The DSA interpretation is also evolving in line with this trend [18]. Molecular pathology and transcript diagnostics are becoming prominent as well [19]. Clinical translation of these trends includes structured transcript tools and molecular classifiers [20], as well as phenotype and archetype approaches [21, 22]. Digital pathology work shows that quantitative estimation of inflammation and chronic lesions is reproducible and clinically significant [23, 24], but also shows that this output requires appropriate link to Banff categories, specimen context, and validation setting [25]. All these developments put continuous histological scores into the reporting problem: the score must stay linked to Banff categories, specimen context, and the validation setting.

A second argument for caution is that the unit of clinical communication is the report, not the score. The report must clarify whether the number allows making a diagnosis, quantifying activity of a diagnosis, assessing chronic injury burden, or asking for additional information. These uses are not interchangeable. The score that is suitable for summary estimation of microvascular injury may not be suitable for making a distinction between alloimmune and viral tubulointerstitial inflammation. Chronicity score can be very useful for severity quantification but still requires a report statement about the type of chronic injury (glomerular, interstitial, tubular, or vascular). The same numerical score, therefore, can be differently reported depending on the lesion system and the clinical question associated with the score.

Literature on external validation of scores and models provides additional warning. Prediction-model guidelines stress that discrimination, intended use, and applicability must be assessed together [26]. Calibration is also crucial for transportation of a model to another environment [27]. Continuous biopsy scores are not the same as clinical prediction models, but the problem of portability is the same: numerical result that works in one cohort can encounter a different pretest environment in another setting. In the case of transplant pathology, the pretest environment is visible through biopsy stream. Proportion of no-rejection biopsies, burden of TCMR/TI, frequency of mixed rejection, and viral nephropathy affect how often the score will be used and how many alternative interpretations will compete for the interpretation.

The research question is quite clear: when the interpretation is limited to the aggregate values of the cohort, which continuous rejection scores have sufficient evidence to be directly reported, which require local workload and component annotation, and which should be reported with adjunct information? The goal of this work is not to introduce a new diagnostic category or change Banff categories, but to link each continuous score with the cohort workload, task stability, and lesion composition that define the way of reporting.

2 Materials and Methods

2.1 Study design and numerical material

The numerical data was derived from the continuous kidney allograft rejection indices by Vaulet et al. [7]. The calculation file included the following information: cohort sizes, diagnostic category counts, AUC values for discrimination tasks, index equations, and lesion ceiling contribution for three biopsy cohorts: 6272 derivation biopsies, 11,043 European validation biopsies, and 2185 United States validation biopsies. No individual biopsy records were accessed, generated, reconstructed, or simulated.

The data was organized in three linked tables before interpretation. First table quantified diagnostic workload using major rejection domains as cases per 100 biopsies. Second table characterized task behavior using phenotype frequency, mean AUC, AUC range, stable discrimination load, portability-floor score, and ancillary signal deficit. Third table quantified score composition using lesion ceiling shares in terms of entropy, effective lesion number, and dominant lesion share. This organization was based on the requirement of a pathology report to know not only whether the score is discriminatory, but also how frequently the score will be encountered and whether a high score is balanced or lesion-dominant.

Diagnostic categories were classified into reporting domains of direct clinical relevance. AMR/MVI domain contained the AMR, MVI, and probable AMR categories. Cellular-inflammation domain contained the TCMR, borderline TCMR, isolated intimal arteritis, and PVAN categories if the goal was to estimate tubulointerstitial inflammation workload. Mixed rejection was retained as separate category since coexistence of antibody-mediated and cellular processes changes the task. This classification was used not to redefine Banff diagnoses, but to quantify

the reporting pressure created by broad diagnostic domains.

2.2 Cohort workload and displacement

The diagnostic workload for cohort c and domain k was calculated as

$$L_{c,k} = 100\,\frac{n_{c,k}}{N_c}, \tag{1}$$

where $n_{c,k}$ is the number of biopsies in the domain and N_c is the number of biopsies in the cohort. This value represents the number of cases per 100 biopsies and may be directly observed using a reporting service. Domains that carry a load of 55 cases per 100 biopsies occur much more frequently than domains that carry a load of 5 cases per 100 biopsies, even if both have good discrimination.

The workloads for each domain were normalized in relation to the four reporting domains to assess the displacement within the cohort. Jensen-Shannon divergence was then determined by

$$\begin{aligned}\mathrm{JSD}(P_c, P_0) &= \frac{1}{2}D_{\mathrm{KL}}(P_c \parallel M) + \frac{1}{2}D_{\mathrm{KL}}(P_0 \parallel M),\\ M &= \frac{1}{2}(P_c + P_0).\end{aligned} \tag{2}$$

Here P_0 is the derivation vector and P_c is a validation vector. The ratio has finite values for the probability vectors and stays constant upon swapping the two vectors. Here, it represents the distance between the diagnostic load of the validation sample from that of the derivation sample.

For a counterpart load ratio, we have

$$\Lambda_{c,k} = \frac{L_{c,k}}{L_{0,k}}, \tag{3}$$

where $L_{0,k}$ represents the derivation work load for the same domain. This helps to determine the direction of the displacement. A higher value of TCMR/TI domain suggests a higher cell inflammation work load while a high value for mixed rejection shows a greater requirement for joint consideration instead of single axis evaluation.

2.3 Task retention and residual signal

The stable discrimination work load for each task was determined by

$$\mathrm{SDL}_t = 100\bar{p}_t(2\bar{A}_t - 1)(1 - R_t), \tag{4}$$

where $\bar{p}_t$ represents the average frequency of phenotypes, $\bar{A}_t$ represents the mean AUC among the cohort, and R_t is the range of AUC. The value of SDL will be high when the task is frequent, highly discriminated by the score, and consistent across cohorts. Otherwise, the values will be low due to the rarity of the phenotype, a low mean of AUC, and differences in discrimination from one validation environment to another. AUC was considered as a ranking-discrimination statistic.

Since mean AUC may hide the worst validation scenario, another measure named portability-floor was also estimated as:

$$\mathrm{PFS}_t = 100\bar{p}_t\{2\min_c(A_{c,t}) - 1\}. \tag{5}$$

This metric assesses how much task evidence persists with the lowest measured cohort AUC being set as the operating point. Hence, this metric is stricter compared to any performance measure averaged over time. High values of PFS indicate a wider applicability of reporting, whereas low values imply that verification is required for the task to be considered standard.

The residual signal in each case was captured by additional signal deficit:

$$\mathrm{ASD}_t = 100(1 - \bar{A}_t)(1 + R_t). \tag{6}$$

ASD is unweighted by frequency. How much diagnostic information could lie beyond the continuous histology score in the particular phenotype is being addressed. This becomes particularly relevant in case of etiologic diagnoses like PVAN, in which the presence of viruses, immunohistochemistry, and clinical virology play the deciding factor despite valuable information from histology.

2.4 Lesion composition and effective number

Lesion ceilings were transformed into proportions, $q_{jm} = Q_{jm}/100$, for index j and lesion constituent m. Lesion entropy is computed by

$$H_j = -\sum_m q_{jm}\log(q_{jm}), \tag{7}$$

and the effective lesion number was calculated as

$$E_j = \exp(H_j). \tag{8}$$

Effective lesion number is defined by the number of equal-weighted components that would lead to the

same entropy. A value close to the number of components suggests an even distribution, while a lower value signifies that the distribution is more skewed to fewer components. The application of entropy in this case follows the logic of effective diversities; the effective number is simpler to understand than entropy on its own [29, 30].

The proportion of dominant lesions was obtained from

$$D_j = 100 \max_m(q_{jm}). \tag{9}$$

The dominant proportion of this value indicates how much of the theoretical maximum of the score is contributed by the largest lesion component. High values for the dominant proportion do not necessarily invalidate the score since it only indicates that the total value must be reported along with the lesion components statement because one lesion component alone may significantly affect the theoretical maximum of the score.

2.5 Reporting assignment

Assignment of the reporting method used the calculated workload values without including any diagnostic category. Composite reporting was directly assigned when there was strong task retention and balanced lesion composition. Component-aware local reporting was assigned when there was strong discrimination but workload or lesion dominance could modify the interpretation of the score. Assignment of adjunct-evidence reporting occurred when the diagnostic phenotype included additional information other than the continuous histology score. The assignment process described the association between the score and reporting methods based on observed workload, task behavior, and lesion composition.

3 Results

3.1 Cohort workload

The three cohorts demonstrated substantial differences in their diagnostic workload. The derivation cohort demonstrated 73.50 no-rejection, 11.50 AMR/MVI-axis, 19.32 TCMR/TI including PVAN, and 1.45 mixed-rejection cases per 100 biopsies. The European validation cohort presented 61.02 no-rejection, 17.75 AMR/MVI, 21.51 TCMR/TI including PVAN, and 2.16 mixed-rejection workload. The United States validation cohort revealed 40.23 no-rejection, 55.47 TCMR/TI including PVAN, and 5.45 mixed-rejection cases per 100 biopsies.

The diagnostic workload Table 1 is the answer to the first part of the reporting question. Unlike the derivation cohort, the United States validation cohort is not a scaled replica of the former. In the latter, the no-rejection workload is reduced by about half while the TCMR/TI including PVAN workload is increased by nearly threefold. Consequently, a cellular inflammation score will be seen far more frequently in this setting, and each high-value score has to be interpreted within a broad differential of diagnoses that include classical TCMR, borderline inflammation, PVAN associated inflammation, isolated arteritis and mixed rejection.

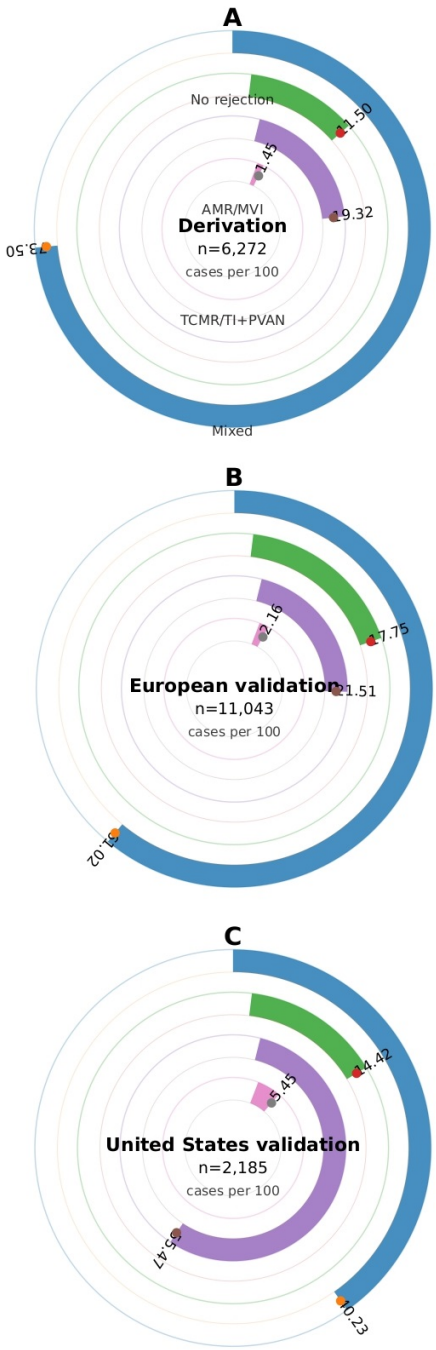
Table 1. Diagnostic workload per 100 biopsies.

Cohort	No rejection	AMR/MVI axis	TCMR/TI incl. PVAN	Mixed rejection
Derivation	73.50	11.50	19.32	1.45
European validation	61.02	17.75	21.51	2.16
United States validation	40.23	14.42	55.47	5.45

The workload fingerprints shown in Figure 1 provide this same information in a graphical format. The rings have been normalized independently to cases per 100 biopsies so that any expansion of one ring does not overshadow the rest. The workload fingerprint of the United States cohort has the largest redistribution of rings with TCMR/TI including PVAN expanding and the no-rejection ring shrinking. This particular visual evidence clearly makes it necessary to consider cellular-inflammation reporting a workload issue locally.

What is equally important is the separation between AMR/MVI and TCMR/TI workloads. AMR/MVI is a relatively more frequent occurrence in the European validation cohort compared to the derivation cohort; however, the overall displacement of the European cohort is small. On the other hand, the United States cohort shows a difference in the profile of the biopsy stream. Reporting is thus not only an issue of the increased frequency of a category. It is an issue that results in a score being used in a service where inflammatory lesions make up a bigger percentage of daily diagnostic workload.

The sheet of calculations in Figure 2 ensures transparency of the analysis. First, the band of calculations transforms cohort numbers into



Independent diagnostic-pressure rings; each ring is scaled to 100 biopsies.

Figure 1. Cohort workload fingerprints.

A Diagnostic-pressure calculation

$$L_{c,k} = 100 n_{c,k} / N_c$$

Cohort	N	No rej.	AMR/MVI	TCMR/TI+PVAN	Mixed
Derivation	6272	73.50	11.50	19.32	1.45
Europe	11043	61.02	17.75	21.51	2.16
United States	2185	40.23	14.42	55.47	5.45

major diagnostic axes; cases per 100 biopsies

B Task-portability calculation

$$PFS_t = 100 \bar{p}_t \{2 \min_c (A_{c,t}) - 1\} \quad ASD_t = 100(1 - \bar{A}_t)(1 + R_t)$$

Task	Mean freq	Mean AUC	PFS	ASD
No rejection / any other	58.25	0.973	53.59	2.75
Any rejection / no rejection	41.75	0.957	37.57	4.38
Borderline-TCMR / none	28.21	0.970	25.39	3.09
TCMR / no TCMR	14.98	0.987	14.38	1.35
AMR/MVI / neither	13.13	0.977	12.34	2.36
PVAN / no PVAN	2.91	0.823	1.75	18.37

selected rows from the portability table

C Index-composition calculation

$$q_{jm} = Q_{jm}/100 \quad H_j = -\sum q_{jm} \log(q_{jm}) \quad E_j = \exp(H_j)$$

Index	Comp.	H	E	Dominant %
AMR/MVI	4	1.380	3.98	28.1
TCMR/TI	4	1.196	3.31	46.3
Activity	6	1.783	5.95	17.6
Chronicity	4	1.332	3.79	40.0

entropy and dominant lesion share

Figure 2. Aggregate calculation sheet.

workload, the second one interprets the frequency of phenotypes and AUC behavior into task retention, and the third one interprets the proportions of lesion ceiling into component balance. Conclusions can thus be derived from the numbers displayed on the figure and not from the reconstructed patient-level data.

3.2 Cohort displacement

The profile of load ratios determined the direction of validation-cohort displacement. The European validation cohort had load ratios of 1.54 for AMR/MVI and 1.49 for mixed rejection and a ratio of 1.11 for TCMR/TI including PVAN, which was close to the derivation ratio. The validation cohort from the United States exhibited a different pattern. It had a load ratio of 0.55 for no-rejection, 2.87 for TCMR/TI including PVAN, and 3.76 for mixed rejection (Table 2).

Table 2. Cohort displacement relative to the derivation cohort.

Cohort	Jensen-Shannon divergence	Mean absolute log2 load shift	Largest load ratio
Derivation	0.0000	0.00	1.00
European validation	0.0094	0.41	1.54 for AMR/MVI
United States validation	0.0997	1.16	3.76 for Mixed rejection

The results of displacement analysis reveal that the validation cohort of the United States is a more stringent application site. The Jensen-Shannon divergence of 0.0997 is about ten times larger than the European one of 0.0094. The mean absolute log2 displacement of 1.16 for the United States is almost three times greater than the corresponding European

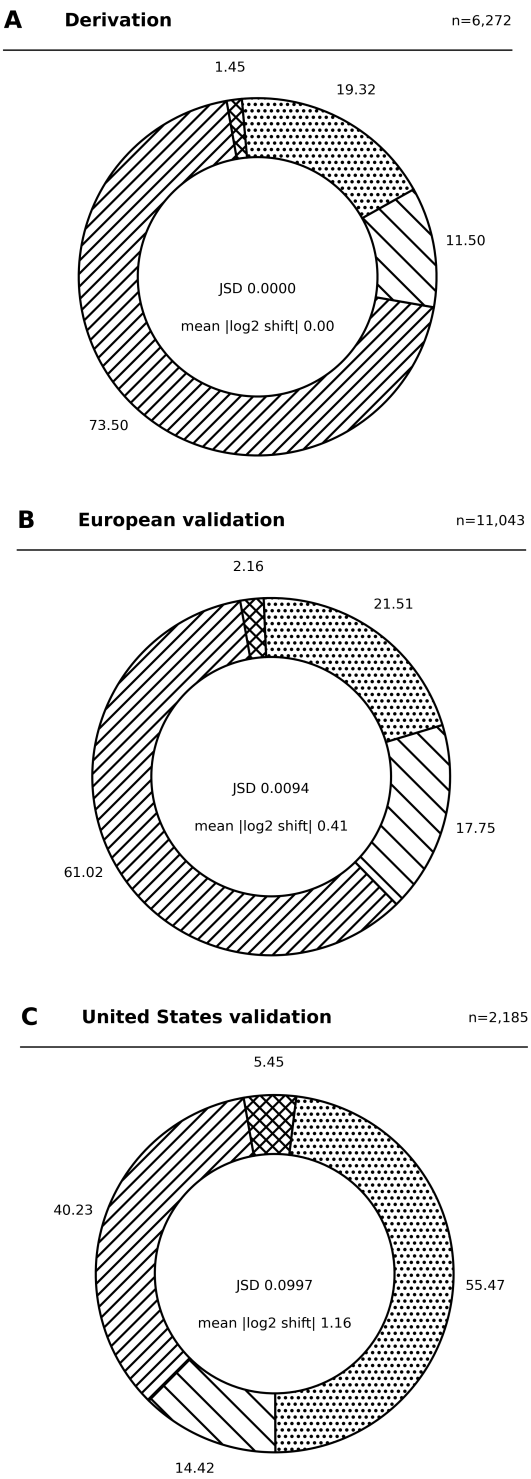


Figure 3. Axis displacement imprints.

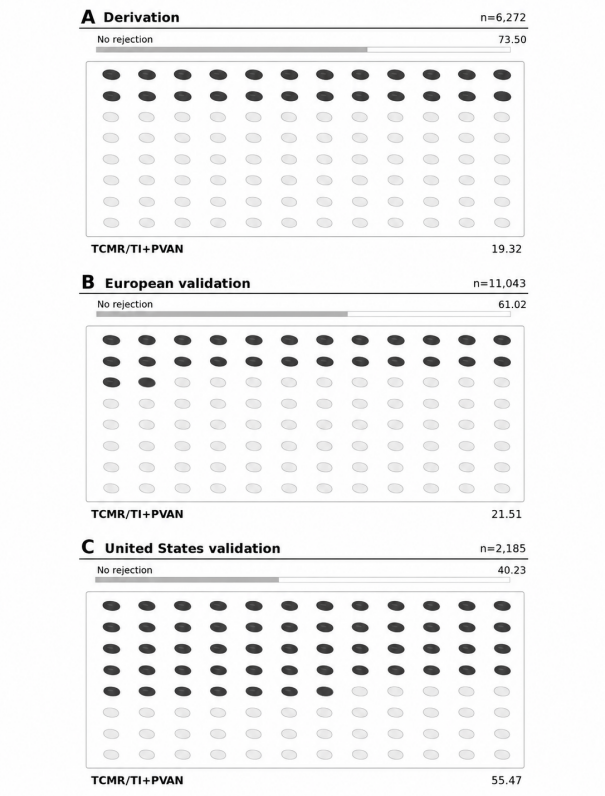


Figure 4. TCMR/TI workload field.

one of 0.41. Neither of these statistics suggests any deficiency in the scores in the United States validation cohort. Rather, it reveals that the context of interpretation in the new diagnostic environment differs from the derivation one, particularly with regard to cellular inflammation and mixed rejection.

The pressure imprints in Figure 3 illustrate the direction of displacement. While the European imprint is similar to the derivation pattern, the United States imprint moves towards TCMR/TI including PVAN and mixed rejection. The reason is that when applying the test in a displaced population, it needs to have more explanation than the test which was applied in a population with diagnostic work-load similar to the derivation sample.

The 100-biopsy fields in Figure 4 represent the key cohort transition. The derivation and European validation fields had similar TCMR/TI including PVAN density, while the United States field was very crowded. This is the numeric justification of the local-reporting principle: the high TCMR/TI value in a high-workload cohort should not be considered as a stand-alone number since the differential diagnosis of such a score is more complex and frequent.

3.3 Task retention and residual signal

Broad rejection-separation tasks showed the highest values of retained reporting. There were no rejections versus any other diagnosis with the mean phenotype frequency of 58.25 per 100 biopsies, the mean AUC of 0.973, the AUC range of 0.03, the SDL of 53.49, the PFS of 53.59, and the ASD of 2.75. Any rejections versus

no rejections had the mean phenotype frequency of 41.75, the mean AUC of 0.957, the AUC range of 0.01, the SDL of 37.75, the PFS of 37.57, and the ASD of 4.38. Such tasks are broad, frequent, and robust against the observed cohorts.

The rejection-specific tasks demonstrated more complicated reporting implications. There was a borderline TCMR/TCMR versus none with the mean phenotype frequency of 28.21, the mean AUC of 0.970, the SDL of 25.72, and the PFS of 25.39. It is the most robust rejection-specific task in terms of retained values. There was also a TCMR versus no TCMR with the lower phenotype frequency of 14.98 but the highest mean AUC among the listed tasks equal to 0.987. Thus, there were the SDL and PFS values of 14.43 and 14.38. AMR/MVI versus neither and probable AMR/MVI/AMR versus none retained good values with PFS of 12.34 and 13.39.

This difference makes it clear that it is impossible to assign one reporting status to all continuous scores. A task can have high AUC, but have low retained reporting weight because of a low frequency of the corresponding phenotype. For example, mixed rejection has the mean AUC 0.973, but it has the mean frequency 3.02 per 100 biopsies and the clinical meaning of the score is composite. Therefore, it is appropriate for specialist integrative interpretation, not for general language of reporting.

Comparison of broad tasks and specific tasks is of great clinical importance. No rejection vs any other diagnosis is a task of high retention because of its stability and frequency, but it is the reporting claim of low resolution because it shows that biopsy distinguishes from no-rejection state. Borderline TCMR/TCMR is a specific and actionable claim, but the interpretation of this task is dependent on the specifics of biopsy practice and borderline between inflammation, borderline TCMR, and other inflammatory states. Thus, the table of tasks does not estimate the importance of diagnoses, but the ability to make reporting claims with burden of evidence.

Figure 5 displays the retention orbit for frequency, PFS, AUC, and ASD in a single plot. The broader rejection tasks exhibit the largest retained orbits. Borderline TCMR/TCMR is the most specific rejection task, while PVAN is differentiated from the other rejection phenotypes by the radial ASD region. The differentiation does not imply that PVAN is unimportant. On the contrary, it reflects the fact that PVAN requires etiological information beyond what an inflammation-rejection score can provide.

The PVAN row presents the most interesting exception in the table. PVAN has the lowest PFS and the highest ASD values, implying that the continuous inflammation score has the greatest signal left per case of this phenotype. The finding makes clinical sense because PVAN is based on the viral infection and tissue damage as opposed to rejection pathology alone. Reporting the result as a problem of rejection score may be misleading with regard to the etiology of the disease. The result should be reported as a rejection score together with the viral evidence.

The PVAN evidence seal in Figure 6 provides clinical readability to the table result. The role of histology as a component of the diagnostic signal is maintained, but the ASD of 18.37 is significantly higher than those of the rejection tasks. Hence, in the case of a report with a continuously elevated inflammation score close to the PVAN, the report should provide the information on the viral evidence used for its interpretation: BK viral load, SV40 immunohistochemistry in case of its performance and the class of polyomavirus nephropathy.

3.4 Lesion composition

An analysis of lesion composition distinguished balanced scores from lesion-dominant scores. The AMR/MVI index consists of four components with lesion entropy of 1.380, effective lesion number of 3.98 and dominant lesion share of 28.1%. The effective lesion number is close to the number of four components which suggests that AMR/MVI is a balanced composite in terms of its theoretical ceiling. Thus, its direct reportability is justified provided that the report maintains the terminology of Banff AMR/MVI index and the connection with antibodies.

The activity index is the most balanced score of all. It consists of six components with entropy of 1.783, effective lesion number of 5.95 and dominant lesion share of 17.6%. The close-to-one effective lesion number is indicative of broad distribution of actively injured tissue and not one particular lesion which is the strongest composition-based support for the use of direct composite reporting.

TCMR/TI and chronicity indices needed an alternative interpretation. TCMR/TI index contains four components and its effective lesion number equals 3.31, while its dominant lesion share is 46.3%.

Table 3. Task-retention quantities.

Task	Mean frequency	Mean AUC	AUC range	SDL	PFS	ASD per 100 cases
No rejection vs any other	58.25	0.973	0.03	53.49	53.59	2.75
Any rejection vs No rejection	41.75	0.957	0.01	37.75	37.57	4.38
Borderline TCMR/TCMR vs none	28.21	0.970	0.03	25.72	25.39	3.09
TCMR vs No TCMR	14.98	0.987	0.01	14.43	14.38	1.35
Probable AMR/MVI/AMR vs none	14.55	0.960	0.00	13.39	13.39	4.00
AMR/MVI vs neither	13.13	0.977	0.01	12.39	12.34	2.36
AMR vs No AMR	8.37	0.970	0.02	7.71	7.70	3.06
Mixed rejection vs non-mixed	3.02	0.973	0.02	2.80	2.78	2.72
PVAN vs No PVAN	2.91	0.823	0.04	1.81	1.75	18.37

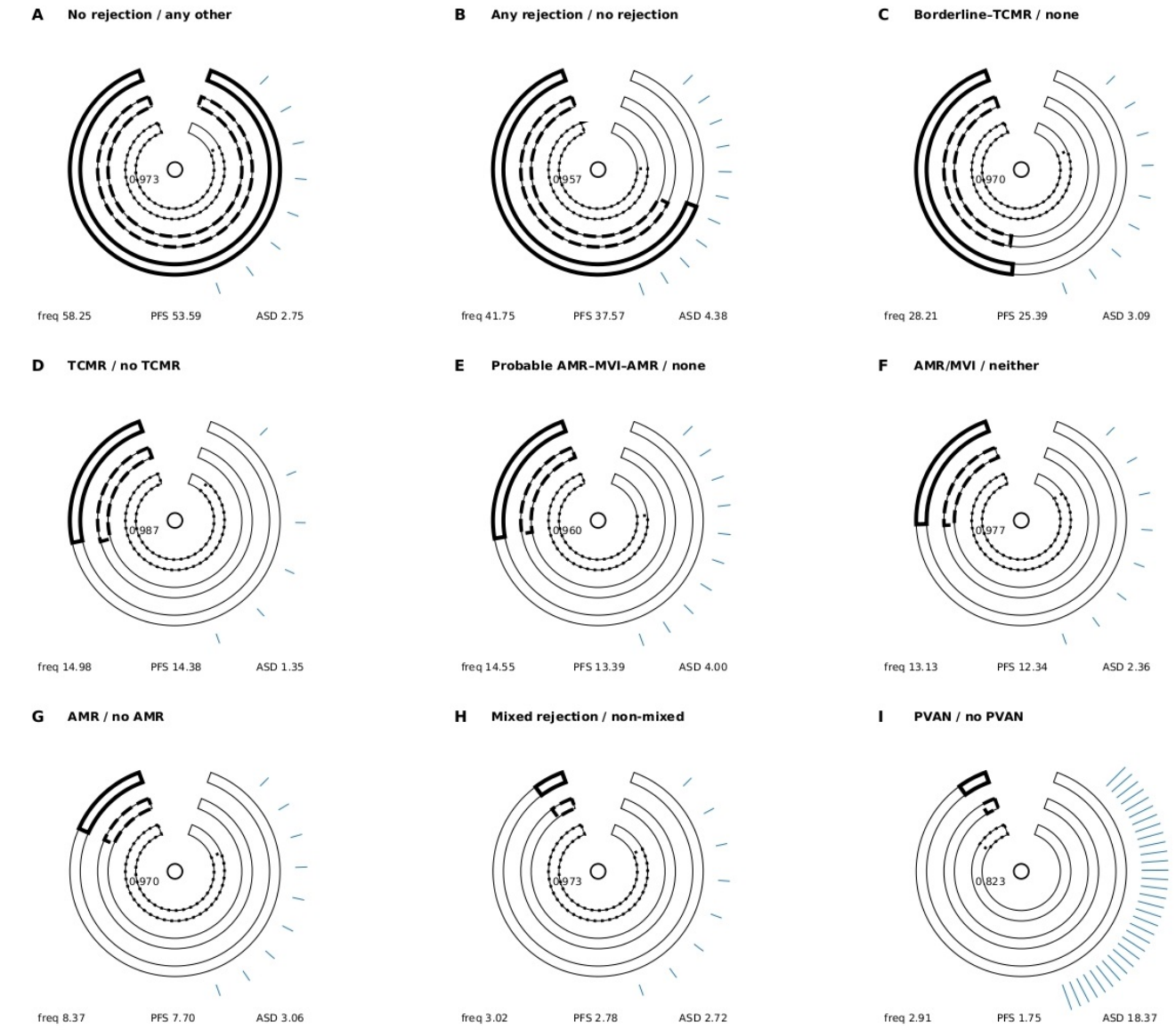


Figure 5. Task-retention orbit.

The chronicity index has an effective lesion number of 3.79 out of four components, however, chronic glomerulopathy is a component contributing to 40.0% of its theoretical ceiling (Table 4).

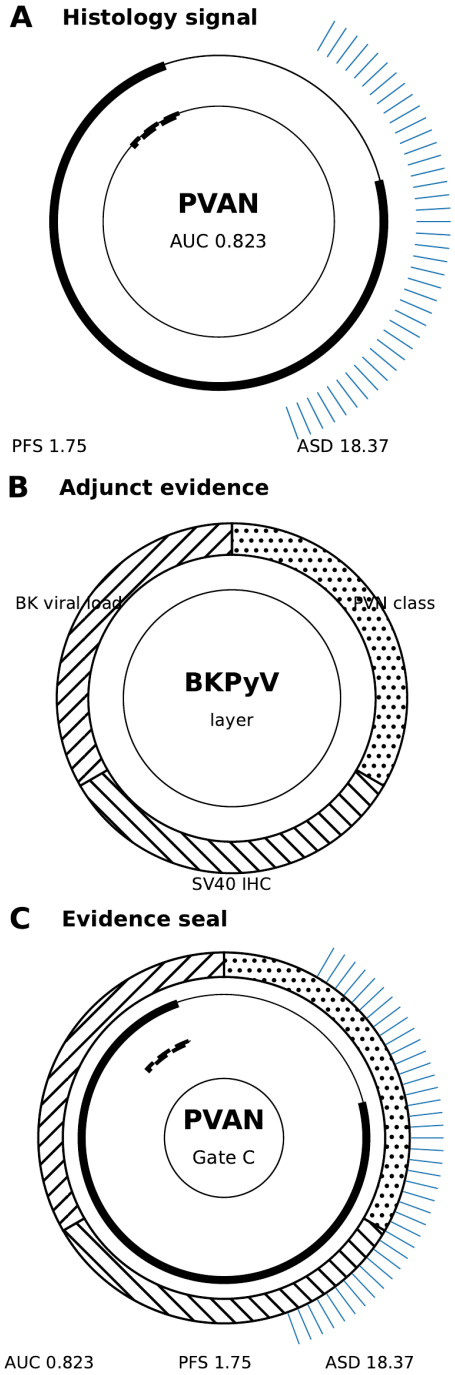


Figure 6. PVAN evidence seal.

Table 4. Lesion composition of continuous scores.

Index	Components	Lesion entropy	Effective lesion number	Dominant lesion share(%)
AMR/MVI	4	1.380	3.98	28.1
TCMR/TI	4	1.196	3.31	46.3
Activity	6	1.783	5.95	17.6
Chronicity	4	1.332	3.79	40.0

The lesion-composition table provides an answer to the second component of the research question.

AMR/MVI and activity have sufficient component balances such that they can be considered as composite scores. TCMR/TI and chronicity do not lose their values; however, the total value of these variables depends more heavily on the dominance of the component values. The reporting implication is clear: If the TCMR/TI value is high, then its name should state if the score is caused by arteritis, tubulitis, or interstitial inflammation, while if the chronicity value is high, then the cause should be chronic glomerulopathy.

In Figure 7, lesion mosaic converts ceiling shares to a tissue map image. While AMR/MVI and activity exhibit more spread out regions, TCMR/TI and chronicity display larger predominant regions. The image is very significant in daily reporting since it could be misleading for the clinician to know the total score without knowing which component produced it.

Entropy fingerprints of Figure 8 provide a second perspective on the same information about lesion balance. Component slots indicate how many lesions are involved in the score, retained filling indicates how many lesions are actually represented, and the dominant-share blade is the point where a single lesion makes a significant contribution to the ceiling effect. Thus, the panels of TCMR/TI and chronicity scores also advocate for the adoption of component-aware reporting despite the clinical utility of the continuous scores.

The balanced AMR/MVI and activity information helps determine the concept of a direct composite score. The term "direct composite" does not imply diagnostic independence; it implies that the overall score can be used as an abbreviated indication of disease severity as long as there is no significant ceiling effect driven by a single lesion, and the task is associated with adequate evidence across different cohorts. However, the report should still contain the diagnostic class, the pattern of component lesions, and the necessary supporting evidence of the Banff criteria.

Conversely, the TCMR/TI and chronicity scores define the other side of the issue. These scores are not considered invalid, as they have a significant number of lesions and reflect a relevant burden of injury. Their deficiency is interpretive concentration: a high value of TCMR/TI could be mainly due to intimal arteritis, while a high chronicity value could mainly depend on chronic glomerulopathy. In both cases, the total score

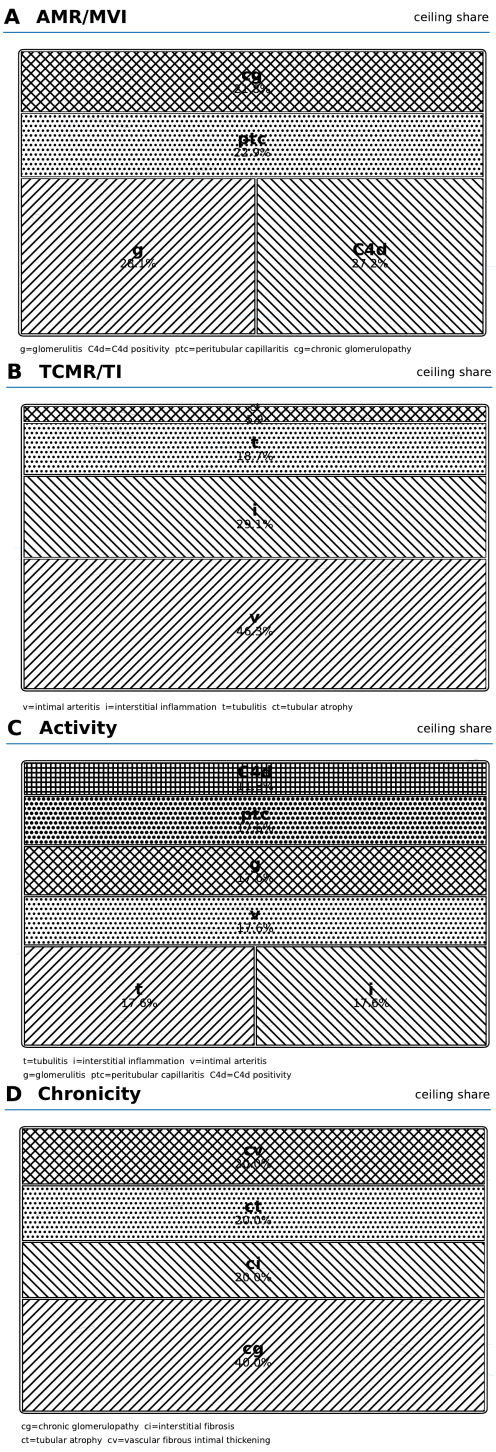


Figure 7. Lesion-ceiling mosaic.

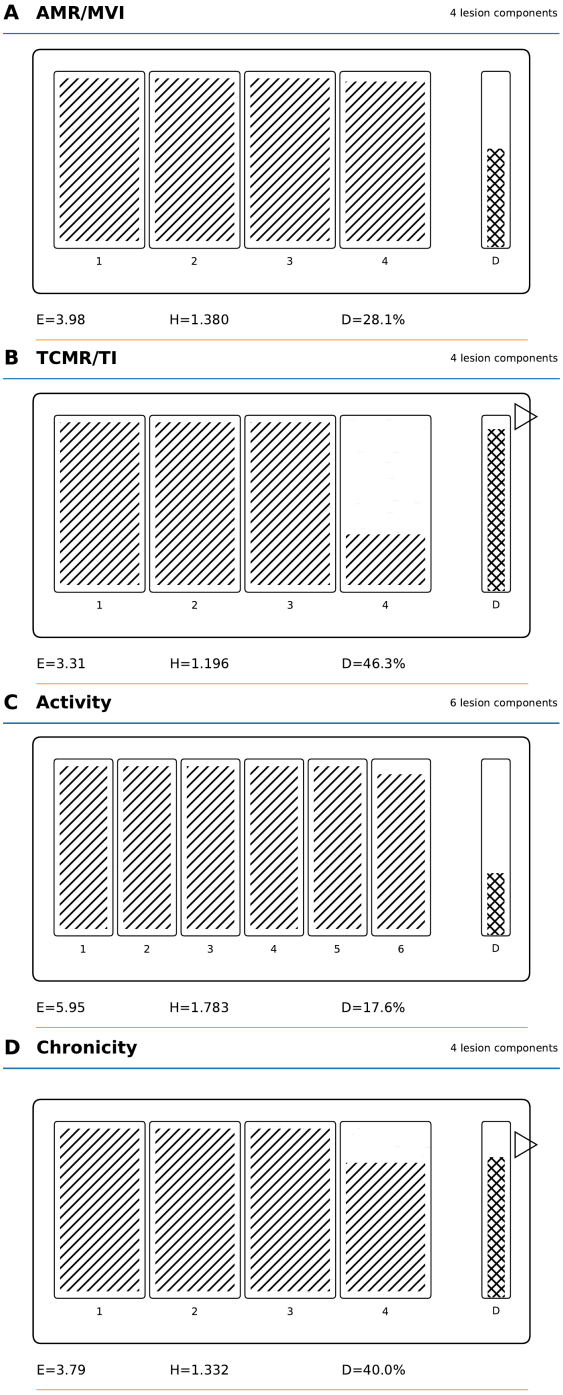


Figure 8. Entropy fingerprints.

becomes more informative when combined with a brief statement about the components.

3.5 Reporting assignment

The final reporting assignment assigned each score into the reporting mode appropriate for the evidence behind the score. AMR/MVI score became direct composite reporter as it had excellent AMR/MVI task retention and balanced lesion structure. Activity became direct composite reporter due to its lesion effective number being 5.95 out of six lesions and dominant lesion share being 17.6%.

TCMR/TI became local component-aware reporter. The assignment is not based on poor discrimination, as the TCMR task had mean AUC of 0.987 and PFS of 14.38. The assignment was done based on clinical utility sensitivity, as United States validation cohort

had TCMR/TI including PVAN load ratio of 2.87, and TCMR/TI score had dominant lesion share of 46.3%. Chronicity became local component-aware reporter due to chronic glomerulopathy making up 40.0% of the ceiling.

PVAN became adjunct-evidence reporter. The PFS of PVAN was 1.75 and ASD of PVAN was 18.37. This does not diminish the significance of PVAN recognition on biopsy, as this result means that PVAN should not be reported as if a rejection-inflammatory score was an etiologic classifier. Mixed rejection became integrative review reporter due to low mean frequency, but high United States load ratio of 3.76 (Table 5).

Table 5. Reporting assignment of scores and phenotypes.

Score or phenotype	Quantitative basis	Reporting role	Report language
AMR/MVI	PFS 12.34; effective lesion number 3.98	Direct composite	Report as a quantitative microvascular injury summary with Banff and antibody context.
Activity	Effective lesion number 5.95; dominant lesion share 17.6%	Direct composite	Report as an overall activity score while preserving lesion-level Banff interpretation.
TCMR/TI	United States TCMR/TI load ratio 2.87; dominant lesion share 46.3%	Local component-aware	State local workload context and name the dominant cellular lesion pattern.
Borderline TCMR/TCMR	SDL 25.72; PFS 25.39	Local component-aware	Use as a high-priority cellular-inflammation task with local threshold awareness.
Chronicity	Dominant chronic-glomerulopathy share 40.0%	Local component-aware	State whether chronicity is broad or driven by glomerular chronic injury.
Mixed rejection	Mean frequency 3.02; United States load ratio 3.76	Integrative review	Interpret through AMR and TCMR components rather than a single-axis label.
PVAN	PFS 1.75; ASD 18.37	Adjunct evidence required	Require BK viral evidence, SV40 staining when performed, and PVN class.

These tables and figures create a series of decision-making processes for reporting. The workload table defines the frequency of occurrence for each diagnostic domain; the displacement table confirms whether the validation setting is similar to that of derivation; the task-retention table identifies the discrimination tasks with adequate evidence; the lesion-composition table determines the balance of the total score; and the reporting-assignment table converts the numerical results into report language. Each graphic figure provides a compact test for the corresponding process. This series of analyses maintains the focus on reporting adequacy rather than on creation of a new scoring system.

The reporting-assignment table provides a direct answer to the research question. There is no single reporting status for a numerical score family. Direct composite language should be used if the task is retained and the score is compositionally balanced. Local component-aware language is indicated if the score is discriminative while the cohort workload or lesion dominance is variable. Adjunct evidence language should be used when the phenotype is etiologic rather than histologic.

4 Discussion

The results show that continuous rejection scores must be adopted in biopsy reports as reporting claims are defined for every score. The main message is not that continuous scores are inaccurate. The main message is that their reporting claims are different depending on the diagnostic task. AMR/MVI and activity qualify as balanced composite scores with strong retained use. TCMR/TI and chronicity have diagnostic potential, but it is conditional on local biopsy workloads and lesion patterns. PVAN belongs to other categories of score reporting because viral evidence is crucial for its diagnosis.

The results clarify the role of AUC in transplant pathology reporting. AUC is an important concept, but not a sole justification for reporting scores. High AUC for TCMR versus no TCMR cannot guarantee that reporting sentence for the score will be the same in the centers where TCMR/TI including PVAN comprises more than half of the workload. Likewise, high AUC for mixed rejection cannot transform it into general screening tool for the center where the mean frequency is only 3.02 per 100 biopsies. The score must be evaluated based on the task performed by it.

The results concerning the influence of workload on the cohort validation become especially important for the practical application of the scores. While the European validation cohort had almost the same diagnostic composition as the derivation workload, the US cohort was quite different. Therefore, portability of the score cannot be measured by only one label. A score can be validated statistically, but still require different wording in the report according to the workload of the center. Thus, transplant pathology services considering continuous scores should estimate their no-rejection, AMR/MVI, TCMR/TI including PVAN, mixed-rejection, and PVAN workloads before deciding on the report wording.

The AMR/MVI result favors cautious direct reporting of the score. The score has the balanced lesion composition and strong task retention, making it a reasonable quantitative representation of microvascular injury. Nonetheless, the direct reporting does not imply the independent reporting. The score must be connected with Banff AMR/MVI

criteria, donor specific antibodies assessment, C4d status, transcriptomic data or clinical evidence in case it is included in the diagnostic process. Such interpretation is in line with the current interest of Banff in MVI, DSA-negative states and MVI phenotypes beyond classical AMR [4, 8]. Also, the idea of standardized antibody evidence in transplant diagnostics [31] fits the current approach.

The TCMR/TI result requires a bit more caution. The score has strong discriminative power, however, the US workload and dominant lesion change the interpretation of the score. High TCMR/TI score cannot serve as a standalone number because it can indicate arteritis, tubulitis, interstitial inflammation, inflammation in scarred cortex, viral inflammation or mixed disease. Hence, report should specify the lesion pattern and, if needed, explain that local biopsy flow is enriched with diseases associated with cellular inflammation.

Chronicity score also requires component annotation. The lesion composition was high, however, 40.0% chronic glomerulopathy implies that high score can be driven by the glomerular chronic lesion. It is important because chronic glomerulopathy can provide different history of the disease than interstitial fibrosis and tubular atrophy in general. The studies on the inflammation in the scarred area support the necessity of component-level information even with total scores [11, 12]. Also, the digital quantification of chronic and inflammatory lesions highlights the same issue [24].

The clear boundaries of continuous rejection score reporting are provided by PVAN. While the high ASD does not mean the lack of utility of the histology, it implies that the rejected hypothesis is narrower. A rejection-inflammatory score is not sufficient to classify PVAN [13, 14]. Also, the type of PVN and clinical context of the disease remain necessary [15, 16]. In this way, the report is protected against misinterpretation of the viral inflammation as alloimmune rejection.

From the practical perspective, the results provide recommendations about the adoption of the quantitative histology. Before adding continuous score to the report, pathology service should decide on its role: direct composite score, local component-aware score or etiological phenotype requiring additional evidence. The direct composite score can be reported briefly. The local component-aware score requires lesion comment. Etiological phenotype requires the report of laboratory or immunohistochemical evidence. The staged approach maintains the benefits of continuous score reporting and keeps it clinically interpretable.

The major limitation of the study is that cohort-level data are used rather than individual-level. It prevents recalibration, subgroup or outcome modeling, and center-specific threshold estimation. However, this limitation is transformed into the advantage: all conclusions of the study are based solely on the cohort-level values (number of patients and biopsies, AUC values, lesion ceiling shares and derived measures shown in the tables). Thus, the results can be considered as reporting recommendations and not as the basis for the treatment algorithm or alternative to the Banff diagnostic categories.

The secondary limitation is that local practice can be different from three cohorts used in the analysis. The protocol biopsy use, biopsy indication, immunosuppressive regimen, viral monitoring, DSA testing and molecular diagnostics can all change the composition of findings in a pathology service. Therefore, the study does not suggest any fixed reporting sentence for every score. Instead, it identifies which parts of the report should change in case of different workload: local context for TCMR/TI and mixed rejection, component annotation for TCMR/TI and chronicity, and adjunct evidence for PVAN.

The overall conclusion is that continuous rejection scores can improve, but not confuse transplant biopsy reporting. The score should not be considered as the replacement for diagnostic reasoning. It should be quantitative addition to the report only in the case of the supported reporting claim. This approach corresponds to the Banff practice, antibody and transcript diagnostics, BK virus guidance and digital pathology reports.

The clinically important advantage of the assignment of scores is that it separates the score validity and the score reporting sentence. TCMR/TI is not moved into the local component-aware reporting not because it is poor, but because a good score can be misinterpreted in the case of high local workload and dominant component. PVAN is not moved into the adjunct evidence reporting not because the histology is irrelevant, but because viral etiology cannot be estimated with the help of the rejection-inflammatory

score. In such a way, two common mistakes are avoided: discarding useful scores for the sake of universality, and use of useful scores beyond the claim.

The conclusion is also based on the numerical hierarchy, not on the narrative preferences. Direct reporting scores have both task or composition support. Local reporting scores have strong use, but require additional information. Adjunct evidence phenotype has the etiological signal that remains outside the score. This hierarchy provides a practical way for transplant service to report quantitatively without losing diagnostic accuracy of Banff categories.

5 Conclusion

The major research question was which continuous kidney allograft rejection scores could be reported directly in a report, which scores required local interpretation of individual components' involvement and which scores needed an additional proof outside the scope of the number. The response to the question is very specific. In case of AMR/MVI and activity scores there is enough transparency in the task retention and balance between lesions to present composite scores. At the same time in case of TCMR/TI, chronicity scores, borderline TCMR/TCMR situations and mixed cases of rejection the total numbers cannot be presented. They require knowledge about the local workload context and explanations of individual component contribution. In case of PVAN it is not possible to interpret score based on the continuous inflammatory metric alone. It is necessary to have an interpretation of the score confirmed by viral proof and PVN classification. To summarize, this conclusion serves as a reporting guide rather than a diagnostic category. A continuous score should be reported only if the wording of the score corresponds to the workload of the cohort, retained evidence of the task and lesion composition which is reflected in the number.

References

[1] Roufosse, C., Simmonds, N., Clahsen-van Groningen, M., Haas, M., Henriksen, K. J., Horsfield, C., ... & Becker, J. U. (2018). A 2018 reference guide to the Banff classification of renal allograft pathology. *Transplantation, 102*(11), 1795-1814.

[2] Jeong, H. J. (2020). Diagnosis of renal transplant rejection: Banff classification and beyond. *Kidney research and clinical practice, 39*(1), 17.

[3] Mengel, M., Mannon, R. B., & Feng, S. (2024). The Banff Process—reloaded: a joint initiative from the Banff Foundation for Allograft Pathology and the American Journal of Transplantation. *American Journal of Transplantation, 24*(3), 325-327.

[4] Naesens, M., Roufosse, C., Haas, M., Lefaucheur, C., Mannon, R. B., Adam, B. A., ... & Mengel, M. (2024). The Banff 2022 Kidney Meeting Report: reappraisal of microvascular inflammation and the role of biopsy-based transcript diagnostics. *American Journal of Transplantation, 24*(3), 338-349.

[5] Roufosse, C., Naesens, M., Haas, M., Lefaucheur, C., Mannon, R. B., Afrouzian, M., ... & Mengel, M. (2024). The Banff 2022 Kidney Meeting Work Plan: data-driven refinement of the Banff Classification for renal allografts. *American journal of transplantation, 24*(3), 350-361.

[6] Naesens, M., Roufosse, C., Cornell, L. D., Haas, M., Mannon, R. B., Afrouzian, M., ... & Mengel, M. (2026). The Banff 2024 Kidney Meeting Report: Rejection as a spectrum of phenotypes and focus on differential diagnostic reasoning. *American Journal of Transplantation*.

[7] Vaulet, T., Koshy, P., Wellekens, K., Aubert, O., Bottomley, C., Callemeyn, J., ... & Naesens, M. (2025). Continuous indices to assess the phenotypic spectrum of kidney transplant rejection. *Nature Communications, 16*(1), 10417.

[8] Varol, H., Wagenmakers, A., Hoeft, K., Callemeyn, J., Bodewes, R., Bramer, W., ... & Clahsen-Van Groningen, M. C. (2025). Expanding the scope of microvascular inflammation: unveiling its presence beyond antibody-mediated rejection into T-cell mediated contexts. *Transplant International, 37*, 13464.

[9] Piñeiro, G. J., Montagud-Marrahi, E., Ríos, J., Ventura-Aguiar, P., Cucchiari, D., Revuelta, I., ... & Diekmann, F. (2021). Influence of persistent inflammation in follow-up biopsies after antibody-mediated rejection in kidney transplantation. *Frontiers in medicine, 8*, 761919.

[10] Haas, M., Loupy, A., Lefaucheur, C., Roufosse, C., Glotz, D., Seron, D. A., ... & Mengel, M. (2018). 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.

[11] Lefaucheur, C., Gosset, C., Rabant, M., Viglietti, D., Verine, J., Aubert, O., ... & Loupy, A. (2018). T cell–mediated rejection is a major determinant of inflammation in scarred areas in kidney allografts. *American Journal of Transplantation, 18*(2), 377-390.

[12] Nankivell, B. J., Shingde, M., Keung, K. L., Fung, C. L. S., Borrows, R. J., O'Connell, P. J., & Chapman, J. R. (2018). The causes, significance and consequences of

inflammatory fibrosis in kidney transplantation: the Banff i-IFTA lesion. *American journal of transplantation, 18*(2), 364-376.

[13] Nickeleit, V., Singh, H. K., Randhawa, P., Drachenberg, C. B., Bhatnagar, R., Bracamonte, E., ... & Seshan, S. V. (2018). The Banff Working Group classification of definitive polyomavirus nephropathy: morphologic definitions and clinical correlations. *Journal of the American Society of Nephrology, 29*(2), 680-693.

[14] Kowalewska, J., El Moudden, I., Perkowska-Ptasinska, A., Kapp, M. E., Fogo, A. B., Lin, M. Y., ... & McCune, T. R. (2021). Assessment of the Banff Working Group classification of definitive BK polyomavirus nephropathy. *Transplant International, 34*(11), 2286-2296.

[15] Kotton, C. N., Kamar, N., Wojciechowski, D., Eder, M., Hopfer, H., Randhawa, P., ... & Transplantation Society International BK Polyomavirus Consensus Group. (2024). The second international consensus guidelines on the management of BK polyomavirus in kidney transplantation. *Transplantation, 108*(9), 1834-1866.

[16] Al-Talib, M., Welberry-Smith, M., Macdonald, A., & Griffin, S. (2025). BK Polyomavirus-associated nephropathy–diagnostic and treatment standard. *Nephrology Dialysis Transplantation, 40*(4), 651-661.

[17] Tambur, A. R., Campbell, P., Chong, A. S., Feng, S., Ford, M. L., Gebel, H., ... & Nickerson, P. (2020). Sensitization in transplantation: assessment of risk (STAR) 2019 Working Group Meeting Report. *American Journal of Transplantation, 20*(10), 2652-2668.

[18] Lefaucheur, C., Louis, K., Morris, A. B., Taupin, J. L., Nickerson, P., Tambur, A. R., ... & Levitsky, J. (2023). Clinical recommendations for posttransplant assessment of anti–HLA (Human Leukocyte Antigen) donor-specific antibodies: A Sensitization in Transplantation: Assessment of Risk consensus document. *American Journal of Transplantation, 23*(1), 115-132.

[19] Toulza, F., Dominy, K., Willicombe, M., Beadle, J., Santos, E., Cook, H. T., ... & Roufosse, C. (2022). Diagnostic application of transcripts associated with antibody-mediated rejection in kidney transplant biopsies. *Nephrology Dialysis Transplantation, 37*(8), 1576-1584.

[20] Beadle, J., Papadaki, A., Toulza, F., Santos, E., Willicombe, M., McLean, A., ... & Roufosse, C. (2023). Application of the Banff Human Organ Transplant Panel to kidney transplant biopsies with features suspicious for antibody-mediated rejection. *Kidney International, 104*(3), 526-541.

[21] Halloran, P. F., Reeve, J. P., Pereira, A. B., Hidalgo, L. G., & Famulski, K. S. (2014). Antibody-mediated rejection, T cell–mediated rejection, and the injury-repair response: new insights from the Genome Canada studies of kidney transplant biopsies. *Kidney international, 85*(2), 258-264.

[22] Zhang, H., Haun, R. S., Collin, F., Cassol, C., Napier, J. O., Wilson, J., ... & Coley, S. M. (2024). Development and validation of a multiclass model defining molecular archetypes of kidney transplant rejection: a large cohort study of the Banff Human Organ Transplant Gene Expression Panel. *Laboratory Investigation, 104*(3), 100304.

[23] Farris, A. B., Alexander, M. P., Balis, U. G., Barisoni, L., Boor, P., Bülow, R. D., ... & Solez, K. (2023). Banff Digital Pathology Working Group: image bank, artificial intelligence algorithm, and challenge trial developments.

[24] Hermsen, M., Ciompi, F., Adefidipe, A., Denic, A., Dendooven, A., Smith, B. H., ... & van der Laak, J. A. (2022). Convolutional neural networks for the evaluation of chronic and inflammatory lesions in kidney transplant biopsies. *The American Journal of Pathology, 192*(10), 1418-1432.

[25] Iwadoh, K., & Tonsho, M. (2025). Prospects for Artificial Intelligence-Based Pathological Diagnosis of Renal Transplant Biopsy. Nephron, 149(Suppl. 1), 45-51.

[26] Collins, G. S., Moons, K. G., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., ... & Logullo, P. (2024). TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. bmj, 385.

[27] Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., Steyerberg, E. W., & Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative Bossuyt Patrick Collins Gary S. Macaskill Petra McLernon David J. Moons Karel GM Steyerberg Ewout W. Van Calster Ben van Smeden Maarten Vickers Andrew J. (2019). Calibration: the Achilles heel of predictive analytics. *BMC medicine, 17*(1), 230.

[28] Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine learning, 77*(1), 103-123.

[29] Endres, D. M., & Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information theory, 49*(7), 1858-1860.

[30] Jost, L. (2006). Entropy and diversity. *Oikos, 113*(2), 363-375.

[31] Tambur, A. R., Bestard, O., Campbell, P., Chong, A. S., Crespo, M., Ford, M. L., ... & Nickerson, P. (2023). Sensitization in transplantation: assessment of risk 2022 working group meeting report. *American Journal of Transplantation, 23*(1), 133-149.