Full title: Data-driven derivation and validation of novel phenotypes for acute kidney transplant rejection using semi-supervised clustering

Running title: Data-driven transplant phenotypes

Thibaut Vaulet[1], Gillian Divard[2], Olivier Thaunat[3,4], Evelyne Lerut[5], Aleksandar Senev[6,7], Olivier Aubert[2], Elisabet Van Loon[6], Jasper Callemeyn[6], Marie-Paule Emonds[6,7], Amaryllis Van Craenenbroeck[6,8], Katrien De Vusser[6,8], Ben Sprangers[6,8], Maud Rabeyrin[9], Valerie Dubois[10], Dirk Kuypers[6,8], Maarten De Vos[1,11], Alexandre Loupy[2], Bart De Moor[1], Maarten Naesens[6,8]

[1] ESAT Stadius Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Leuven, Belgium

[2] Université de Paris, INSERM, Paris Translational Research Centre for Organ Transplantation, Paris, France; Kidney Transplant Department, Necker Hospital, Assistance Publique - Hôpitaux de Paris, Paris, France.

[3] French National Institute of Health and Medical Research (Inserm) Unit 1111, Lyon, France

[4] Hospices Civils de Lyon, Edouard Herriot Hospital, Department of Transplantation, Nephrology and Clinical Immunology, Lyon, France

[5] Department of Imaging & Pathology, University Hospitals Leuven, Leuven, Belgium

[6] Department of Microbiology, Immunology and Transplantation, KU Leuven, Leuven, Belgium

[7] Histocompatibility and Immunogenetics Laboratory, Belgian Red Cross—Flanders, Mechelen, Belgium

[8] Department of Nephrology and Kidney Transplantation, University Hospitals Leuven, Leuven, Belgium

[9] Hospices Civils de Lyon, Department of Pathology, Bron, France

[10] French National Blood Service (EFS), HLA Laboratory, Décines-Charpieu, France

[11] Department of Development and Regeneration, University Hospitals Leuven, KU Leuven, Leuven, Belgium

WORDS counts:

- ⇨ **Abstract = 250**
- ⇨ **Main text = 3489 (without methods)**

Correspondence to:

Prof Maarten Naesens, Department of Microbiology, Immunology and Transplantation, KU Leuven, University of Leuven, Belgium. Tel: +32 16 34 45 80; maarten.naesens@kuleuven.be

Keywords : Kidney transplant, acute rejection, Banff, biopsies, machine learning, consensus clustering

ABSTRACT

Background

Over the past decades, an international group of experts iteratively developed a consensus classification of kidney transplant rejection phenotypes, known as the Banff classification. We postulated that data-driven clustering of kidney transplant histological data could simplify the complex and discretionary rules of the Banff classification, while improving the association with graft failure.

Methods

The data consisted of a training set of 3510 kidney transplant biopsies from an observational cohort of 936 recipients. Independent validation of the results was performed on an external set of 3835 biopsies from 1989 patients. Based on acute histological lesion scores and the presence of donor-specific HLA antibodies, stable clustering was achieved based on a consensus of 400 different clustering partitions. Additional information on kidney transplant failure was introduced with a weighted Euclidean distance.

Results

Based upon the proportion of ambiguous clustering, 6 clinically meaningful cluster phenotypes were identified. There was significant overlap with the existing Banff classification (adjusted rand index = 0.48). However, the data-driven approach eliminated intermediate and mixed phenotypes and created acute rejection clusters that each significantly associates with graft failure. Finally, we developed and validated a novel visualization tool to present disease phenotypes and severity in a continuous manner, as complement to the discrete clusters.

Conclusion

We have developed and validated a semi-supervised clustering approach for the identification of clinically meaningful novel phenotypes of kidney transplant rejection. Our approach has the potential to offer a more quantitative evaluation of rejection subtypes and severity, especially in situations where the current histological categorization is ambiguous.

ABBREVIATIONS

ABMR: antibody-mediated rejection

ARI: adjusted rand index

AUROC: area under the ROC curve

HLA-DSA: human leukocyte antigen – donor specific antibody

HR: hazard ratio

k: number of clusters

PAC: proportion of ambiguous clustering

RMST: restricted mean survival time

TCMR: T-cell mediated rejection

Glossary Table

Semi-supervised clustering: intermediate form of data analysis, between the unsupervised and supervised learning frameworks. While being mostly driven by the data patterns (unsupervised learning), external variables provide additional (weak) supervision to guide the clustering.

Consensus clustering: clustering framework where the data is clustered in many partitions based on various initial conditions. The final partition is obtained *via* a consensus function that summarizes the whole set of partitions (e.g. majority voting)

PAC (proportion of ambiguous clustering): In a consensus clustering setting, PAC measures the proportion of pairs of datapoints that demonstrate inconsistent clusters attributions over the set of partitions.

k-mean: clustering algorithm which attempts to partition the data into k clusters based on the nearest centroid (cluster center).

weighted Euclidean distance: modified version of the traditional Euclidean distance where variables are weighted to put more or less emphasis on certain features of the data.

PCA (principal component analysis): unsupervised dimensionality reduction method based on data covariance. PCA is often used to project data into a two-dimensional space for visualization purpose.

kNN (k-nearest neighbours): non-parametric algorithm used for supervised learning. Provided a meaningful distance metric, this algorithm uses the k-nearest instances to a query datapoint to perform a prediction, typically through averaging or majority voting.

ARI (adjusted rand index): metric used to assess the degree of similarity of two different partitions from the same data, adjusted for random permutations. A score of 1 indicates a perfect overlap of the two partitions, whereas a score of 0 indicates a random partitioning.

INTRODUCTION

Kidney transplant biopsies are crucial in the follow-up of patients after transplantation. Both at time of graft dysfunction (indication biopsies) and at time of stable graft function (protocol biopsies), the histological evaluation of these biopsies enables to distinguish rejection mechanisms from other injury processes and to orient the appropriate therapeutic interventions. Over the past decades, an international group of experts has developed a consensus classification of kidney transplant rejection phenotypes, known as the Banff classification.[1–3]

The Banff classification relies on the histological evaluation of a set of well-defined lesions, further translated into semiquantitative, ordinal lesion scores.[4] The diagnostic classification process consists of a set of *if-then* rules that map conditional clauses based on lesion scores to a diagnosis category. Currently, the Banff classification encompasses 5 main categories[3] : (1) Normal biopsy or non-specific changes; (2) Antibody-mediated changes (ABMR); (3) Borderline changes; (4) T-cell mediated rejection (TCMR); and (5) Polyomavirus nephropathy. Several of these categories are further subdivided into subtypes. This classification was developed iteratively, based on studies that examine the associations between lesions and risk factors like donor-specific HLA antibodies, between lesions and graft failure, and among lesions themselves.[5–7] Banff diagnostic categories are not mutually exclusive and Banff lesions are not specific for disease processes, which leads to overlapping diagnoses and mixed rejection phenotypes. Although this reflects a histological reality, the clinical interpretation of this complex categorization process is difficult, leading to unstable clinical decisions.

Instead of this iterative consensus process for disease classification, data-driven mathematical modelling of the multidimensional histological data could be appropriate. Such approach could refine the thresholds for the diagnostic phenotypes, simplify the complex and discretionary *if-then* rules, avoid the issue of mixed phenotypes, and yield new phenotypes and disease reclassification. Categorizing data into groups without pre-existing labels is commonly referred as unsupervised clustering.[8] Although the

resulting clusters (reclassified disease phenotypes) might be valid from a mathematical perspective, there is no guarantee that they will show relevant association with external outcome variables. To overcome this, introducing information on outcome in the clustering process could be of interest. [9–11] Whether such mathematical modelling approach would also be applicable to the classification of kidney transplant rejection, has not been evaluated yet.

On the basis of these considerations, we built and externally validated a model for mathematical reclassification of acute kidney transplant rejection, based on the integration of the set of inflammatory lesions in kidney transplant biopsies, informed by graft failure, in a retrospective observational cohort study.

METHODS

Data

Patients and biopsies

For the training cohort, all consecutive adult recipients of a kidney transplant at the University Hospitals-Leuven between March 2004, the start of the protocol biopsy program, and February 2013 were eligible for this study (n=1137). A minimal of 5 years follow-up at time of data extraction (March 2018) was required. Recipients of combined transplantation (n=113) or kidney transplantation after another solid organ transplantation (n=24) were excluded. All transplants were performed with negative complement-dependent cytotoxicity crossmatches. The clinical data were collected during routine clinical follow-up in electronic medical records, which were used for clinical patient management and directly linked to the SAS database from which the research database was extracted. The standard immunosuppressive maintenance regimen consisted of tacrolimus, mycophenolate and corticosteroids[12]. The histological data consisted of all 3622 kidney transplant biopsies performed at the Leuven University Hospitals between April 2004 and February 2015 in 949 patients. Biopsies were performed upon medical indication (indication biopsies at time of graft dysfunction) or as part of an established follow-up protocol (protocol biopsies).[13] Biopsies with missing lesion scores were excluded (n=112), due to missing HLA-DSA (n=73) and/or missing the score of C4d deposition in peritubular capillaries (n=40). 3510 biopsies from 936 recipients remained available for analysis. This study was approved by the Ethical Committee of the University Hospitals Leuven (S64006).

For the validation cohort, the electronic database of Lyon University Hospitals (registration #AC-2016-2706) and the Paris Transplant Group were screened with the same selection criteria as detailed above. Between January 2007 and December 2015 for the Lyon dataset, and between March 2009 and October 2019 for the Paris dataset, respectively 1356 (from 726 transplants) and 2479 biopsies (from 1304 transplants) were included as an independent validation set, performed either for indication or as part of

the routine follow-up at 3 months and 12 months post transplantation). Only complete data were included. Clinical, histological and immunological data were extracted from these databases, anonymized, and transmitted to Leuven to be used as an external independent validation cohort.

Histological scoring

In the training cohort, all post-transplant kidney allograft biopsies performed in this cohort, until the time of data extraction in December 2018, were included. One pathologist (EL) reviewed all biopsies, independent of clinical information to avoid bias. The severity of the histological lesions was semi-quantitatively scored according to the Banff categories with a small deviation for C4d thresholds.[12] The set of individual Banff lesions (N=14) represents either acute or chronic injury processes. We focused on the following 7 acute Banff lesions, with semiquantitative scores reflecting disease activity, in concordance with the Banff guidelines[4]: tubulitis (t; score 0 to 3), interstitial inflammation (i; score 0 to 3), glomerulitis (g; score 0 to 3), intimal arteritis (v; score 0 to 3), C4d deposition in peritubular capillaries (C4d; score 0 to 3), peritubular capillaritis (ptc; score 0 to 3), thrombotic microangiopathy (TMA; present vs. absent). We considered transplant glomerulopathy (cg; score 0 to 3), interstitial fibrosis (ci; score 0 to 3), tubular atrophy (ct; score 0 to 3), vascular intimal thickening (cv; score 0 to 3), mesangial matrix increase (mm; score 0 to 3), arteriolar hyalinosis (ah; score 0 to 3) and glomerulosclerosis (gs; score 0 to 3) as chronic lesions and did not take these lesions into account in the classification of acute rejection phenotypes. As the presence of HLA-DSA is a defining feature in the Banff diagnosis of ABMR, HLA-DSA was also considered in the clustering process (present vs. absent), as defined previously for this cohort.[14]

The biopsies were classified into acute rejection categories based on the criteria as defined by the most recent Banff 2019 consensus[3]. Overall, each biopsy was assigned to one of the six following categories based on the Banff acute rejection phenotype: (1) No rejection, (2) Borderline changes, (3) TCMR, (4) ABMR, (5) Mixed borderline rejection and (6) Mixed rejection. Borderline changes were diagnosed as foci

of tubulitis (t > 0) with minor interstitial inflammation (i1) or moderate-severe interstitial inflammation (i2 or i3) with mild (t1) tubulitis. Antibody-mediated rejection (ABMR) was diagnosed by the presence of the three Banff criteria for either acute or chronic active ABMR according to the Banff 2019 classification, but not taking into account potential non-HLA antibodies or gene expression changes. Due to lack of information on i-IFTA and total-i scores, chronic T-cell mediated rejection was not considered separately. We labeled biopsies presenting an overlap of ABMR and TCMR as Mixed rejection and the biopsies with an overlap of ABMR and Borderline changes as Mixed borderline rejection.

Data analysis

Semi-supervised clustering strategy

We scaled each histological lesion score (feature) into the unit interval. We adapted semi-supervised learning from,[10] where additional information was used to facilitate the creation of clinically meaningful clusters. Specifically, Bair and Tibshirani[10] used the Cox scores from univariate models to perform a feature selection prior to clustering, whereas we used the Cox score to weigh the features. We chose k-means as the core algorithm for the clustering process because of its straightforward implementation, its efficiency, its ability to accommodate the weighting of features and the possibility to classify new biopsies into non-overlapping clusters. The information from the death-censored kidney transplant survival outcome was introduced with a weighted Euclidean distance to provide additional guidance during the clustering process. Each feature was weighted with the normalized coefficient's z score of univariate Cox models, adjusted for clustered data, i.e. repeated biopsies from the same patients, using a *sandwich* variance estimate. Features with a higher weight contribute more heavily to the notion of dissimilarity between clusters than low weight features which will be less relevant to the definition of a cluster. Although guided by external survival information, the clustering task remains mostly unsupervised as the lesion scores patterns are the most influential driving force in the final clusters.

Consensus clustering

We used consensus clustering[15] based on 400 clustering partitions of the data with different random initializations of the k-means algorithm seed and a different subsampling (80%) of the original data, similar to the approach used by Monti.[16] For the clustering process, all biopsies were considered independent. We used the nearest centroid method to assign a cluster label to the remaining 20% of out-of-bag biopsies for each partition. The final consensus clustering was achieved through majority voting along the 400 partitions. To avoid introducing biases in the clustering process by the overrepresentation of protocol biopsies, we adopted a scheme where indication biopsies and protocol biopsies were weighted based on the inverse of their total proportion in the dataset. Cluster profiles were reported using the normalized mean value of lesions, or for binary features the percentage of biopsies with the feature present. We also report the proportion of each original lesions score. Where appropriate, individual lesions scores were compared between a pair of clusters with a χ^2 test. The degree of similarity between two different partitions of the data was evaluated with the adjusted rand index (ARI). This index accounts for overlapping partitions due to chance. It varies from -1 to 1, an ARI of 0 meaning random partitioning. A decision tree was trained on the cluster-labelled data to mimic the internal clustering process. The tree was generated using the Gini criterion with a minimum of 10 biopsies per leaf.

Tuning of parameters

To define the optimal number of clusters, we used the proportion of ambiguous clustering (PAC)[17] to assess the stability of our results at different values of k, i.e. the number of clusters, with thresholds set at 10.0% and 90.0% of consensual clustering. Intuitively, PAC measures the proportion of all possible pairs of biopsies from the whole dataset which demonstrate inconsistent clusters attributions over the 400 partitions. The lower the PAC, the more stable the clustering across different conditions. We

discarded very low value of k, since they only create a restricted number of clusters (typically No rejection vs. any rejection with k=2), which is not helpful to describe different phenotypes.

Biopsy stability

In order to identify biopsies that are part of pairs with an unstable cluster assignment over the set of clustering partitions, we developed an empirical individual stability score based on the consensus matrix. The consensus matrix C is a $n \times n$ matrix, where n is the total number of biopsies and entries $C_{i,j}$ represent the proportion of times biopsies i and j are clustered in the same cluster over the whole set of different partitions. The stability score s_i for biopsy *i* was defined as $s_i = \frac{1}{N/2}\sum_{j=1}^{N} |c_{i,j} - 0.5|$ where $c_{i,j}$ represented the value from the consensus matrix at the ith row and j^{th} column and *N* the total number of biopsies. Intuitively, a theoretical score of 1 reflects that the biopsy was consistently part of biopsy pairs that are either clustered together in the same cluster or clustered in different clusters. A theoretical score of 0 would reflect a biopsy that forms ambiguous pairs with any other biopsy.

Survival analysis

Graft survival times are reported as number of days until graft failure, calculated from each biopsy date. Patients were administratively censored at the of last follow-up date or at time of death. Survival curves are plotted with Kaplan-Meier estimators along with the 95% confidence interval. To avoid artificially increasing the incidence of transplant failure events due to repeated biopsies in a given individual, survival times from repeated biopsies in a given cluster were averaged for each patient. Pairwise comparison of survival curves was performed using Cox modelling and hazard ratio (HR) with 95% confidence interval. Because potentially, proportional hazard assumption violations might bias the HR, we also report the restricted mean survival times (RMST)[18] at 5 and 10 years and its confidence interval at 95%. This measure can be interpreted as the mean survival time without event within a pre-defined

time range, representing the area under the survival curve up to a pre-defined time-point. We also report the differences in RMST (DRMST) with a baseline category, which estimates the difference in average event-free survival, in years, between a given category and the baseline group.

Visualization

Principal component analysis (PCA) was performed on the Cox-score weighted acute lesions scores and the first two components were used for 2D visualization purposes. To better visualize the heterogeneity in the acute lesion scores, we developed a two-dimensional plot using polar coordinates, with the radius calculated as the sum of re-weighted acute lesions scores, scaled to the unit interval (from 0 to 1), and the theta angle is a scaled version (for visual purposes) of the second component of the PCA, which is directly related to the main rejection phenotype. Because the sum of lesions is directly related to graft failure due to the individual weighting of lesions scores, this approach combines the severity and the phenotype trend in one single plot.

All analyses have been performed with Python 3.6.[19] A web application where others can upload their own patient data and derive the clusters from the individual Banff lesion scores is available at https://*rejectclass.pythonanywhere.com.*

RESULTS

Patient and biopsy characteristics

Descriptive patient (N=936) and biopsy (N=3510) data of the training cohort are shown in Table 1. On average, 3.75 biopsies (range = 1 to 11) were performed per patient. Of the 773 indication biopsies, 644 (83.3%) were performed within the first year of transplantation (median at 22 days post-transplantation), and 129 (16.7%) after one year. HLA-DSA were present at the time of 468 (13.3%) biopsies.

Semi-supervised clustering of rejection phenotypes

Fully unsupervised clustering of our biopsy cohort (N=3510) yielded an optimum of 4 different clusters, based on the proportion of ambiguous clustering (S1 Fig). Compared to cluster 1 (essentially normal biopsies), the three other clusters associated significantly with impaired graft survival. However, their histological and clinical relevance were less clear, as none of these three clusters were defined based on microcirculation inflammation and antibody activity (glomerulitis, peritubular capillaritis and C4d), suggesting that the number of clusters was insufficient to reflect the clinical reality and previous knowledge on the relevance of these lesions and ABMR. Increasing the number of clusters created clusters that were no longer associated with impaired graft survival compared to cluster 1 (S1 Table).

To optimize the clinical significance of the clusters, we applied a semi-supervised clustering approach, weighing the histological features with survival information. The optimal number of clusters (*k*) was six, based on the proportion of ambiguous clustering (S1 Table). We labeled the six identified clusters from 1 to 6, according to the overall association with graft failure (Fig 1; S2 Table). Biopsies in cluster 1 were dominated by 0 scores for the lesions, in cluster 2 by high g scores, and in cluster 3 by t and i. Clusters 1 to 3 were HLA-DSA negative, while all biopsies included in clusters 4 to 6 were in patients with HLA-DSA.

Biopsies in cluster 1 had no or very limited inflammation, and good outcome, and could be considered as No Rejection. In cluster 2, all cases had moderate to severe glomerulitis in the absence of HLA-DSA, sometimes accompanied by tubulo-interstitial inflammation and peritubular capillaritis. These cases of glomerulitis in the absence of HLA-DSA are currently not fully understood and not reflected in the current Banff classification, yet associate with impaired graft outcome. Cluster 3 is characterised by moderate to severe degrees of tubulo-interstitial inflammation, reminding of acute TCMR in the Banff classification. In cluster 4, no or only very limited inflammation is noted, sometimes C4d deposition in peritubular capillaries. All cases in cluster 4 were HLA-DSA positive, which appeared to be a risk factor for graft failure in comparison to cluster 1, even in the absence of extensive inflammation. Biopsies in cluster 5 were HLA-DSA positive with high g scores and could be considered to reflect active ABMR. Biopsies in cluster 6 were HLA-DSA positive with high t and i scores, often combined with g and ptc, and could suggest "Mixed rejection". Biopsies in cluster 1 and 4 were most often protocol biopsies, while clusters 2, 3 and 5 were similarly distributed between protocol and indication biopsies (S3 Table). Cluster 6 was observed most often in indication biopsies and had worst eGFR and highest proteinuria.

Although we focused on acute histological lesions, these lesions often co-occurred with chronic lesions (S2 Fig). With k smaller or greater than six, we observed larger PAC (S1 Table). Increasing k did not drastically reshape previously found clusters, but rather added new clusters while conserving the similar centroids of the clusters derived at lower k. With k=7, we observed the separation of cluster 1 in two clusters based on the t lesion (and to a lesser extent i). However, the survival curves from those two sub-clusters were largely overlapping (log-rank test p-value 0.97), illustrating that also from clinical perspective, the optimal number of clusters was k=6. The ARI was similar between the various k values and the minimal distance between two centroids also decreased with greater k.

Based on the consensus matrix, the average stability scores per cluster were 0.98, 0.98, 0.99, 0.98 and 0.99 for clusters 1, 3, 4, 5 and 6, respectively. In comparison to the other clusters, cluster 2, characterized by glomerulitis in the absence of HLA-DSA, was less stable with an average stability score of 0.75. 44 biopsies had a low stability score (< 0.5): 30 biopsies from cluster 2 (29.7%) and 14 biopsies from cluster 1 (0.5%). Because k-mean is a distance-based algorithm, it is possible to compute relative distances to the closest clusters' boundary. If a biopsy is nearer to a cluster centroid than it is from the 2nd closest centroid, the relative distance will be small. On the other hand, a biopsy that is near the clusters' boundary will get a relative distance approaching 1, translating an almost equidistant position (S3 Fig). Biopsies with low stability scores were mostly found on the cluster edges.

Comparison of disease clusters with Banff 2019 rules

There was important overlap between the clusters and the Banff categories with an ARI of 0.48 (Table 2; Supplemental Results). Due to its distance-based approach, the clustering algorithm led to better separation of the biopsies than the Banff 2019 classification, as also illustrated in the plots of PCA applied on the weighted acute lesions scores (Fig 2A). Although all lesions were taken into account simultaneously to assign each biopsy to a cluster, a decision tree could be derived, based on the four main driving forces, g, HLA-DSA, i and t (S4 Fig). This decision tree assigned the correct cluster with 97.0% of balanced accuracy, which confirmed the dominance of these 4 lesions in the phenotype reclassification. The 3% misclassified cases related to 24 biopsies.

Quantitative visual presentation of disease clusters

As expected, the superposition of the 6 disease clusters on the 2-D polar plot aligned better visually with the mathematical disease reclassification than with the different Banff 2019 phenotypes (Fig 2B). Biopsies projected with a negative angle were mostly associated with Banff TCMR, whereas those with a positive angle represented Banff ABMR. Biopsies with mixed rejection phenotypes were projected

around 0°. When plotting individual lesions, and also combinations of those (g+ptc = microcirculation inflammation) and (i+t = tubulo-interstitial inflammation)(S5 Fig), the PCA and the theta values associated with these two major components (microcirculation inflammation vs. tubulo-interstitial inflammation) driving the disease reclassification. The radius on the plot was higher in indication biopsies compared to protocol biopsies (mean ± sd : 0.22 ± 0.23 vs. 0.08 ± 0.13 respectively, student t test $p<0.0001$), illustrating more inflamed biopsies at time of graft dysfunction than at time of stable graft function (S6 Fig).

Association between disease clusters and graft failure

During follow-up, 125 grafts failed, at a median of 3.67 years (1 day to 12 years) after transplantation. 9.1%, 22.4%, 25.0%, 30.0%, 37.7% and 50.0% of grafts failed within the first 5 years after the biopsy in respectively cluster 1 to 6. The disease clusters 2 to 6 all associated with an increased risk of graft failure in comparison the cluster 1 (Fig 1 and Table 3). Although Banff rejection categories had significant association with graft failure, except for Borderline changes, the clusters weighted average in DRMST at 5 and 10 years were higher than the weighted average DRMST from the Banff classification (respectively 0.46 and 1.25 years for the clusters vs. 0.29 and 0.72 years for the Banff categories), illustrating an overall better discrimination in terms of graft failure (Table 3). Furthermore, we observed an asymmetry between the first three and last three clusters, based on HLA-DSA status. Hazard ratios on the HLA-DSA negative/HLA-DSA positive pair of clusters reported the following values: cluster 1 vs cluster 4 : 2.84 (CI 95% 1.80-4.30; $p<0.0001$), cluster 2 vs. cluster 5 : 2.02 (1.00-4.12 ; $p= 0.051$), cluster 3 vs. cluster 6: 2.41 (CI 95% 1.35-4.30 ; $p= 0.003$) (S7 Fig). The survival outcome of each cluster did not depend on the adjustment method for repeated biopsies per patient (S8 Fig). The clustering of biopsies led to improved prediction of graft failure, compared to the Banff classification (Supplemental Results).

The radius on the polar plot of each biopsy associated independently with graft failure, with an AUROC for 2- and 5-year post-biopsy graft survival of respectively 0.70 (95% CI 0.66-0.73) and 0.69 (95% CI 0.67-0.72), respectively. Biopsies projected on the outer ranges of the radius had higher inflammatory lesion scores, and significantly worse survival compared to biopsies near the center of the polar plot (Fig 3A). Similar association with graft failure was obtained when we predicted graft failure for each biopsy separately, from the information available on the nearest neighborhood, calculated using the weighted Euclidean distance. For example, Fig 3B displays the survival probability at 5 years post-biopsy, estimated from local Kaplan-Meier estimates based on 40 nearest neighbors. With this local approach, solely based on the lesion scores and HLA-DSA status and not taking into account graft functional data or post-transplant time, the AUROC of the probability for graft failure were 0.72 (95% CI 0.68-0.74) and 0.70 (95% CI 0.67-0.73), respectively at 2- and 5-years post-biopsy.

External validation

Using the features weights and the cluster centroids obtained from the consensus clustering process, we are able to classify any new biopsy into one of the 6 previously described clusters. We applied this algorithm, starting from the lesion scores and HLA-DSA status only, without information on graft survival, to an external dataset of 3835 biopsies from Lyon University Hospital (N=1356) and the Paris Transplant Group (N=2479)(S4 Table). Note that this dataset did not include thrombi in its variables. We therefore imputed this feature from the mean value of our training data. A comparison of the final clusters proportions between the two centers is presented in S9 Fig. Similar to the training set, biopsies from the external validation set were largely dominated by non-inflamed cluster 1. The main difference in cluster distribution was a higher proportion of cluster 4 biopsies in the external dataset compared to the Leuven dataset (26.0% vs 8.7%, $p<0.0001$), explained by a larger prevalence of HLA-DSA positive biopsies in the external data. Logically, the proportion of lesions within each clusters of the external validation set were

very similar to the clusters obtained from the original data. There was also a similar association of the clusters with graft failure (S10 Fig).

A polar plot illustrates the full overlap in the histological presentations between the training and validation cohorts (S11 Fig). Although the proportion of biopsies performed upon indication was notably higher in the validation cohort (22.0% vs 37.7%, χ^2 test $p<0.0001$), the overall distribution of inflammation, estimated using the radius on the polar plot, was comparable between the training and validation datasets (S12 Fig). Comparing the clusters obtained on the validation dataset with the Banff categories, we obtained an ARI of 0.35. While maintaining a large overlap between the clustering method and the Banff classification (S5 Table), it demonstrates a higher reclassification rate in the validation dataset.

DISCUSSION

Using a semi-supervised and data-driven approach on 7345 post-transplant kidney biopsies with re-weighting of acute histological lesions, we derived and validated six distinct, clinically meaningful, phenotypic clusters. This mathematical clustering approach was fundamentally different from the iterative Banff classification process, which relies on a set of clinically derived *if-then* rules. Nevertheless, both in the training and the validation cohort, the novel phenotypes for kidney transplant rejection had a good degree of similarity with the Banff rejection categories, while redistributing intermediate and mixed phenotypes and maintaining the association with graft failure. The novel rejection phenotypes led to improved prediction of graft failure compared to the Banff classification. For integration of the novel phenotypic clustering with disease severity, and to move away from the black-white disease categorization, we proposed and validated a method for easily interpretable two-dimensional visual and quantitative presentation of the multidimensional histologic data.

Despite the similarity between the novel clusters and the Banff categories, we showed statistically improved prediction of graft failure with the clustering approach than when using the Banff categories, especially in ambiguous situations, like Borderline changes or mixed rejection phenotypes. The association between (non)-inflamed clusters and graft survival remained present even when the biopsies were stratified according to the rejection or non-rejection categories defined by Banff. An example of the clinical impact of this is e.g. the lack of cluster reflecting Banff Borderline changes. Borderline changes are not reflected in a separate cluster, but most often (79.8%) classified to non-inflamed cluster 1, with best post-transplant graft survival. Using this clustering approach therefore may solve the clinically difficult issue of how to deal with minimal tubulo-interstitial inflammation, below the current thresholds for TCMR. Also, the clustering algorithm proposed a novel phenotype, which is driven by glomerulitis in the absence of HLA-DSA. Although the causes of this phenotype are currently unknown, this resembles the phenotype described in recent publications on HLA-DSA negative microcirculation inflammation.[12,20–

[22] This phenotype is currently not recognized in the Banff classification[6] and should be worked out in greater detail with respect to pathophysiology, risk factors and clinical presentation. Finally, cluster 6 represents cases with mixed rejection phenotypes (ABMR with TCMR or Borderline changes), which is not recognized as separate category in the Banff classification, but representing a clinical dilemma.[23]

Our clustering method directly relies on distance computation and provides a clinically relevant similarity metric to compare biopsies without concomitant clinical data besides HLA-DSA. For instance, we demonstrated and validated that local survival prediction based on the nearest biopsies exhibited prognostic value solely based on the histological lesions and HLA-DSA status, thus not taking into account graft functional parameters or demographic factors relevant for outcome.[24] Relying on this ad-hoc distance metric and to move beyond the black-white clustering approach, we developed an intuitive two-dimensional visualization tool, enabling to plot newly performed post-transplant biopsies and rapidly assess the disease severity along with the dominant phenotype of neighboring biopsies. Because the k-mean algorithm is a hard-clustering algorithm, biopsies near the clusters' boundaries get strictly allocated to one of the two neighbor clusters, preventing an overlap of diagnoses. This explains for instance that the mixed rejection biopsies are now split into one of the major clusters based on their dominant lesions. However, contrasting with the Banff categorization, our clustering system can provide some degree of certainty regarding the classification, as expressed in term of the relative distance to the closest cluster prototype (centroid). As a time-independent approach, our method is intended solely for reclassification of rejection (clustering algorithm and theta angle on the polar plot) and assessment of disease severity (radius on the polar plot). Our analysis on the accuracy of the local survival prediction needs to be seen as support for the clinical validity of the location of each sample on the polar plot and does not suggest clinical utility of the local survival prediction as a prognostic tool on its own. For prognostication, more granular tools are becoming available, such as the iBox prediction score,[24] which also integrate time post-transplantation and graft functional parameters into the models. Finally,

diagnosis of other relevant disease phenotypes like glomerulonephritis or polyomavirus nephropathy are based on other parameters that are currently not included in the algorithm. These diseases should not be evaluated with our system solely intended for reclassification of rejection phenotypes.

As chronic histological lesions in kidney transplant biopsies are non-specific,[4] we focused solely on acute inflammatory lesions to derive the novel rejection phenotypes. The evolution from active/early stage disease, to chronic active, and finally chronic inactive forms of the same disease, was therefore not assessed and can be considered for future developments reclassification system. In addition, our approach fully depends on the quality of the histological assessment, which is pathologist-dependent and therefore not fully reproducible.[25,26] More objective data, such as computerized imaging data or molecular expression data, or information on e.g. non-HLA antibodies and other immune risk factors, could further improve the reproducibility and accuracy of our system. Next, although the clusters described are sound biologically/clinically, whether treatment decisions based on clusters instead of based on Banff diagnosis will yield better outcome cannot be tested in this retrospective study. Similarly, data-driven algorithms do not assess pathophysiological mechanisms, hence no causal relations can be deducted from any cluster. Besides these clinical aspects, also some technical limitations warrant discussion. In concordance to the method described earlier,[10] we used the whole dataset to compute the lesion score weights. The more data available to compute the weights, the more precise their estimation will be. In this semi-supervised setting, weights overfitting is less detrimental than in a purely supervised approach. Despites its good performance, the k-mean algorithm remains simplistic. More elaborated core clustering algorithms, such as model-based or fuzzy clustering methods, could benefit the current approach and warrants additional studies.

Although we described a meaningful data-driven alternative to classify kidney transplant biopsies, and although our system has benefits over the current Banff categories, we do not suggest replacing the

existing Banff classification with this algorithm but use it in addition to Banff categorization, especially in cases that are difficult to categorize according to Banff. The clinical or scientific utility of our approach needs to be shown in further studies that validate the improved clinical decision-making with regards to rejection treatment. Clinical implementation will depend on further external validation and detailed discussion at future Banff meetings and international consensus. The underlying risk factors and clinical presentations of each of the clusters still needs to be evaluated in greater depth, including information on HLA-DSA subtypes and profiles, non-HLA antibodies etc. Inference on treatment decisions could not be made on our cohort, given the fact that Banff-rejection cases were treated with high-dose corticosteroids, and that cases of ABMR were treated with antibody-targeted therapies only very rarely[12]. Nevertheless, the current study highlights the potential of using the full scale of lesion grades for classification of kidney transplant biopsies, rather than using discretionary cut-off values. In the era of increasing availability of morphometric[27] or molecular[2] data, advanced statistical analysis and machine learning, with many resources to handle high-dimensional continuous variables,[28] the existing expert-based consensus of *if-then* rules could be further improved using our approaches.

Conclusion

We have developed and validated a semi-supervised clustering approach for the identification of clinically meaningful novel phenotypes for kidney transplant rejection, based on individual lesion scores. This approach has potential to offer a more quantitative evaluation of rejection subtypes and severity, especially in situations where the current histological categorization is ambiguous.

Author contributions

TV and MN designed the study and the analysis plan. EL, AS, EVL, JC, MPE, BS, VD, MR, DK, OT and MN were involved in clinical data collection and data quality control. TV did the statistical analyses and created the Figs and tables, with input from BDM and MN. TV and MN interpreted the results and wrote the article, and all coauthors revised and approved it.

Declaration of interests

The authors declare no competing interests.

Acknowledgments

The authors thank the centers of the Leuven Collaborative Group for Renal Transplantation, the clinicians and surgeons, nursing staff and the patients. OT is grateful to the Dr Dijou and Dr Picard from the Department of Pathology of the Hospices Civils de Lyon for their contribution to the histological follow-up of kidney transplant patients. OT is indebted to the members of the Laboratoire de Recherche Translationnelle en Immunologie des Greffes from Edouard Herriot Hospital for their help during data collection.

Financial Disclosure Statement

This work is supported by The Research Foundation Flanders (FWO) and the Flanders Innovation & Entrepreneurship agency (VLAIO), with a TBM project (grant no. IWT.150199) and by a C3 internal grant from the KU Leuven (grant no. C32/17/049). MN and BS are senior clinical investigators of The Research Foundation Flanders (FWO) (1844019N and 1842919N, respectively). TV, EVL, JC hold a fellowship grant (1S93918N, 1143919N and 1196119N, respectively) from The Research Foundation Flanders (FWO). OT is supported by the Agence Nationale pour la Recherche (ANR-16-CE17-0007-01), the Fondation pour la Recherche médicale (PME20180639518), and the Etablissement Français du Sang. BDM is supported by KU Leuven Research Fund (projects C16/15/059, C32/16/013, C24/18/022), Industrial Research Fund (Fellowship 13-0260) and several Leuven Research and Development bilateral industrial projects, Flemish Government Agencies: FWO (EOS Project no 30468160 (SeLMA), SBO project I013218N. BDM also received funding from the Flemish Government (AI Research Program), VLAIO (City of Things (COT.2018.018) Industrial Projects (HBC.2018.0405), and the European Commission. This project has

received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 885682). AL is supported by French national research agency (INSERM) ATIP Avenir. OA hold a fellowship grant from the Fondation Bettencourt Schueller. GD hold a fellowship grant from the Fondation pour la Recherche Médicale.

TABLE OF CONTENTS FOR THE SUPPLEMENTAL MATERIAL

Supplemental Results

S1 Table. Performance indices for a range of k values used to define the optimal number of clusters in the semi-supervised clustering algorithm

S2 Table. Details of cluster composition according to the individual Banff lesions scores and donor-specific HLA antibodies (N=3510 biopsies of the derivation cohort).

S3 Table. Distribution of biopsies among clusters and stratification into protocol vs indication biopsies (N=3510 biopsies of the derivation cohort).

S4 Table. Demographic, clinical and histological characteristics of the patients and biopsies included in the validation dataset.

S5 Table. Contingency tables comparing the Banff 2019 diagnosis and the 6 clusters obtained on the external validation dataset.

S1 Fig. Distribution of the individual acute lesion scores in the clusters using an unweighted approach, and post-biopsy Kaplan-Meier graft survival curves relative to cluster 1 of the derivation cohort

S2 Fig. Distribution of chronic lesions in the 6 acute lesion clusters

S3 Fig. Relative distances to the closest clusters' boundary

S4 Fig. Decision tree of the clustering process

S5 Fig. Various combinations of lesions scores displayed on the polar plots.

S6 Fig. Comparison of indication vs. protocol biopsies, as superposed on the polar plot.

S7 Fig. Post-biopsy graft survival in the three DSA-/DSA+ pair of clusters.

S8 Fig. Post-biopsy graft survival in the six clusters, according to the adjustment method for repeated biopsies per patient.

S9 Fig. Comparison of cluster proportion per center

S10 Fig. Distribution of the individual acute lesion scores in the different clusters, and post-biopsy Kaplan-Meier graft survival curves relative to cluster 1 of the external validation cohort

S11 Fig. Overlay of the data from Leuven and the external dataset in the polar plot, according to the six clusters identified in the derivation cohort.

S12 Fig. Distribution of the radius from the polar plot

S13 Fig. Association with graft survival in the polar plot visualization tool based on the validation data

References

1. Solez, K. *et al.* International standardization of criteria for the histologic diagnosis of renal allograft rejection: the Banff working classification of kidney transplant pathology. *Kidney Int.* **44**, 411–422 (1993).
2. Haas, M. *et al.* The Banff 2017 Kidney Meeting Report: Revised diagnostic criteria for chronic active T cell–mediated rejection, antibody-mediated rejection, and prospects for integrative endpoints for next-generation clinical trials. *Am. J. Transplant.* **18**, 293–307 (2018).
3. Loupy, A. *et al.* The Banff 2019 Kidney Meeting Report (I): Updates on and clarification of criteria for T cell– and antibody-mediated rejection. *Am. J. Transplant.* doi:10.1111/ajt.15898.
4. Roufosse, C. *et al.* A 2018 Reference Guide to the Banff Classification of Renal Allograft Pathology: *Transplantation* **102**, 1795–1814 (2018).
5. Racusen, L. C. *et al.* Antibody-mediated rejection criteria - an addition to the Banff 97 classification of renal allograft rejection. *Am. J. Transplant. Off. J. Am. Soc. Transplant. Am. Soc. Transpl. Surg.* **3**, 708–714 (2003).
6. Haas, M. *et al.* Banff 2013 meeting report: inclusion of c4d-negative antibody-mediated rejection and antibody-associated arterial lesions. *Am. J. Transplant. Off. J. Am. Soc. Transplant. Am. Soc. Transpl. Surg.* **14**, 272–283 (2014).
7. Loupy, A. *et al.* The Banff 2015 Kidney Meeting Report: Current Challenges in Rejection Classification and Prospects for Adopting Molecular Pathology. *Am. J. Transplant. Off. J. Am. Soc. Transplant. Am. Soc. Transpl. Surg.* **17**, 28–41 (2017).
8. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. (Springer Science & Business Media, 2009).
9. Bullinger, L. *et al.* Use of Gene-Expression Profiling to Identify Prognostic Subclasses in Adult Acute Myeloid Leukemia. *N. Engl. J. Med.* **350**, 1605–1616 (2004).

10. Bair, E. & Tibshirani, R. Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data. *PLOS Biol.* **2**, e108 (2004).
11. Kickingereder, P. *et al.* Radiomic Profiling of Glioblastoma: Identifying an Imaging Predictor of Patient Survival with Improved Performance over Established Clinical and Radiologic Risk Models. *Radiology* **280**, 880–889 (2016).
12. Senev, A. *et al.* Histological picture of antibody-mediated rejection without donor-specific anti-HLA antibodies: Clinical presentation and implications for outcome. *Am. J. Transplant. Off. J. Am. Soc. Transplant. Am. Soc. Transpl. Surg.* **19**, 763–780 (2019).
13. Coemans, M. *et al.* Occurrence of Diabetic Nephropathy After Renal Transplantation Despite Intensive Glycemic Control: An Observational Cohort Study. *Diabetes Care* (2019) doi:10.2337/dc18-1936.
14. Senev, A. *et al.* Specificity, strength, and evolution of pretransplant donor-specific HLA antibodies determine outcome after kidney transplantation. *Am. J. Transplant.* **19**, 3100–3113 (2019).
15. Strehl, A. & Ghosh, J. Cluster ensembles --- a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **3**, 583–617 (2003).
16. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Mach. Learn.* **52**, 91–118 (2003).
17. Şenbabaoğlu, Y., Michailidis, G. & Li, J. Z. Critical limitations of consensus clustering in class discovery. *Sci. Rep.* **4**, (2014).
18. Royston, P. & Parmar, M. K. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med. Res. Methodol.* **13**, 152 (2013).
19. Rossum, G. van. *Python tutorial*. (1995).

20. Koenig, A. *et al.* Missing self triggers NK cell-mediated chronic vascular rejection of solid organ transplants. *Nat. Commun.* **10**, 1–17 (2019).
21. Bestard, O. & Grinyó, J. Refinement of humoral rejection effector mechanisms to identify specific pathogenic histological lesions with different graft outcomes. *Am. J. Transplant.* **19**, 952–953 (2019).
22. Callemeyn, J. *et al.* Transcriptional Changes in Kidney Allografts with Histology of Antibody-Mediated Rejection without Anti-HLA Donor-Specific Antibodies. *J. Am. Soc. Nephrol. JASN* (2020) doi:10.1681/ASN.2020030306.
23. Madill-Thomsen, K. *et al.* Discrepancy analysis comparing molecular and histology diagnoses in kidney transplant biopsies. *Am. J. Transplant. Off. J. Am. Soc. Transplant. Am. Soc. Transpl. Surg.* (2019) doi:10.1111/ajt.15752.
24. Loupy, A. *et al.* Prediction system for risk of allograft loss in patients receiving kidney transplants: international derivation and validation study. *BMJ* **366**, (2019).
25. Furness, P. N. *et al.* International variation in histologic grading is large, and persistent feedback does not improve reproducibility. *Am. J. Surg. Pathol.* **27**, 805–810 (2003).
26. Smith, B. *et al.* A method to reduce variability in scoring antibody-mediated rejection in renal allografts: implications for clinical trials - a retrospective study. *Transpl. Int. Off. J. Eur. Soc. Organ Transplant.* **32**, 173–183 (2019).
27. Sicard, A. *et al.* Computer-assisted topological analysis of renal allograft inflammation adds to risk evaluation at diagnosis of humoral rejection. *Kidney Int.* **92**, 214–226 (2017).
28. Fröhlich, H. *et al.* From hype to reality: data science enabling personalized medicine. *BMC Med.* **16**, 150 (2018).

TABLE 1. Demographic, clinical and histological characteristics of the patients and biopsies included.

Cohort characteristics	Total (N=936)
Donor demographics	
Donor type	
Donation after brain death, N (%)	726 (77.6)
Donation after cardiac death, N (%)	153 (16.3)
Living donation, N (%)	57 (6.1)
Age (years), mean ± SD	47.7 ± 14.7
Male, N (%)	497 (53.1)
Diabetes, N (%)	24 (2.6)
Recipient demographics	
Age (years), mean ± SD	53.5 ± 13.3
Male, N (%)	572 (61.1)
Ethnicity	
Caucasian, N (%)	920 (92.3)
African, N (%)	12 (1.3)
Asian, N (%)	3 (0.3)
Hispanic, N (%)	1 (0.1)
BMI (kg/m^2), mean (range)	25.4 (4.5)
Pre-transplant donor-specific HLA antibodies, N (%)	408 (11.6%)
Repeat transplantation, N (%)	141 (15)
Cold ischemia time (hours), mean ± SD	14.2 ± 5.7
Total number of HLA A/B/DR mismatches, mean ± SD	2.8 (1.3)
Biopsy characteristics	**Total (N=3510)**
Banff 2019 diagnosis	
No rejection, N (%)	2671 (76.1)
Borderline changes, N (%)	333 (9.5)
TCMR, N (%)	314 (8.9)
ABMR, N (%)	110 (3.1)
Mixed rejection (ABMR + TCMR), N (%)	61 (1.7)
Mixed borderline rejection (ABMR + borderline changes), N (%)	21 (0.6)
Indication biopsies, N (%)	N=773 (22.0)
Days since transplantation, median (interquartile range)	22 (8-96)
eGFR at day of biopsy, median (interquartile range)	19.8 (10.9-29.0)
Protocol biopsies, N (%)	N=2737 (78.0)
3 months, N (%)	823 (30.1)
12 months, N (%)	759 (27.7)
24 months, N (%)	639 (23.3)
36 months, N (%)	205 (7.5)
48 months, N (%)	22 (0.8)
60 months, N (%)	289 (7.6)
Days since transplant, median (interquartile range)	377 (100-752)
eGFR at day of biopsy, median (interquartile range)	46.4 (36.5-57.8)

TABLE 2. Contingency tables comparing the Banff 2019 diagnosis and the 6 clusters derived from semi-supervised learning. Proportions represent the distribution in the clusters per Banff category (N=3510 biopsies).

Banff 2019 diagnosis	N	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
No rejection	2659	2387 (89.8%)	53 (2.0%)	4 (0.2%)	215 (8.1%)	0 (0.0%)	0 (0.0%)
Borderline changes	327	261 (79.8%)	9 (2.8%)	26 (8.0%)	23 (7.0%)	0 (0.0%)	8 (2.4%)
TCMR	285	48 (16.8%)	25 (8.8%)	184 (64.6%)	5 (1.8%)	0 (0.0%)	23 (8.1%)
ABMR	122	8 (6.6%)	4 (3.3%)	0 (0.0%)	56 (45.9%)	53 (43.4%)	1 (0.8%)
Mixed borderline rejection	27	1 (3.7%)	3 (11.1%)	1 (3.7%)	3 (11.1%)	15 (55.6%)	4 (14.8%)
Mixed rejection	90	5 (5.6%)	7 (7.8%)	16 (17.8%)	5 (5.6%)	27 (30.0%)	30 (33.3%)
Total	**3510**	**2710 (77.2%)**	**101 (2.9%)**	**231 (6.6%)**	**307 (8.7%)**	**95 (2.7%)**	**66 (1.9%)**

TABLE 3. Graft survival, restricted mean survival time (RMST) and difference in RMST (DMRST) at 5- and 10-years post-biopsy, according to each cluster and each Banff diagnostic category (N=3510).

Banff diagnosis	% Graft survival at 5 years post-biopsy	% Graft survival at 10 years post-biopsy	RMST at 5 years post-biopsy (95% CI)	RMST at 10 years post-biopsy (95% CI)	HR vs. No rejection	HR p-value vs. No rejection	DRMST at 5 years vs. No rejection (95% CI)	DRMST at 10 years vs. No rejection (95% CI)
No rejection	89.5%	51.0%	4.74 (4.63-4.85)	9.01 (8.56-9.46)	-	-	-	-
Borderline changes	83.1%	42.3%	4.66 (4.47-4.85)	8.87 (8.04-9.7)	1.27 (0.88-1.84)	0.201	0.08 (-0.06-0.22)	0.14 (-0.24-0.53)
TCMR	75.5%	41.2%	4.50 (4.27-4.74)	8.46 (7.7-9.22)	1.66 (1.17-2.36)	0.004	0.24 (0.05-0.42)	0.55 (0.09-1.01)
ABMR	70.2%	22.0%	4.26 (3.89-4.63)	7.63 (6.25-9.02)	2.63 (1.65-4.21)	<0.0001	0.48 (0.15-0.81)	1.38 (0.51-2.25)
Mixed borderline rejection	63.6%	8.3%	3.55 (4.06-4.57)	6.74 (4.06-7.72)	4.26 (2.29-7.94)	<0.0001	0.67 (0.06-1.28)	2.13 (0.65-3.61)
Mixed rejection	59.2%	20.5%	3.98 (3.48-4.48)	7.03 (5.74-8.32)	3.24 (2.08-5.05)	<0.0001	0.76 (0.34-1.18)	1.98 (1.00-2.96)
Average	-	-	-	-		-	0.45 (0.11-0.78)	1.24 (0.40-2.07)
Weighted average	-	-	-	-		-	0.29 (0.06-0.53)	0.72 (0.14-0.93)
Cluster	**% Graft survival at 5 years post-biopsy**	**% Graft survival at 10 years post-biopsy**	**RMST at 5 years post-biopsy (95% CI)**	**RMST at 10 years post-biopsy (95% CI)**	**HR vs. cluster 1**	**HR p-value vs. cluster 1**	**DRMST at 5 years vs. cluster 1 (95% CI)**	**DRMST at 10 years vs. cluster 1 (95% CI)**
Cluster 1	90.9%	54.6%	4.76 (4.65-4.87)	9.09 (8.63-9.56)	-	-	-	-
Cluster 2	77.6%	33.3%	4.42 (4.01-4.84)	8.28 (6.98-9.59)	1.98 (1.15-3.43)	0.014	0.34 (0.02-0.70)	0.81 (-0.03-1.65)
Cluster 3	75.0%	39.8%	4.54 (4.30-4.79)	8.56 (7.75-9.36)	1.72 (1.17-2.52)	0.005	0.22 (0.02-0.41)	0.53 (0.05-1.02)
Cluster 4	70.0%	28.6%	4.23 (3.87-4.58)	7.67 (6.36-8.98)	2.84 (1.88-4.30)	<0.0001	0.53 (0.24-0.82)	1.42 (0.69-2.15)
Cluster 5	62.3%	6.1%	4.06 (3.48-4.64)	6.84 (5.04-8.64)	4.17 (2.48-7.03)	<0.0001	0.70 (0.26-1.14)	2.25 (1.07-3.42)
Cluster 6	50.0%	6.2%	3.96 (3.43-4.48)	6.78 (4.96-8.59)	4.37 (2.59-7.35)	<0.0001	0.80 (0.31-1.29)	2.31 (1.11-3.52)
Average	-	-	-	-		-	0.52 (0.17-0.87)	1.46 (0.58-2.35)
Weighted average	-	-	-	-		-	0.46 (0.09-0.69)	1.25 (0.48-2.01)