Deep learning-based classification of kidney transplant pathology: a retrospective, multicentre, proof-of-concept study

Kers, J.; Bulow, R.D.; Klinkhammer, B.M.; Breimer, G.E.; Fontana, F.; Abiola, A.A.; ... ; Boor, P.

Deep learning-based classification of kidney transplant pathology: a retrospective, multicentre, proof-of-concept study

Jesper Kers, Roman D Bülow*, Barbara M Klinkhammer, Gerben E Breimer, Francesco Fontana, Adeyemi Adefidipe Abiola, Rianne Hofstraat, Garry L Corthals, Hessel Peters-Sengers, Sonja Djudjaj, Saskia von Stillfried, David L Hölscher, Tobias T Pieters, Arjan D van Zuilen, Frederike J Bemelman, Azam S Nurmohamed, Maarten Naesens, Joris J T H Roelofs, Sandrine Florquin, Jürgen Floege, Tri Q Nguyen, Jakob N Kather†, Peter Boor†*

Summary

Background Histopathological assessment of transplant biopsies is currently the standard method to diagnose allograft rejection and can help guide patient management, but it is one of the most challenging areas of pathology, requiring considerable expertise, time, and effort. We aimed to analyse the utility of deep learning to preclassify histology of kidney allograft biopsies into three main broad categories (ie, normal, rejection, and other diseases) as a potential biopsy triage system focusing on transplant rejection.

Methods We performed a retrospective, multicentre, proof-of-concept study using 5844 digital whole slide images of kidney allograft biopsies from 1948 patients. Kidney allograft biopsy samples were identified by a database search in the Departments of Pathology of the Amsterdam UMC, Amsterdam, Netherlands (1130 patients) and the University Medical Center Utrecht, Utrecht, Netherlands (717 patients). 101 consecutive kidney transplant biopsies were identified in the archive of the Institute of Pathology, RWTH Aachen University Hospital, Aachen, Germany. Convolutional neural networks (CNNs) were trained to classify allograft biopsies as normal, rejection, or other diseases. Three times cross-validation (1847 patients) and deployment on an external real-world cohort (101 patients) were used for validation. Area under the receiver operating characteristic curve (AUROC) was used as the main performance metric (the primary endpoint to assess CNN performance).

Findings Serial CNNs, first classifying kidney allograft biopsies as normal (AUROC 0·87 [ten times bootstrapped CI 0·85–0·88]) and disease (0·87 [0·86–0·88]), followed by a second CNN classifying biopsies classified as disease into rejection (0·75 [0·73–0·76]) and other diseases (0·75 [0·72–0·77]), showed similar AUROC in cross-validation and deployment on independent real-world data (first CNN normal AUROC 0·83 [0·80–0·85], disease 0·83 [0·73–0·91]; second CNN rejection 0·61 [0·51–0·70], other diseases 0·61 [0·50–0·74]). A single CNN classifying biopsies as normal, rejection, or other diseases showed similar performance in cross-validation (normal AUROC 0·80 [0·73–0·84], rejection 0·76 [0·66–0·80], other diseases 0·50 [0·36–0·57]) and generalised well for normal and rejection classes in the real-world data. Visualisation techniques highlighted rejection-relevant areas of biopsies in the tubulointerstitium.

Interpretation This study showed that deep learning-based classification of transplant biopsies could support pathological diagnostics of kidney allograft rejection.

Funding European Research Council; German Research Foundation; German Federal Ministries of Education and Research, Health, and Economic Affairs and Energy; Dutch Kidney Foundation; Human(e) AI Research Priority Area of the University of Amsterdam; and Max-Eder Programme of German Cancer Aid.

Introduction

Although kidney transplantation is the most frequently performed solid organ transplantation worldwide, there is a major shortage of organs for transplantation.[1] This shortage renders long-term allograft survival, particularly concerning allograft rejection, an important goal in patient management.

Histopathological assessment of allograft biopsies remains an essential tool in diagnosing organ rejection, thereby guiding the treatment and management of patients who have received a transplant.[2] Despite the international efforts to improve and standardise assessment of kidney allograft pathology within the Banff classification,[3] evaluation of transplant biopsies remains challenging and time-consuming. Although the Banff classification provides a scoring system for allograft pathology, this approach remains semiquantitative and subjective, and is often confounded by interobserver variability.[4] Also, pathology is experiencing a decrease in workforce as fewer young physicians aspire to become

Lancet Digit Health 2022; 4: e18–26

Published Online November 15, 2021 https://doi.org/10.1016/S2589-7500(21)00211-9

See Comment page e2

*Joint first authors

†Joint senior authors

Department of Pathology (J Kers PhD, F Fontana PhD, A A Abiola MD, R Hofstraat MSc, Prof G L Corthals PhD, J J T H Roelofs PhD, Prof S Florquin MD), **Center for Experimental and Molecular Medicine** (H Peters-Sengers PhD), **and Renal Transplant Unit** (Prof F J Bemelman MD, A S Nurmohamed PhD), **Amsterdam UMC, University of Amsterdam, Amsterdam, Netherlands; Department of Pathology, Leiden Transplant Center, Leiden University Medical Center, Leiden, Netherlands** (J Kers); **Van 't Hoff Institute for Molecular Sciences, University of Amsterdam, Amsterdam, Netherlands** (J Kers, R Hofstraat, Prof G L Corthals); **Institute of Pathology** (R D Bülow MD, B M Klinkhammer PhD, S Djudjaj PhD, S von Stillfried MD, D L Hölscher, Prof P Boor PhD), **Department of Nephrology and Immunology** (Prof J Floege MD, Prof P Boor), **and Department of Medicine III** (J N Kather MD), **RWTH Aachen University Hospital, Aachen, Germany; Department of Pathology** (G E Breimer PhD, T Q Nguyen PhD) **and Department of Nephrology and Hypertension** (T T Pieters MD, A D van Zuilen PhD), **University Medical Center Utrecht, Utrecht, Netherlands; Nephrology and Dialysis Unit,**

University Hospital of Modena, Modena, Italy (F Fontana); **Department of Morbid Anatomy and Forensic Medicine, Obafemi Awolowo University Teaching Hospitals Complex, Ile-Ife, Nigeria** (A A Abiola); **Department of Nephrology and Renal Transplantation, University Hospitals Leuven, Leuven, Belgium** (Prof M Naesens MD); **Department of Microbiology, Immunology, and Transplantation, KU Leuven, Leuven, Belgium** (Prof M Naesens); **Pathology and Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK** (J N Kather); **Medical Oncology, National Center for Tumor Diseases, University Hospital Heidelberg, Heidelberg, Germany** (J N Kather)

Correspondence to:
Prof Peter Boor, Institute of Pathology, RWTH Aachen University Hospital, Aachen 52074, Germany
pboor@ukaachen.de

or

Dr J Kers, Department of Pathology, Amsterdam UMC, University of Amsterdam, Amsterdam 1105 AZ, Netherlands
j.kers@amsterdamumc.nl

See Online for appendix

Research in context

Evidence before this study

We searched Pubmed and Web of Science on Jan 1, 2021, using the terms "Deep Learning" OR "Machine Learning" AND "Kidney" AND "Transplantation" from database inception to Dec 31, 2020, with no language restrictions. Our search yielded 78 results; additionally, we used the bibliographies of the retrieved articles for literature review. We found several studies investigating machine or deep learning for kidney diseases using clinical parameters and some studies on segmentation of kidney histology images. However, we did not identify any studies that investigated the diagnostic classification of kidney allograft histopathology.

Added value of this study

To our knowledge, this is the first study to investigate deep learning-based classification of kidney allograft histopathology, particularly focusing on transplant rejection, based on whole slide images of allograft biopsies alone. This is the largest retrospective multicentre study to date to analyse the potential of deep learning in kidney transplant pathology. This study also provided an assessment of the models' predictions using several visualisation techniques.

Implications of all the available evidence

The developed deep learning-based models could serve as a basis to create a decision support system for pathologists, augmenting kidney transplant diagnostics. Such a system could potentially reduce assessment subjectivity and variation of kidney allograft histopathology.

pathologists.[5] Tools that assist in transplant histopathology diagnostics, potentially providing reproducible quantitative data, could be one approach to tackle these challenges. The ongoing digital transformation of pathology enables the effective application of artificial intelligence and particularly deep learning, showing great promise for such tools to be used in the near future.[6–8]

Convolutional neural networks (CNNs) are a specific type of deep learning neural network that are particularly suited for image analysis and computer vision.[9] CNNs have widely been applied in image-based medical diagnostics, particularly in radiology,[10,11] and also increasingly in oncological surgical pathology.[12–14] CNNs can extract subtle patterns from cancer histopathology images, detecting molecular subclasses of tumours in an end-to-end way—ie, training CNNs directly on raw image data without manually defining intermediate steps.[12,15] Currently, the very few applications of deep learning in transplant histopathology and nephrology mostly focus on automated semantic segmentation of histology into different histological compartments.[16–18] Additionally, prediction tools for allograft loss[19] and molecular archetypal analysis of kidney allograft rejection[20] have been developed. However, no end-to-end deep learning biomarkers are yet available in transplant histopathology.

We aimed to develop and validate CNNs for automated preclassification of kidney allograft biopsies using digital biopsies and their pathologist-derived diagnoses as ground truth, and to analyse the potential utility of these CNNs as a biopsy triage system focusing on transplant rejection.

Methods

Study design and participants

We performed a retrospective, multicentre, proof-of-concept study and identified diagnostic kidney allograft biopsy samples by a database search in the Departments of Pathology of the Amsterdam UMC, Amsterdam, Netherlands (n=1130) and the University Medical Center Utrecht, Utrecht, Netherlands (n=717; appendix pp 3–5). 101 consecutive kidney transplant biopsies were identified in the archive of the Institute of Pathology, RWTH Aachen University Hospital, Aachen, Germany (figure 1, appendix p 5). All three institutions are kidney transplant centres. Physical glass slides (produced from formalin-fixed paraffin-embedded tissue) of one periodic acid Schiff (PAS), one haematoxylin and eosin (H&E), and one Jones silver stain per kidney biopsy were digitised with identical scan resolutions (appendix pp 3–6). 5844 digital whole slide images from 1948 kidney transplant biopsy samples were tessellated into image tiles used to either train a single CNN or two serial CNNs (figure 1A). All experiments were done in accordance with the Declaration of Helsinki and were approved by the local ethics and privacy committees (Amsterdam 19.260; Utrecht 19.482; Aachen EK315/19). The need for informed consent was waived by the local ethics and privacy committees.

Ground truth

Allograft biopsy samples were centrally assigned to the classes normal (Banff category 1); rejection, comprised of antibody-mediated rejection, T-cell-mediated rejection, and mixed rejection (Banff categories 2–4), including cases where there was suspicion of antibody-mediated rejection and borderline T-cell-mediated rejection; or other diseases (Banff category 5), on the basis of the 2019 update of the Banff criteria,[21] by an experienced transplant nephropathologist who assessed the Banff lesions and final diagnoses within the pathology reports (appendix p 26). Samples with another diagnosis in addition to rejection were classified as rejection because we aimed to potentially identify all biopsies with histological signs of rejection. Pathological

diagnoses in the written reports were a consensus of three (at Amsterdam UMC) or two nephropathologists (at Utrecht and Aachen), after discussion at weekly multidisciplinary nephrology-pathology consensus meetings at the respective tertiary expertise centre for kidney transplantation. The additional class of disease, used for experiments with the serial CNN, included all samples classified as rejection and other diseases.

Deep learning analyses

Details on generation of tiles and their preprocessing, CNN training, and CNN performance analyses and visualisation techniques are provided in the appendix (pp 6–8). The primary endpoint to assess CNN performance was the area under the receiver operating characteristic curve (AUROC) with ten times bootstrapped CIs. Additionally, we show precision-recall curves, and report the area under the precision-recall curve (AUPRC).

The models were trained on the Amsterdam and Utrecht cohorts (figure 1B) and validated on the external real-world Aachen cohort (figure 1C). The two serial CNNs were trained to distinguish the classes normal (Banff category 1) versus disease (all other Banff categories) by the first CNN, followed by the second CNN to distinguish rejection (Banff categories 2–4) versus other diseases (including Banff category 5) in the disease class only (figure 1D). The single CNN was trained to distinguish in parallel the classes normal, rejection, and other diseases (figure 1E). A complete list of included histopathological diagnoses within the classes is given in the appendix (pp 24–25). Initially, we investigated which histological stains (H&E, PAS, or Jones silver) should be used for optimal performance. The highest performance, assessed by mean AUROC, was achieved using all stains combined (figure 1A). Therefore, we used all stains for all following analyses. Next, we investigated several CNN architectures (ie, ResNet18, ResNet50, ResNet101, ShuffleNet, and Inceptionv3) for their respective performance. Although there were almost no differences in mean AUROCs achieved in three times internal cross-validation on the Amsterdam cohort, we could observe differences in the generalisability of the models when deployed on the Utrecht cohort, to which the CNNs had not been exposed previously. The three ResNets achieved the highest mean AUROCs on the Amsterdam cohort, but did not generalise as well as the Inceptionv3 architecture (appendix p 9). Prioritising generalisability, we used the Inceptionv3 architecture throughout the rest of the study.

As T-cell-mediated rejection and BK polyomavirus nephropathy can be particularly difficult to differentiate given the largely overlapping pathology, we investigated network confusion for biopsies with these diagnoses. Given the diagnostic importance of cortical kidney tissue, we evaluated the performance on the basis of the amount of cortical tissue.

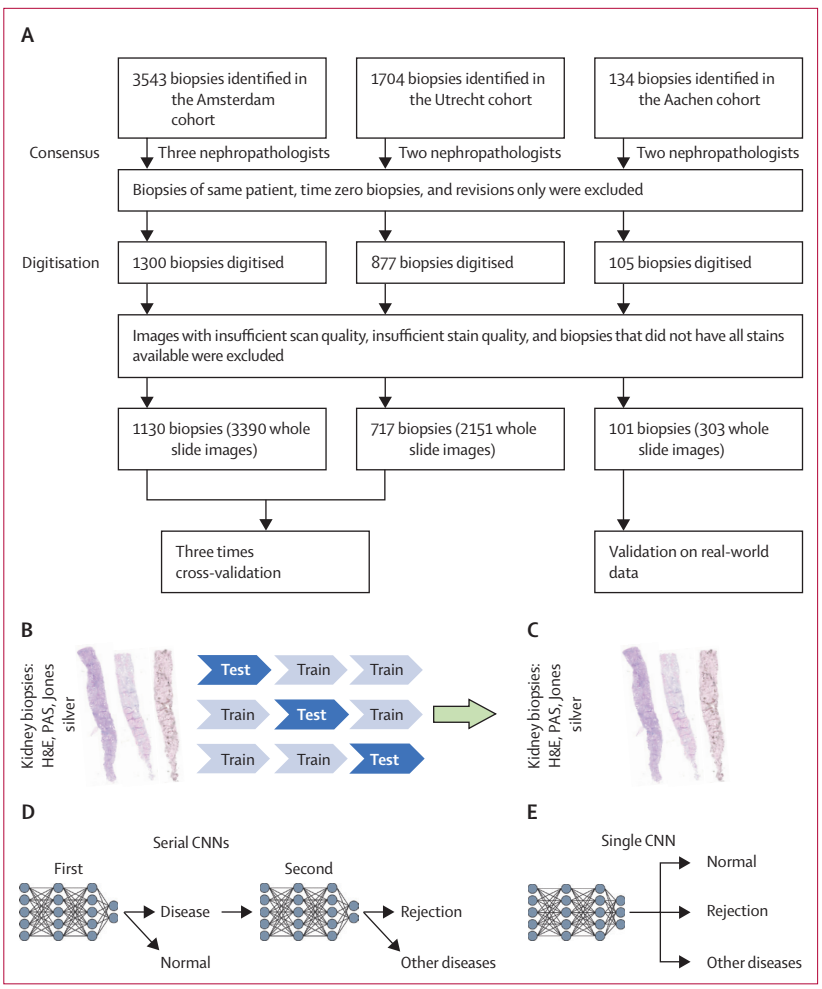

Figure 1: **Flowchart of the generation of the allograft biopsy cohorts and study plan**
A database search was performed in all centres to identify all kidney allograft biopsies and samples were excluded on the basis of predefined criteria. The remaining biopsies were digitised, followed by a manual quality check resulting in the exclusion of samples with insufficient quality for further processing. The final number of samples and whole slide images are shown at the bottom of the flowchart (A). The Amsterdam and Utrecht cohorts were used for model development, and performance in these cohorts was assessed in three times cross-validation (B). The Aachen cohort was used as an unseen external validation set to assess generalisability (C). This split of cohorts was used in two approaches: a series of two CNNs first classifying biopsies into normal and disease and then classifying samples in the disease class into rejection or other diseases (D), and one single network classifying biopsies into normal, rejection, and other diseases (E). H&E=haematoxylin and eosin. PAS=periodic acid Schiff. CNN=convolutional neural network.

To visualise the basis of classification on the individual most predictive tiles of each class, we applied Occlusion Sensitivity[22]and gradient-weighted Class Activation Mapping (gradCAM).[23]

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

A single Inceptionv3 CNN trained on kidney allograft pathology (on the Amsterdam and Utrecht samples

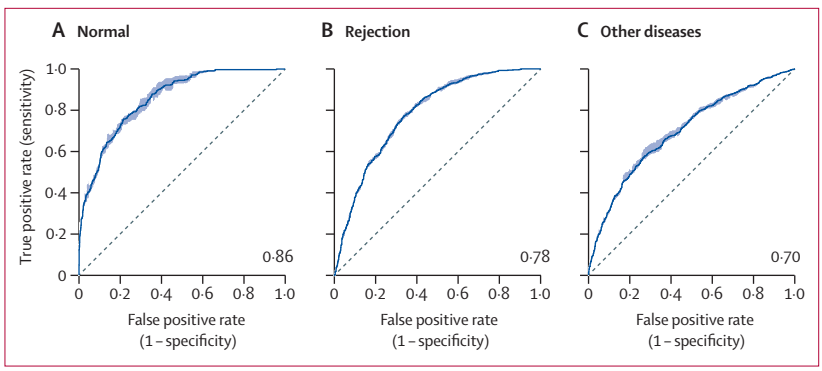

Figure 2: **Single CNN performance evaluated in three times cross-validation**
Patient-level receiver operating characteristic curves for the normal (A), rejection (B), and other diseases (C) classes. In this analysis, the cohorts from Amsterdam and Utrecht were used in a combined fashion resulting in 347 samples classed as normal, 664 classed as rejection, and 836 classed as other diseases (total n=1847). CNN=convolutional neural network.

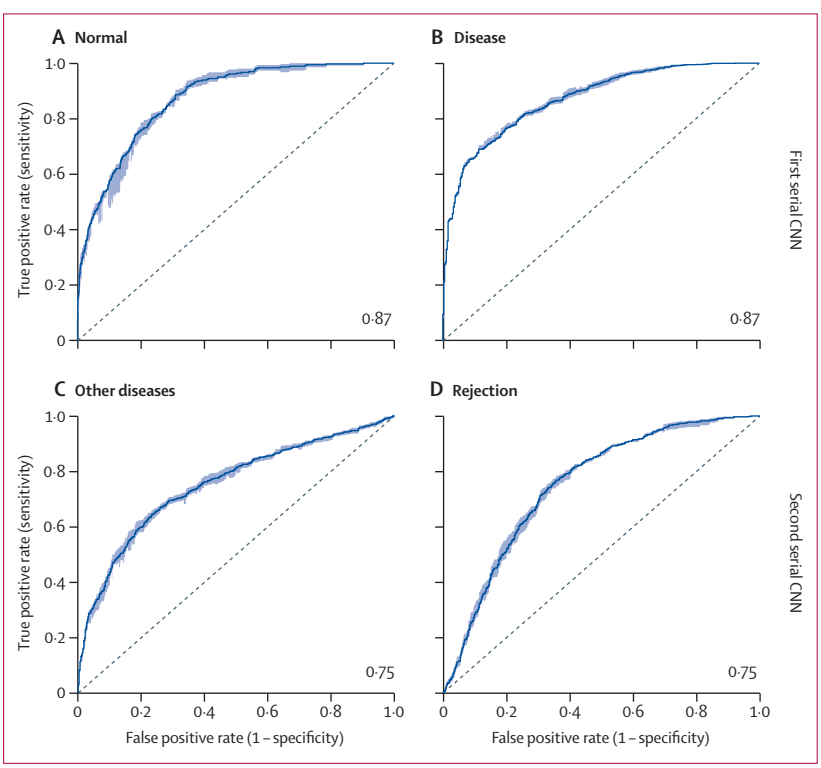

Figure 3: **Serial CNN performance evaluated in three times cross-validation**
Patient-level receiver operating characteristic curves for the normal (A) and disease (B) classes (normal n=347, disease n=1500, total n=1847), as performed by the first serial model, and for the other diseases (C) and rejection (D) classes (other diseases n=836, rejection n=664, total n=1500), as performed by the second serial model. In this analysis, the cohorts from Amsterdam and Utrecht were used in a combined fashion. CNN=convolutional neural network.

combined) achieved AUROCs of 0·86 (ten times bootstrapped CI 0·85–0·87) for the normal class, 0·78 (0·77–0·79) for the rejection class, and 0·70 (0·68–0·72) for the other diseases class (figure 2).

Next, we trained two CNNs (on the Amsterdam and Utrecht samples combined), first classifying biopsies into normal (AUROC 0·87 [CI 0·85–0·88]) and disease (0·87 [0·86–0·88]; figure 3A, B), and then classifying the biopsies classified as disease into rejection (0·75 [0·73–0·76]) or other diseases (0·75 [0·72–0·77]; figure 3C, D). Precision-recall curves and the mean AUPRC for all models showed a good trade-off between precision and recall, especially for classifying biopsies as normal (appendix p 10). A confusion matrix comparing pathologist and single CNN-derived classes showed confusion primarily between the rejection and other diseases classes (appendix p 11). A large number of BK polyomavirus nephropathy cases were misclassified as rejection (appendix p 12). The analyses of all pathological diagnoses of the misclassified cases showed that all rejection types were misclassified by the single CNN (appendix p 27).

In the evaluation of performance on the basis of the amount of cortical tissue, the performance was similar between subgroups with varying glomerular numbers (appendix pp 13, 28).

In the model validation, when deploying the single CNN on external kidney allograft biopsies (Aachen samples) similar AUROCs (0·80 [ten times bootstrapped CI 0·73–0·84] for the normal class, 0·76 [0·66–0·80] for the rejection class, and 0·50 [0·36–0·57] for the other diseases class) were achieved, showing good generalisability for the normal and rejection classes (figure 4). However, no generalisability was found for the diverse class of other diseases. Visualisation of the basis of the single CNN model's predictions revealed that in the normal and rejection classes, large areas of the biopsy core were highly predictive, but only a smaller amount of tiles were highly predictive in the other diseases group (figure 4). The different distributions of highly predictive tiles is likely to reflect focal pathology within one biopsy and heterogeneity of phenotypes within the respective classes. We next investigated whether such prediction maps point to relevant diagnostic regions for rejection. We extracted highly predictive regions from whole slide images correctly classified as rejection (appendix p 14), which showed interstitial inflammation, tubulitis, and peritubular capillaritis, all of which are diagnostic lesions of the Banff classification for kidney allograft pathology.

In the normal class, tubular cross-sections, thin interstitium, and normal peritubular capillaries were highlighted using both Occlusion Sensitivity and gradCAM (figure 4C, appendix p 15). For the rejection class, interstitial lymphocytic infiltrates and injured tubuli were important (figure 4F, appendix p 15). In the most predictive tile of the other diseases class, altered tubulointerstitium and intratubular material were important for the prediction (figure 4I, appendix p 15).

When we deployed both serial CNNs on the Aachen cohort, the first serial CNN achieved AUROCs of 0·83 (ten times bootstrapped CI 0·80–0·85) for the normal class and 0·83 (0·73–0·91) for the disease class (figure 5), also indicating good generalisability. The second serial CNN generalised less well with AUROCs of 0·61 (0·50–0·74) for the other diseases class and 0·61

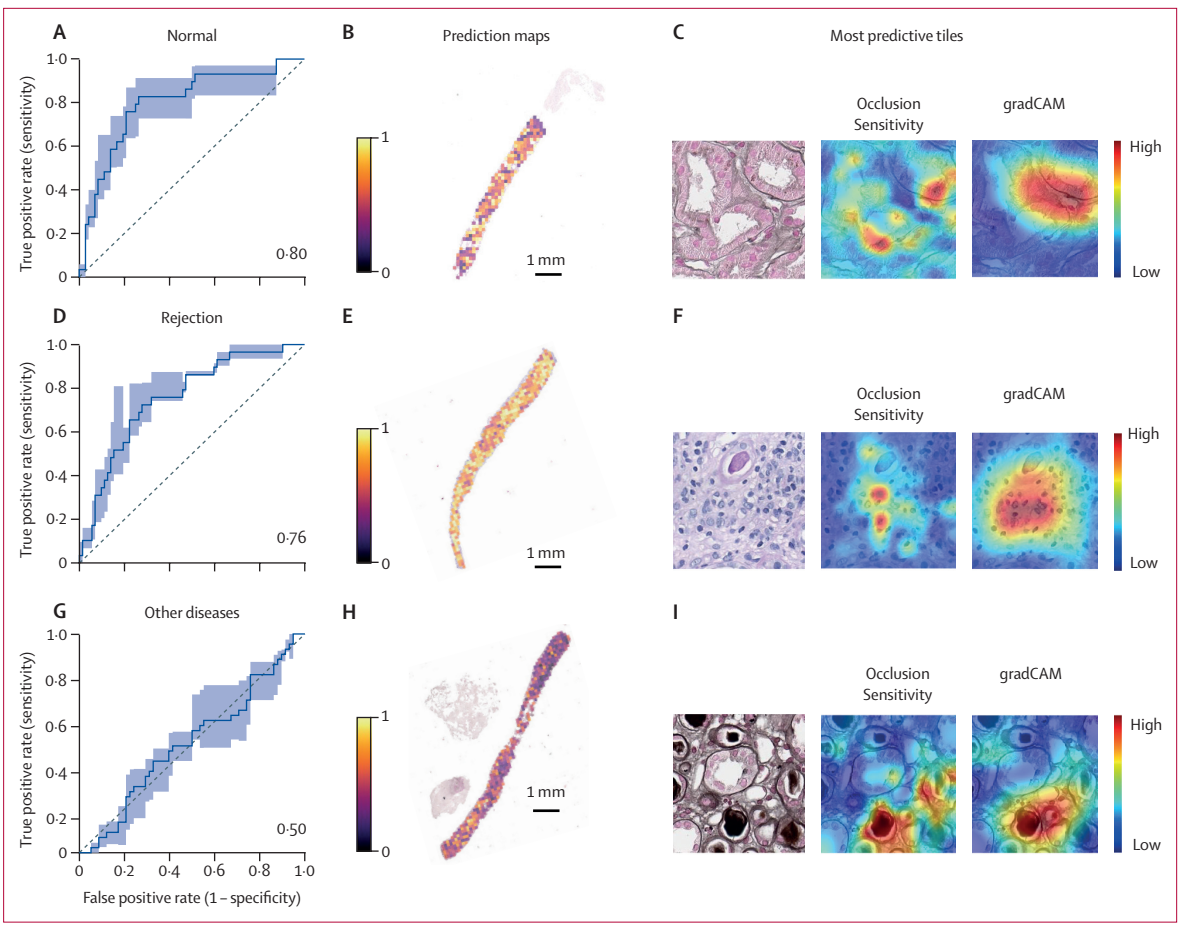

Figure 4: **Single CNN performance in the external Aachen cohort**
Patient-level receiver operating characteristic curves for the normal class (A), rejection class (D), and other diseases class (G). (B) Representative prediction map for the normal class mapping the predictive value of respective tiles to their parent whole slide image. Tile-level visualisations of important image areas of the normal class (C), rejection class (F), and other diseases class (I), using Occlusion Sensitivity and gradCAM on the most predictive tile. Representative prediction maps for the rejection class (E) and other diseases class (H). 29 samples were in the normal class, 43 in the other diseases class, and 29 in the rejection class (total n=101). Tile edge length is 128 µm for each tile. CNN=convolutional neural network. gradCAM=gradient-weighted Class Activation Mapping.

(0·51–0·70) for the rejection class (figure 5). Highly predictive tiles seem to be distributed across the entire biopsy cores for all classes. When using Occlusion Sensitivity and gradCAM, thin interstitium and empty peritubular capillaries were highlighted (figure 5C, appendix p 16). The most predictive tile of the disease class showed prominent immune cell infiltrates, some of which were particularly important for classifying this tile as diseased (figure 5F, appendix p 16). The most predictive tile and visualisation of the class other diseases showed interstitial fibrosis and simplified tubular epithelial cells, in line with Banff category 5 included in this class (figure 5I, appendix p 16). In the rejection class, the most predictive tile showed prominent interstitial immune cell infiltration. Only some immune infiltrates were highlighted using Occlusion Sensitivity and gradCAM; however, no obvious visual distinction was possible between important and irrelevant immune infiltrates (figure 5L, appendix p 16). Precision-recall curves and the mean AUPRC for the models deployed on the Aachen cohort showed generalisability of all classes other than the diverse class of other diseases (appendix p 17).

Various applications of the CNNs could be envisioned in a digitised pathology workflow. This assumption is supported by the short time required for the scanning (approximately 2 min per slide) and inference of the models (approximately 10 min using standard hardware for three stains). The most predictive tiles of a patient could potentially be used to facilitate kidney allograft diagnostics by a pathologist—eg, by using the 81 most predictive tiles (9×9 matrix) from correctly classified samples of each class (appendix pp 18–20). Using the optimal operating threshold (0·68) of the first serial CNN resulted in a mean sensitivity of 95·03% (ten times bootstrapped CI 93·80–95·52) and mean specificity of 45·39% (43·27–49·86) for detection of the disease class, with a mean positive predictive value of 87·91%

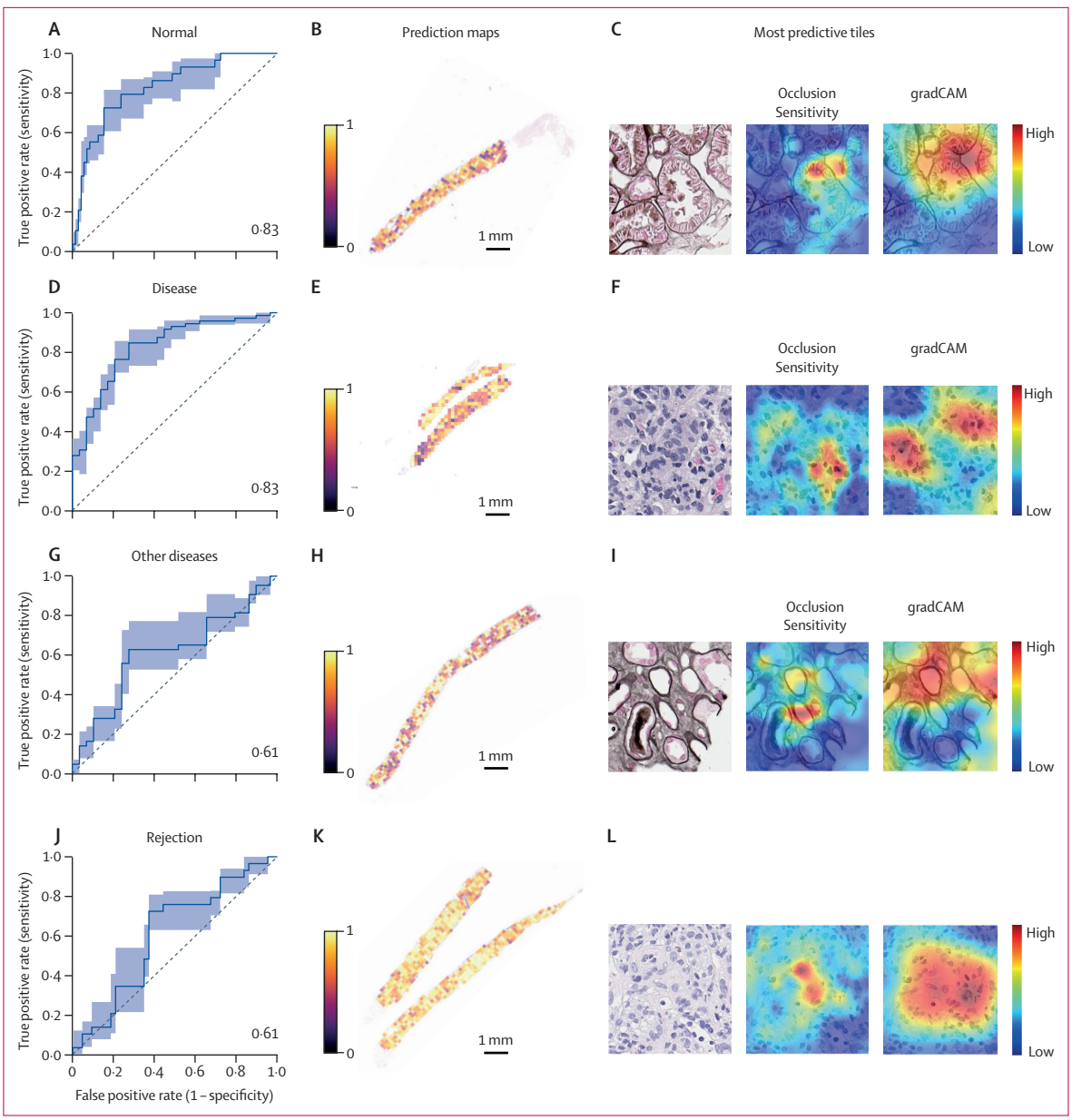

Figure 5: **Serial CNN performance in the external Aachen cohort**

Patient-level receiver operating characteristic curves for the normal (A) and disease (D) classes (disease n=72, normal n=29, total n=101), and for the other diseases (G) and rejection (J) classes (other diseases n=43, rejection n=29, total n=72). (B) Representative prediction map for the normal class mapping the predictive value of respective tiles to their parent whole slide image. (C) Visualisation of the basis for prediction of the class normal using Occlusion Sensitivity and gradCAM on the most predictive tile. Representative prediction maps for the disease class (E), other diseases class (H), and rejection class (K). Tile-level visualisations of important image areas of the disease class (F), other diseases class (I), and rejection class (L), using Occlusion Sensitivity and gradCAM on the most predictive tile. Tile edge length is 128 µm for each tile. CNN=convolutional neural network. gradCAM=gradient-weighted Class Activation Mapping.

(86·33–88·52) and mean negative predictive value of 67·86% (59·03–74·37; appendix p 21).

Discussion

Kidney allograft pathology diagnostics is essential to guide the treatment of patients who receive transplants. However, it is also one of the most challenging fields in diagnostic pathology that could strongly benefit from supportive systems to augment the diagnostic process. In this study, we showed that even in this highly complex use case, deep learning-based classification and visualisation of kidney allograft biopsies could potentially provide a useful diagnostic support system for pathologists. We focused on the classification of the three main overarching classes in kidney transplants (ie, normal, rejection, and other diseases). This approach

provides a basis for future studies to develop specialised models for classification of rejection and the numerous other allograft diseases. We assume that such classification algorithms require specific dedicated models, which could be serially switched after using more general models as proposed here. Despite the already large datasets used, the current performance of these deep learning-based models is not sufficient to replace human pathologists. A histological report contains much more information than the pure diagnostic class (ie, the extent of the canonical Banff lesions). Instead of replacing human pathologists, we argue that these models have the potential to improve pathologists' performance; eg, by using visualisation techniques, in the sense of augmented intelligence and improved digital pathology workflows, but this remains to be evaluated.

Most current deep learning studies in histology, particularly in cancer, only use a single stain per patient (most commonly H&E). For correct diagnostics in kidney allograft pathology, several stains are needed. The use of several stains helps nephropathologists to detect specific diagnostic features more easily and provide correct diagnoses. Accordingly, our CNNs performed better when different stainings were used for training instead of using a single stain. This finding might be because many kidney allograft pathologies are focal—ie, only observed in a given area and perhaps in only one section. Regarding focality, the inclusion of available consecutive H&E or PAS slides, or both, might be an interesting approach to further improve the performance. Another hypothesis is that each stain might contain stain-specific features relevant for the discrimination of the classes. This hypothesis suggests that even in the case of deep learning-augmented allograft diagnostics, several histological stainings and biopsy sections would be required, but the use of deep learning can point the pathologist to the diagnostically most predictive areas and thereby potentially save time.

Distinguishing between normal and diseased biopsies was the easiest task for the CNNs, given that the morphological difference between these classes is the highest. Although the performance was largely similar between the single and the serial CNNs, the single CNN did not generalise well in the external cohort when classifying biopsies with other diseases. This relatively poor generalisability using the single CNN is likely to be because the other diseases class contains many morphologically diverse diseases. Thus, serial CNNs could be useful for specific classification tasks, particularly of very rare pathological features or diseases (eg, BK polyomavirus nephropathy) that have been preclassified (and preselected on the basis of preclassification) by a more general CNN; ie, a general CNN could coarsely classify biopsies and more specialist CNNs can determine the specific features or diseases. The CNNs could be used as prioritisation tools, to triage biopsies that should be assessed with higher urgency and to sort out completely healthy samples for later assessment. In particular, prioritising biopsies with rejection, which often requires fast initiation of treatment, would be helpful.

Introduction of deep learning-based technologies into clinical practice might be hindered by the low explainability of the models, the so-called black box phenomenon.[24] Here, we investigated the localisation of model predictions both at the slide and tile level as one approach of explainability. Remarkably, the most predictive tiles showed some canonical lesions of their respective classes. Visualisation techniques in the most relevant tiles mainly identified alterations of the tubulointerstitium, most of which correlated well with typical pathological findings (eg, healthy tubules for the normal class, or tubulitis or capillaritis for the rejection class). Interstitial fibrosis and tubular atrophy was the leading pathology identified by the visualisation techniques in Banff category 5 for the other diseases class, which was in line with the common development of fibrosis in allograft biopsies. On the other hand, this finding could also be viewed as an unbiased confirmation of some diagnostic criteria (a subset of the Banff criteria) set out by transplant pathologists. However, some important histological lesions were largely missing from highly predictive tiles (and thus also from our tile-level visualisations), particularly those in glomeruli and large arteries. Grouping the biopsies on the basis of glomerular count resulted in only minor differences or no differences in model performance. This analysis indicated that for the given task, glomeruli were not of high importance for the CNNs trained in this study. This low importance might be because the tubulointerstitium represents the majority of tissue, thereby being strongly over-represented in the image tiles used for training, by comparison with tiles containing glomeruli and arteries. Additionally, for the identification of the rejection class, tubulointerstitium contains lesions for both T-cell-mediated rejection and antibody-mediated rejection, which is not the case for glomeruli for example (only antibody-mediated rejection). Misclassified rejection samples included all types of rejection and were most often classified as other diseases, and were only rarely classified as normal.

Visualisation approaches could be used to augment biopsy diagnostics; eg, by pinpointing the focus of pathologists on highly predictive areas (tiles) of a biopsy for rejection, or providing a matrix of most predictive tiles. This augmentation might be particularly useful in non-tumour pathology, in which pathological alterations are not seldom focal and searching for these can be time-consuming. Future approaches using biopsies that have been presegmented into different microanatomic compartments and training specific CNNs for each compartment might potentially improve accuracy and explainability.[16–18] Also, a model solely trained on glomerular images from rejection samples might show higher performance for identification of antibody-mediated rejection.

This study has some limitations. All participating institutions are from western Europe and we did not include data on ethnicity, since collection of this kind of information is illegal in the Netherlands and not routinely performed in Germany. Additionally, some baseline and demographic characteristics, such as the initial nephropathy or HLA mismatches are not available, which might limit transferability of the results to other cohorts. However, this information is not needed for histopathological diagnosis of kidney allograft diseases.[3,21] Before implementation in clinical practice, extending the training datasets to additional centres might lead to the development of more robust models in the future. Variability in staining from only two centres was included in the training set. Although the stainings look different in each centre, our models performed well when deployed on previously unseen data, with different staining appearances and using different whole slide image scanners. The performance concerning staining variability could further improve by adding more data and centres, as well as by deploying additional forms of data augmentation (eg, with generative models).[25] Our focus was not to build a full diagnostic deep learning algorithm that could replace the pathologist, but rather develop a support system that could augment human pathologists. Therefore, we did not differentiate between different types of rejection (ie, T-cell-mediated rejection or antibody-mediated rejection, or mixed).[3] Particularly for antibody-mediated rejection, additional data are required for correct diagnosis; eg, presence of donor-specific antibodies, but also C4d-positive staining of peritubular capillaries. These data are not always available at the time of diagnosis. Such data integration is easy for human pathologists, so our approach extends the idea of computer–human interactions and augmented intelligence. We anticipate that, in the future, such multimodal classification algorithms could be trained similarly to in this study on a multipathologist consensus diagnostic class vote as a surrogate gold standard or alternatively, the archetypal molecular diagnostic class.[20] Combining such molecular classification systems with histological deep learning analyses might lead to a more reproducible and granular classification system of allograft pathologies. Unfortunately, sufficiently large datasets are currently not available and will require considerable efforts from the transplant community in order to develop these. To automatically assess the canonical lesions of the Banff classification, combinations of CNNs performing instance segmentation (detection and delineation of each individual object in an image) of relevant compartments (eg, capillaries, tubuli, vessels, and glomeruli)[16–18] with models detecting individual inflammatory cells would be of high interest. Some tiles containing medulla were found within this 9×9 matrix in a sample classified as other diseases. A possible explanation for this could be that the CNNs might confuse patches of medulla for interstitial fibrosis, since medulla naturally contains more extracellular matrix. Another example of implementation into the clinical workflow in digital pathology could include initial filtering and prioritising of biopsies with pathological changes for the pathologist. Potentially, models trained and analysing exclusively the cortical tissue could overcome this problem. Another limitation is the retrospective design of our study. However, before prospective trials can be done, studies such as ours are required to understand the potential and feasibility of such deep learning-based support systems and identify potential pitfalls. Retrospective studies will be required to prepare for the successful design of prospective trials because, for example, there is currently no accepted way to calculate sample size for the development of deep learning algorithms.

In conclusion, our findings suggest the feasibility of using CNNs for automated and reproducible preclassification of kidney allograft biopsies, potentially augmenting allograft biopsy diagnostics by computer–human interaction.

Contributors

JK, RDB, JNK, and PB conceived and designed the study. JK, TQN, SD, SvS, GEB, FF, AAA, RH, GLC, HP-S, TTP, ADvZ, FJB, ASN, MN, JF, DLH, and RDB collected and digitised patient cohorts. JK, TQN, JJTHR, SF, and RDB assigned patients to classes. RDB and BMK performed annotations and quality controls. RDB and JNK performed deep learning analyses and visualisations. JK, RDB, JNK, and PB wrote the initial draft of the manuscript. JK and RDB created the figures. All authors subsequently read and revised the manuscript and read and approved the final version. All authors had access to all the data in the study and JK, JNK, RDB, and PB verified the data and had final responsibility for the decision to submit for publication.

Declaration of interests

JNK reports consulting roles for Owkin France and Panakeia (UK), outside of the submitted work; and honoraria for lectures from Merck Sharp & Dohme and Eisai and honoraria for participation in advisory board meetings of Merck Sharp & Dohme and Bayer, outside of the submitted work. All other authors declare no competing interests.

Data sharing

For deep learning analysis we used our in-house open source codes which have been described previously.[13,24] All code used was run using MATLAB R2019b and MATLAB R2020a and is freely accessible, with user instructions, at https://github.com/jnkather/DeepHistology. Slide-level heatmaps were generated using QuPath version 0.2.0; tile-level visualisations using Occlusion Sensitivity and gradCAM were performed using MATLAB R2020a. Whole slide images cannot be made publicly available due to regulatory reasons. Models and data will be made available to interested research partners on reasonable request to the corresponding author; the prerequisite for this is a data transfer agreement, approved by the legal departments of the requesting researcher and by all legal departments of the institutions that provided data for the study, and an ethics clearance.

Acknowledgments

This study was funded by the German Research Foundation (Project numbers 322900939, 454024652, 432698239 to PB, 432698239 to SD), the European Research Council (Consolidator Grant AIM.imaging.CKD, number 101001791 to PB), the German Federal Ministries of Education and Research (STOP-FSGS-01GM1901A to PB and SD), Health and Economic Affairs and Energy (DEEP LIVER number ZMVI1-2520DAT111 to PB and JNK; EMPAIA number 01MK2002A to PB), the Dutch Kidney Foundation (17OKG23: DEEPGRAFT project to JK), the Human(e) AI Research Priority Area of the University of

Amsterdam (to JK), and the Max-Eder Programme of German Cancer Aid (grant number 70113864 to JNK).

References

1 Bastani B. The present and future of transplant organ shortage: some potential remedies. *J Nephrol* 2020; **33:** 277–88.

2 Nankivell BJ, Kuypers DR. Diagnosis and prevention of chronic kidney allograft loss. *Lancet* 2011; **378:** 1428–37.

3 Roufosse C, Simmonds N, Clahsen-van Groningen M, et al. A 2018 reference guide to the Banff classification of renal allograft pathology. *Transplantation* 2018; **102:** 1795–814.

4 Marcussen N, Olsen TS, Benediktsson H, Racusen L, Solez K. Reproducibility of the Banff classification of renal allograft pathology. Inter- and intraobserver variation. *Transplantation* 1995; **60:** 1083–89.

5 Robboy SJ, Weintraub S, Horvath AE, et al. Pathologist workforce in the United States: I. Development of a predictive model to examine factors influencing supply. *Arch Pathol Lab Med* 2013; **137:** 1723–32.

6 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019; **25:** 44–56.

7 Boor P. Artificial intelligence in nephropathology. *Nat Rev Nephrol* 2020; **16:** 4–6.

8 Becker JU, Mayerich D, Padmanabhan M, et al. Artificial intelligence and machine learning in nephropathology. *Kidney Int* 2020; **98:** 65–75.

9 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521:** 436–44.

10 Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer* 2018; **18:** 500–10.

11 Chilamkurthy S, Ghosh R, Tanamala S, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *Lancet* 2018; **392:** 2388–96.

12 Kather JN, Heij LR, Grabsch HI, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat Can* 2020; **1:** 789–99.

13 Bulten W, Pinckaers H, van Boven H, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol* 2020; **21:** 233–41.

14 Pantanowitz L, Quiroga-Garza GM, Bien L, et al. An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. *Lancet Digit Health* 2020; **2:** e407–16.

15 Echle A, Grabsch HI, Quirke P, et al. Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning. *Gastroenterology* 2020; **159:** 1406–16.e11.

16 Hermsen M, de Bel T, den Boer M, et al. Deep learning-based histopathologic assessment of kidney tissue. *J Am Soc Nephrol* 2019; **30:** 1968–79.

17 Jayapandian CP, Chen Y, Janowczyk AR, et al. Development and evaluation of deep learning-based segmentation of histologic structures in the kidney cortex with multiple histologic stains. *Kidney Int* 2021; **99:** 86–101.

18 Bouteldja N, Klinkhammer BM, Bülow RD, et al. Deep learning-based segmentation and quantification in experimental kidney histopathology. *J Am Soc Nephrol* 2021; **32:** 52–68.

19 Loupy A, Aubert O, Orandi BJ, et al. Prediction system for risk of allograft loss in patients receiving kidney transplants: international derivation and validation study. *BMJ* 2019; **366:** l4923.

20 Reeve J, Böhmig GA, Eskandary F, et al. Assessing rejection-related disease in kidney transplant biopsies based on archetypal analysis of molecular phenotypes. *JCI Insight* 2017; **2:** 94197.

21 Loupy A, Haas M, Roufosse C, 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* 2020; **20:** 2318–31.

22 Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. *arXiv* 2013; published online Nov 12. http://arxiv.org/abs/1311.2901 (preprint).

23 Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis* 2020; **128:** 336–59.

24 Wang F, Kaushal R, Khullar D. Should health care demand interpretable artificial intelligence or accept "black box" medicine? *Ann Intern Med* 2020; **172:** 59–60.

25 Krause J, Grabsch HI, Kloor M, et al. Deep learning detects genetic alterations in cancer histology generated by adversarial networks. *J Pathol* 2021; **254:** 70–79.