Structured description of the MONKEY challenge

SUMMARY

Item 1: Title

Machine learning for optimal detection of inflammatory cells in the kidney (MONKEY)

Item 2: Abstract

The Banff classification is the standard for histopathologic assessment of transplant kidney biopsies. It consists of 17 Banff Lesion Scores (BLS), 10 of which focus on the presence and extent of inflammatory cells in different kidney compartments. Most BLS are graded semi-quantitatively as mild, moderate, and severe based on the number of inflammatory cells within the corresponding compartment. As the diagnosis and subsequent treatment decision depend on the result of the different BLS, it is of utter importance that the assessment of these individual BLS is objective and consistent.

However, in daily practice, the reproducibility of Banff scoring is mediocre and time-consuming. Therefore, the development of automated biopsy assessment holds great potential to reduce pathologists' workload and increase scoring consistency. Three years ago, the Computational Pathology Group (CPG) of Radboudumc (Nijmegen, Netherlands) started the DIAGGRAFT project. Funded by the Dutch Kidney Foundation, the project aims to develop, validate, and implement deep learning for objective and reproducible assessment of histopathological features of renal allograft pathology according to the Banff Classification (Banff Lesion Scores) in renal transplant biopsies.

A previously developed segmentation algorithm by Hermsen et al. already enables us to segment kidney tissue in the different compartments with weighted mean Dice coefficients for all classes of 0.80 and 0.84. For the automatic assessment of BLS, additional algorithms are required to correctly detect/segment lymphocytes and monocytes, according to the Banff Classification. In a previous study, we developed an AI model for lymphocyte detection in immunohistochemically stained tissue sections, showing an F1-score of 0.78 and the highest agreement with manual evaluation ($\kappa = 0.72$). In contrast, the average pathologist's agreement with the reference standard was $\kappa = 0.64$. This model was trained on 83 glass slides of breast (33 slides), prostate (22 slides), and colon (28 slides) cancer specimens from nine different medical centers in the Netherlands. For BLS scoring in routine diagnostics, we need to develop an AI-model that can detect/segment both lymphocytes and monocytes in Periodic acid-Schiff (PAS) stained kidney transplant biopsies. This will be the focus of the MONKEY challenge.

In further studies, the model resulting from the MONKEY challenge will be integrated within the existing structure segmentation model and validated with pathologists for assessing kidney transplant biopsies according to the Banff Classification to be used in routine diagnostics in the future.

Item 3: Keywords

Kidney transplant biopsies, Cell detection, Inflammation detection

CHALLENGE ORGANIZATION

Item 4: Organizers

Core organization team:

- **Linda Studer**, Department of Pathology, Radboudumc, Nijmegen, The Netherlands
- **Dominique van Midden**, Department of Pathology, Radboudumc, Nijmegen, The Netherlands
- **Prof. Luuk Hilbrans**, Department of Nephrology, Radboudumc, Nijmegen, The Netherlands
- **MD PhD Jesper Kers**, Department of Pathology, Amsterdam University Medical Centers, and Center for Analytical Sciences Amsterdam, Van 't Hoff Institute for Molecular Sciences, University of Amsterdam, Amsterdam, the Netherlands | Department of Pathology, Leiden University Medical Center, Leiden, the Netherlands
- **PhD Fazael Ayatollahi**, Department of Pathology, Radboudumc, Nijmegen, The Netherlands
- **Prof. Jeroen van der Laak**, Department of Pathology, Radboudumc, Nijmegen, The Netherlands

For additional contributors, see the website: https://monkey.grand-challenge.org/organizers/

Provide information on the **primary contact person.**
Linda Studer. Email address: linda.studer@radboudumc.nl
LinkedIn Profile: https://www.linkedin.com/in/linda-studer/

Item 5: Lifecycle type

This challenge will have two cycles.

1. **Challenge cycle:** This is the initial cycle. We will have a development and validation phase with a live leaderboard running for about 4.5 months, followed by a final test phase and an announcement of the winners.
2. **Open submission cycle:** after the announcement of the winners, the challenge will reopen for submissions and be supported for up to 5 years

Item 6: Challenge venue and platform

a) Report the **event** (e.g., conference) that is **associated** with the challenge (if any).
We are applying to become a MIDL-associated challenge.
b) Report the **platform** (e.g., grand-challenge.org) used to run the challenge:
We are using Grand-challenge.org.
c) Provide the **URL** for the challenge website (if any): monkey.grand-challenge.org

Item 7: Participation policies

a) Define the **allowed user interaction** of the algorithms assessed (e.g. only (semi-) automatic methods allowed).
There is no user interaction.
b) Define the policy on the **usage of training data**. The data used to train algorithms

may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.
Additional data and pre-trained networks are allowed. The data sources must be reported, and either the data or the model weights must be publicly available under a permissive license.

c) Define the **participation policy for members of the organizers' institutes**. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.
Members of our institution are allowed to participate like regular contestants but are, of course, not permitted to access any of the test data on our local data shares. Members from the core organizing team are, however, excluded.

d) Define the **award policy**. Provide details with respect to challenge prizes.
The total price money available is 3.250 EUR. It will be divided between the two leaderboards (monocyte and lymphocyte detection vs. mono-nuclear leukocytes (MNL), i.e., combined).

e) **Define the policy for result announcement.**
Per the MIDL guidelines, we will hold a webinar at the end of the challenge, where the winners will be announced and can present their solutions. We will also publish a journal paper regarding our findings. Following the MIDL challenge guidelines, we will encourage all submitting groups to submit their method as a short paper to MIDL 2025.

f) **Define the publication policy**. In particular, provide details on ...
… who of the participating teams/the participating teams' members qualifies as author
… whether the participating teams may publish their own results separately, and (if so)
… whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).
Up to three members of each leaderboard's top three performing teams will be invited to participate in the challenge paper as consortium authors. Participants of the MONKEY challenge and non-participating researchers using the dataset can publish their own results at any time, separately. Challenge participants are encouraged to submit their solution as a short paper at MIDL 2025. Any such publications must cite this document (BIAS preregistration form for the MONKEY challenge), which will be published on Zenodo with a corresponding DOI. Once a study protocol and/or a challenge paper has been published, they are requested to refer to those publication(s) instead.

Item 8: Submission method

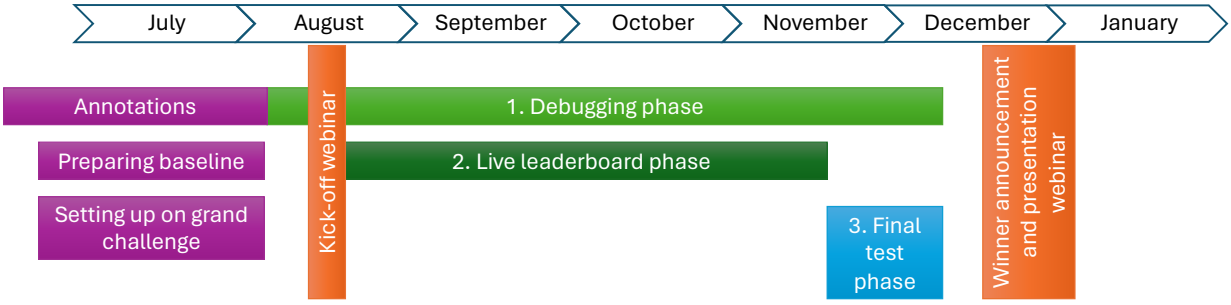
a) **Describe the method used for result submission**. Preferably, provide a link to the submission instructions.
Submissions will be made using Docker containers to grand-challenge.com. We are preparing a tutorial on how to do this, in addition to the documentation by Grand Challenge (found.

b) **Provide information on the possibility for participating teams to evaluate their algorithms** before submitting final results.
There will be a live leaderboard phase, during which participants can see how their algorithm performs on a hold-out validation set of 10 cases. They can also use cross-validation on the training dataset.

Item 9: Challenge schedule

Provide a **timetable** for the challenge.

Timeline



Item 10: Ethics approval

Indicate whether **ethics approval** is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a **reference to the document** of the ethics approval (if available).

Approval No. 2022-13686 from Prof. Dr. P.N.R. Dekhuijzen, Chair of the Institutional Review Board of the Radboud University Medical Center, CMO Radboudumc (METCoost-en-CMO@radboudumc.nl), approved on 31. March 2022.

Item 11: Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit **listing of the license** applied.

The data will be distributed under the CC BY-NC-SA (Attribution-NonCommercial-ShareAlike) license.

Item 12: Code availability

a) Provide information on the **accessibility of the organizers' evaluation software** (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.
   The evaluation code will be provided to the participants via GitHub (https://github.com/computationalpathologygroup/monkey-challenge).

b) In an analogous manner, provide information on the **accessibility of the participating teams' code.**
   The participating team's code and model weights must be available on GitHub (or a similar platform) and must be open access.

Item 13: Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to **sponsoring/ funding** of the challenge. Also, state explicitly who had/will have **access to the test case labels** and when.

This challenge is funded by the Dutch Kidney Foundation (Grant Nr. 21OK+012). The award money is a legacy from former IBEX employee John Theunissen. Any members of the Computational Pathology Team at RadboudUMC can access the test case labels. The test cases and annotations will not be released publicly.

MISSION OF THE CHALLENGE

Item 134: Field(s) of application

State the **main field(s) of application** that the participating algorithms target.

- Diagnosis
- Medical image analysis research

Item 15: Task category(ies) State the task category(ies).

- Classification
- Detection
- Localization

Item 16: Cohorts

We distinguish between the *target cohort* and the *challenge cohort*. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery.
While the challenge could be based on *ex vivo* data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding gender or age (target cohort).
The challenge and the target cohort have the same data collection protocol, a cohort of 121 kidney transplant biopsies. The only difference is that the challenge cohort also contains IHC slides, but the algorithm evaluation will only be conducted on the PAS-stained slides.

Item 17: Imaging modality(ies)

Specify the **imaging technique(s)** applied in the challenge.
For each case, there is a PAS-stained and IHC (double staining for CD3/CD20 and PU.1) (re-) stained whole slide image (WSI), performed in the lab at RadboudUMC.

Item 18: Context information

Provide additional **information given along with the images**. The information may correspond ...

a.) … directly to the **image data** (e.g. tumor volume).
b.) … to the **patient** in general (e.g. gender, medical history).

We will provide a quality score for the IHC slides, the institution from which the biopsy was taken, and the final biopsy diagnosis to give participants an overview of the distribution of different morphologies. The categories are insufficient clues for rejection (normal), ABMR (anti-body mediated rejection), TCMR (T-cell mediated rejection), mixed (ABMR+TCMR), borderline, chronic damage (IFTA), and other (BK virus nephropathy, necrosis).

Item 19: Target entity(ies)

a.) Describe the **data origin**, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and

challenge cohort.
All data originates from transplant kidney biopsies, which are used to assess a transplant organ's health and define the treatment strategy that ensures the longevity of the donor organ.

b.) Describe the **algorithm target**, i.e. the structure(s)/ subject(s)/ object(s)/ component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.
The challenge is focused on detecting and differentiating inflammatory cells in PAS-stained WSI. The inflammatory cells in question are monocytes and lymphocytes. The challenge also includes detection without distinction between the cells, called MNL (mono-nuclear leukocytes). The expected outputs are point predictions of the cell's location in pixels at a spacing of 0.24/µm.

Item 20: Assessment aim(s)

Identify the **property(ies) of the algorithms to be optimized** to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see "Metrics"), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- *Example 1:* Find liver segmentation algorithm for CT images that processes CT images of a certain size in less than a minute on a certain hardware with an error that reflects inter- rater variability of experts.
- *Example 2:* Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below.
Our primary goal is to optimize the performance of inflammatory cell detection, followed by classifying the detected inflammatory cells into two subclasses. We will measure this using the Free Response Operating Characteristic (FROC) analysis (see "Metrics").

CHALLENGE DATA SETS

Item 21: Data source(s)

a.) Specify the **device(s)** used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

b.) Describe relevant details on the imaging process/**data acquisition** for each acquisition device (e.g. image acquisition protocol(s)).
All WSI are scanned with two different scanner settings ("CPG profile" and "diagnostic profile") using a P1000 WSI scanner (3DHistech, Hungary) at RadboudUMC. Different scan profiles result in different color reproduction in the resulting images, for which the developed AI models should preferably be insensitive. For most cases, we also offer the original scan performed at the source institution (Vienna: 3D-Histec Pannoramic 250, Bern: 3D-Histec P1000, Emory: Olympus Nanozoomer, Mayo: Aperio system, UMCU: Hamamatsu XR).

c.) Specify the **center(s)/institute(s)** in which the data was acquired and/or the **data**

providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.
The cases and slides were collected six different pathology departments from 4 countries:
A.) RadboudUMC, Netherlands
B.) UMC Utrecht, Netherlands
C.) Medical University of Vienna, Austria
D.) Mayo Clinic Minnesota, USA
E.) IGMP, University of Bern, Switzerland
F.) Emory University, USA.

d.) Describe relevant **characteristics** (e.g. level of expertise) **of the subjects** (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).
The inclusion criteria for selecting biopsies were discussed among a group of expert renal pathologists. An experienced laboratory technician developed the restaining protocol, and renal pathologists evaluated staining quality and applicability. Student annotators needed to learn how to interpret the IHC staining. A resident pathologist explained this and created a brief guideline on how to annotate lymphocytes and monocytes. We used an existing AI model to automatically create dot annotations based on the chromogen's staining intensity, accelerating the annotation process. An advanced resident pathologist specializing in renal pathology checked all annotated cases.

Item 22: Training and test case characteristics

a.) State what is meant by one **case** in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).
Examples: Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any). A case refers to all information that is available for one particular patient in a specific information (parameter 18). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.
In the training set, one case consists of (i) 2-3 PAS-stained WSI scans: CPG profile, diagnostic profile, and original scan (if available), and (ii) double-stained IHC WSI scan.
A case in the validation (live leaderboard) and final test set refers to one PAS-stained WSI scan (CPG profile).
In all cases, one or more ROIs are annotated.

b.) State the **total number** of training, validation and test cases.
Training: 26+18+20+19 = 83 (centers A, B, C, D)
Validation: 10 (center E)
Test: 9 (center E) + 19 (center F).

c.) Explain **why a total number** of cases and **the specific proportion** of training, validation and test cases was chosen.
We split the cases by center to create a realistic scenario where a model is developed in a new institution. Four centers are used for the training set to ensure enough data was available for training. Center E is split between the validation and test set to have a reference in case certain algorithms perform well during validation and fail during test time. Center F uses a very different type of scanner, which could be challenging to generalize to.

d.) Mention **further important characteristics** of the training, validation and test cases

(e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.
Detection of inflammatory cells is strongly influenced by the presence of specific pathologies in the slides. Being relatively straightforward in biopsies with little pathologies, it may be very hard (also for expert humans) to identify these cells in biopsies with severe scarring. We aimed to collect a similar number of cases and a similar distribution of morphologies from all centers. The split between validation and test for center E also ensures a similar distribution between both subsets.
The study protocol sent to all collaborators specified the following (aiming to collect 20 cases each):

a. No-mild changes:
    i. 2 no rejection or inconclusive
    ii. 2 mild signs of rejection or mild IFTA (<25%)
b. Moderate-severe changes
    i. 2 moderate-severe glomerulitis
    ii. 2 moderate-severe endovasculitis
    iii. 2 moderate-severe tubulitis
    iv. 2 moderate-severe peritubular capillaritis
    v. 2 moderate IFTA (26-50%)
    vi. 2 severe IFTA (>50%)
c. Other alterations
    i. 4 – tubulopathic changes, polyoma/BK, pyelonephritis, ischemic necrosis, etc.

Item 23: Annotation characteristics

a.) Describe the **method for determining the reference annotation**, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include *manual image annotation*, *in silico ground truth generation* and *annotation by automatic methods*.
We use IHC double-staining for monocytes (applying monoclonal antibody PU.1, red) and lymphocytes (CD3/CD20, brown) to guide the annotation process. These slides are either re-stains of the PAS slide or consecutive cuts of the PAS slide. An experienced laboratory technician developed the re-staining protocol, and renal pathologists evaluated staining quality and applicability.

b.) If human annotation was involved, state the **number of annotators**.
6 annotators.

c.) Provide the **instructions given to the annotators** (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the **annotation protocol**.
The ROIs were all selected by DmV.
Student annotators were instructed on how to interpret the IHC staining to annotate the lymphocytes and monocytes by DvM. The annotation process is accelerated by generating automated detections of the lymphocytes and monocytes based on the IHC slide. We used HistoKat fusion from Fraunhofer Mevis to align the original PAS to the re-stained slides. A previously developed model for the automated detection of lymphocytes was used to create automated annotations for both lymphocytes and monocytes (using color deconvolution). The annotators (students + FM) then curate the automated annotations. False positive detections are deleted, and missed monocytes

and lymphocytes are added. After that, DmV reviews all annotations. She also annotates cases with difficult morphologies or where the registration or automated detection on the IHC failed. The same protocol is used for all cases.

d.) Provide **details on the subject(s)/algorithm(s) that annotated** the cases (e.g. information on **level of expertise** such as number of years of professional experience, medically trained or not). Provide the information separately for the training, validation and test cases if necessary.

Four student annotators (VD, MdK, HQ, TdW)
One 5th-year resident pathologist (not specializing in renal pathology, FM)
One 5th-year resident specializing in renal pathology (DvM)

e.) Describe the **method(s) used to merge multiple annotations** for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

N/A

Item 24: Data pre-processing method(s)

Describe the **method(s) used for pre-processing** the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

Whole slide images (WSIs) are scanned using various staining techniques, profiles, and formats, necessitating their conversion to a standard format. All slides are first registered to the corresponding Periodic Acid-Schiff (PAS) staining diagnostic profile using HistokatFusion as the registration tool to achieve this. This ensures that all slides have the same coordinates, allowing annotations to be made and visualized on a common coordinate system. The output of the registration process is a file in the ".sqreg" format, which includes the paths to both the template and reference WSIs. All registered slides are subsequently converted to the TIFF format to standardize the format. The same protocol is used for all cases.

Item 25: Sources of error

a.) Describe the most relevant **possible error sources related to the image annotation**. If possible, **estimate the magnitude** (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

In contrast to most other computational pathology challenges, MONKEY's reference standard is primarily based on highly objective immunohistochemical staining. Also, all annotations are visually checked and corrected if an experienced pathologist deems it necessary. Still, interpretation of IHC staining, varying IHC staining results, and manual operation of a computer mouse will introduce a small amount of annotation noise.

b.) In an analogous manner, describe and quantify **other relevant sources of error**.

Issues from scanning or WSI preparation (artifacts, bad staining, etc.) can negatively impact the image quality. However, since all ROIs are chosen manually, these issues are noticed and can be rectified, i.e., by rescanning or excluding the slide.

ASSESSMENT METHODS

Item 26: Metric(s)

a.) Define the **metric(s) to assess a property of an algorithm**. These metrics should reflect the desired algorithm properties described in *assessment aim(s)* (parameter 20). State which metric(s) were used to compute the ranking(s) (if any).
We will apply the Free Response Operating Characteristic (FROC) analysis. In it, the true positive rate (TPR), a.k.a. sensitivity or recall) is plotted against the average number of false positives (FP) per mm² over all slides. We define a detected cell as true positive (TP) if it lies within a distance margin of a manually annotated cell. The margin is 10µm and 4µm for monocytes and lymphocytes, respectively. For the combined detection, the margin is 7.5µm. Based on this definition, we will compute the TP, FP, and false negatives (FN) and use them in the FROC analysis. From the FROC curve, we derive an "FROC score" by taking sensitivity at five pre-selected values of FP/mm²: [10, 20, 50, 100, 200, 300]. The score computation may be fine-tuned during the challenge to compare the best methods better.

b.) **Justify why** the metric(s) was/were chosen, preferably with reference to the biomedical application.
FROC has been previously used in detection tasks in the TIGER and CAMELYON challenges.

Item 27: Ranking method(s)

a.) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically, the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.
Leaderboard 1 will show the results for the MNL detection (overall inflammation). We will directly use the FROC value. Leaderboard 2 will show the lymphocyte and monocyte detection results, where the FROC will be computed per class and then averaged for the final ranking. The individual values will, however, be visible on the leaderboard.

b.) Describe the method(s) used to manage **submissions with missing results** on test cases.
Missing results will result in a lower performance, as there will be more false negatives.

c.) **Justify why** the described ranking scheme(s) was/were used.
To address the class imbalance between the monocytes and lymphocytes,

Item 28: Statistical analyses

a.) Provide **details for the statistical methods** used in the scope of the challenge analysis. This may include description of the **missing data handling**, details about the assessment of **variability of rankings**, description of any method used to assess **whether the data met the assumptions**, required for the particular statistical approach, or indication of any **software product** that was used for all data analysis methods.
N/A, as no patient data is used.

b.) **Justify why** the described statistical method(s) was/were used.
N/A

Item 29: Further analyses

After the final test stage of the challenge, we are planning to use the best algorithm(s) for future analysis:

- **Susceptibility to image variance:** The original slides from the hold-out test set from center F have also been scanned with a very different scanner, which allows us to assess the impact of scanner variability.

- **Reader study:** For the final milestone of the DIAGGRAFT project, we will organize a reader study with nine pathologists. The goal is for each of them to diagnose 100 cases. They will be requested to give scores for the six Banff lesion scores g, t, ptc, ci, ct, and i, as well as to categorize the biopsy into “T cell-mediated rejection”, “antibody-mediated rejection”, “mixed rejection”, “no specific allograft pathology”, and “other diseases of the allograft”. To compare the influence of computer-aided diagnostics, each pathologist will assess each case twice, once with and once without AI assistance. Thus, we can analyze the inter-observer agreement between pathologists, as well as between the pathologists and the AI.
  Allograft failure and additional donor, recipient, and transplantation factors are also known for this cohort. This allows us to further study the prognostic value of pathologists’ diagnoses with and without AI assistance.