Supplementary Material to:
Next-Generation Morphometry for pathomics-data mining in histopathology

Running title: Automated segmentation and quantification of human kidney histopathology

David L. Hölscher[1,#], Nassim Bouteldja[1,#], Mehdi Joodaki[2], Maria L. Russo[3], Yu-Chia Lan[1], Alireza Vafaei Sadr[1], Mingbo Cheng[2], Vladimir Tesar[4], Saskia v. Stillfried[1], Barbara M. Klinkhammer[1], Jonathan Barratt[5,6], Jürgen Floege[7], Ian SD Roberts[8], Rosanna Coppo[3,9], Ivan G. Costa[2], Roman D. Bülow[1,§], Peter Boor[1,7,§,*]; for the International IgA Nephropathy Network[φ]

1 Institute of Pathology, RWTH Aachen University Clinic, Aachen, Germany
2 Institute for Computational Genomics, RWTH Aachen University Clinic, Aachen, Germany
3 Fondazione Ricerca Molinette, Torino, Italy
4 Department of Nephrology, 1st Faculty of Medicine and General University Hospital, Charles University, Prague, Czech Republic
5 John Walls Renal Unit, University Hospital of Leicester National Health Service Trust, Leicester, United Kingdom
6 Department of Cardiovascular Sciences, University of Leicester, Leicester, United Kingdom
7 Department of Nephrology and Immunology, RWTH Aachen University Clinic, Aachen, Germany
8 Department of Cellular Pathology, Oxford University Hospitals National Health Services Foundation Trust, Oxford, United Kingdom
9 Regina Margherita Children's University Hospital, Torino, Italy
φ Contributor information are provided at the end of the manuscript

Shared first authors
§ Shared senior authors

* Address correspondence to:
Peter Boor, M.D., Ph.D.
Institute of Pathology
RWTH Aachen University Hospital
Pauwelsstrasse 30
52074 Aachen, Germany
Phone: +49 241 80 85227
Fax: +49 241 80 82446
E-mail: pboor@ukaachen.de

Table of Contents

Supplementary Methods

Cohort assembly of the internal cohorts

For all cases from the *AC_B* & *AC_N* cohort the diagnosis, histopathological scores and clinical data were collected when available. Final diagnoses and histopathological scores were gathered from the pathology reports. Histopathological scoring and diagnosis were based on the consensus of at least two trained nephropathologists. The following information on histological structures were collected from the reports when available: number of glomeruli, number of globally sclerotic glomeruli, information regarding the presence and severity of arteriosclerosis, and the extent of interstitial fibrosis and tubular atrophy (IFTA). Furthermore, clinical data including laboratory findings and ICD-10 coded diseases were collected if available within the pathology laboratory information system (Supp. Table 12). Estimated glomerular filtration rate (eGFR) derived by using the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) Equation[1] [ml/min/1.73m^2] and proteinuria [mg/d], were gathered if available at the time of biopsy. Presence of hypertension regardless of aetiology was assessed based on ICD-10 codes. Non-availability of clinical or pathological data did not automatically lead to exclusion (please see exclusion criteria).

Before digitalisation of slides all patients were anonymised and given a unique patient identifier. Overall, 320 biopsy cases from 297 patients

(number of biopsies per patient - mean: 1.08, min: 1, max: 4) were included in the *AC_B* cohort. There was a mean of 3.03 Periodic Acid Schiff (PAS)-stained whole-slide images (WSIs) per biopsy case (min: 1, max: 7).

Annotation process

Ground truth annotations of histological kidney structures were performed by an experienced nephropathologist and two trained medical students using QuPath[2], which is the most widely used open-source software for computational pathology applications. In a second step the nephropathologist corrected all annotations. Annotations of difficult/questionable structures were discussed in regular meetings with another nephropathologist. All annotators were instructed to follow a standard operating procedure with clear cut definitions for all six classes (Suppl. Table 3). Besides, the following procedure was performed to accelerate the time-consuming manual annotation process: First, a quarter of the training dataset was annotated and used to train a prior segmentation model. Then additional patches were selected for annotation, focusing on structures that were insufficiently recognized by the prior model and loaded their model predictions into QuPath to provide accurate starting points for annotation. This procedure strongly reduced annotation efforts as it converted the manual annotation into an annotation correction process. However, it was only implemented for training data, meaning that all data used to assess performance and

validate the models was generated completely by experts. This cycle was repeated twice after two and three quarters of data annotations. To ensure generalisability to external cohorts and applicability in a broad spectrum of diseases, specimens from our internal cohorts that reflect the diversity of kidney histopathology for training and testing were selected (e.g., 21 cases of IgA nephropathy (IgAN), 11 pauci-immune glomerulonephritides (GN), a full list of diagnoses included in training/testing/external validation is provided in Suppl. Table 6).

CNN development - tissue segmentation

An nnU-Net was used for the kidney tissue segmentation task. The nnU-Net has gained popularity due to its superiority in several international biomedical segmentation competitions without the need for application-specific manual intervention[3]. Regions of size 2500x2500µm^2 were extracted in a grid-like manner from our annotated cohorts and downsampled into patches of size 512x512 pixels. Regions containing at least 90% white background were ignored. A batch size of 12 was used for training. Soft Dice loss as well as binary cross-entropy were used as equally weighted loss functions. The initial learning rate of 0.01 was polynomially reduced using a power of 0.9 during a total amount of 1000 epochs. Colour and spatial transformations including Gaussian noise, Gaussian blur, gamma contrast, rotations, and scaling were used for data augmentation. In order to further increase the robustness of the model, a five-fold cross-

validation on the training data was performed to train an ensemble of five different convolutional neural networks (CNNs), and then selected the model providing the lowest internal validation loss as the final one in each respective fold. Given an input image, the final prediction was then obtained by averaging the predictions of all five models, followed by filling holes and removing very small predictions. Overall, 12,034 patches from 1,056 WSIs were extracted and used for the development, testing and validation of the model. Training was conducted only on the internal *AC_B* and *AC_N* cohorts using a random 80% data split on patient level, respectively. The remaining 20% of data from both internal cohorts were used for testing (Supp. Table 15).

CNN development - structure segmentation

A U-Net-like model[4] was trained on random minibatches of size six using RAdam[5] as an optimizer for the kidney structure segmentation task. The initial learning rate of 0.001 was divided by three in case the internal validation loss did not drop for 15 epochs straight. Training terminated once the learning rate fell below 4E-6. Then, the model providing the lowest internal validation loss was chosen as the final configuration. Soft Dice-loss as well as the weighted categorical cross-entropy (WCE) were used as equally weighted loss functions as well. In addition, affine, piecewise affine, elastic, 90-degree rotation, flipping, hue and saturation shifting, and gamma contrast transformations were used for data

augmentation. CNN predictions were post-processed by first applying test-time augmentation to improve the model's robustness, second, removing border class predictions, third, filling prediction holes, fourth, removing too small instance predictions, and fifth, dilating tubular predictions. A border class comprising border pixels of all structures to enable the separation of touching instances from the same class, and thus instance-level analysis was implemented. Border pixels were computed for tubules by dilation using a ball-shape structuring element of radius three pixels. Full borders of arteries and glomeruli were not included into the border class as this would have prevented continuous class transitions, e.g., between tubules and glomeruli at their urinary pole. Instead, arteries and glomeruli were dilated using a radius of seven pixels and only their class-specific overlap was included into the border class. Consequently, the border class mostly represented the tubular basement membranes. Using the WCE, a ten times greater weight was assigned to the border class for improved instance separation. Annotated data from the internal cohorts (*AC_B*: 68 WSIs, 3,162 patches; *AC_N*: 17 WSIs, 431 patches) was split into training (*AC_B*: 54 WSIs, 2,621 patches; *AC_N*: 10 WSIs, 200 patches), internal validation (*AC_B*: 2 WSIs, 78 patches; *AC_N*: 2 WSIs, 30 patches) and test set (*AC_B*: 12 WSIs, 463 patches; *AC_N*: 5 WSIs, 201 patches) on case level. External validation was performed on the held-out data from the *KPMP* and *HuBMAP* cohorts on 240 and 198 patches from five WSIs, respectively (Supp. Table 15).

Quantitative performance metrics

Regular Dice-similarity-coefficients (DSCs) were used to measure performance of the tissue segmentation CNN on WSI-level. Performance of the structure segmentation CNN was assessed by instance Dice-similarity-coefficients (iDSCs) to measure accuracy specifically on instance-level. For each prediction instance, a Dice score measuring the spatial overlap with its maximally overlapping ground-truth instance was computed, which was repeated *vice versa* for each ground truth instance. Thus, the iDSC quantifies the mean predicted area coverage per instance, and ranges like the DSC from 0 (no single ground-truth overlap) to 1 (perfect predictions). F1-Scores and positive predictive values (PPVs) were computed by counting the true positives (correctly segmented instances defined as an iDSC above the threshold of 0.5), false positives and false negatives.

Computation of features

A multitude of features for each individual instance was computed after instance segmentation (Supp. Table 3). We computed the area [μm^2] and diameter [μm] of the cross-section of tubules, glomeruli, glomerular tufts, arteries and their corresponding lumina. To encompass simultaneous changes in glomerular and glomerular tuft area the Tuft-Area-Fraction was calculated by dividing glomerular tuft area through glomerular area

per instance as well as the Bowman's area by subtracting the glomerular tuft area from the glomerular area. The area and diameter of whole arteries, the lumen and the wall consisting of intima and media were calculated (adventitia was not segmented and excluded). Small arterioles (full diameter < 50µm) were excluded due to lower segmentation accuracy to enable accurate morphometry analysis. Distances [µm] between glomeruli and their respective nearest glomerular neighbour, as well as between tubules and their respective nearest structure were also calculated. The circularity, eccentricity, elongation and solidity which were calculated based on common shape extraction techniques[6], each taking values between 0 and 1 were calculated for full glomeruli and glomerular tufts. Instance counts and area percentages were extracted on WSI-level.

Features were calculated in all five cohorts on instance-level, but were summarised on patient-level as well. All instances inherited the diagnostic or clinical label of their parent case, facilitating the comparison of diseased cases to normal cases as well as the analysis of discrepancies between different kidney diseases.

Cox regression models

Time from kidney biopsy to the combined endpoint up to eight years (end-stage kidney disease (ESKD) and/or halfing of initial eGFR assessed at time of biopsy) was analysed in the *VALIGA* cohort by implementing

Cox regression models. 4 cases had to be excluded due to missing data (initial eGFR). eGFR and age were grouped by steps of 10. The proportional hazard assumption was assessed graphically using Schoenfeld residuals. Five features (tubular distance, tubular diameter, tuft circularity, tuft area and bowman area) which were descriptive of disease morphology or kidney function based on eGFR in our internal biopsy cohort were selected as digital biomarkers. Optimal cutoffs for all five features were determined by computing the maximally selected log-rank statistic[7]. No model based on the continuous distribution of features was computed because of the large variation in effect sizes due to different scalings (e.g., 0-1 or 0-60,000 um^2). Normalisation of feature distributions was not performed due to the lacking explainability of the actual changes in histopathology. The five features were then used to fit initial unadjusted models. Next, each model was adjusted for age, sex, initial eGFR and MEST-C score (each component of the score respectively). Additionally, a *Digital Biomarkers* model was constructed, using all five features as predictors which was adjusted for age, sex and eGFR at the time of biopsy. In addition, a *MEST-C* model with the M, E, S, T and C components of the score as covariates was constructed and adjusted for age, sex and initial eGFR. A third, hybrid model combined both previously described regression models. Models were compared to each other based on C-Statistic, Akaike information criterion (AIC) and Bayesian information criterion (BIC).

Data analysis with R

Quantitative feature analyses for each WSI were imported as dataframes into the R Environment and data transformation was performed using the following publicly available packages: tidyverse v1.3.1 (cran.r-project.org/web/packages/tidyverse), dplyr v1.0.8 (cran.r-project.org/web/packages/dplyr), gdata v2.18.0 (cran.r-project.org/web/packages/gdata), and broom v0.7.12 (cran.rstudio.com/web/packages/broom). Statistics were calculated using the kSamples v1.2-9 (cran.rstudio.com/web/packages/kSamples) and PMCMRplus v1.9.3 (cran.rstudio.com/web/packages/PMCMRplus) packages. Survival analysis was performed with the survival v3.3-0 (cran.rstudio.com/web/packages/survival), survminer v0.4.9 (cran.r-project.org/web/packages/survminer) and maxstat v0.7-25 (cran.rstudio.com/web/packages/maxstat) packages.

Plots were created by using the packages: ggplot2 v3.3.3 (cran.r-project.org/web/packages/ggplot2) and cowplot v1.1.1 (cran.r-project.org/web/packages/cowplot).

Supplementary Table 1. Clinical and general patient data for included cases across the five datasets. *AC_B* and *AC_N* are internal cohorts while *KPMP*, *HuBMAP* and *VALIGA* are external cohorts. In the internal cohorts presence of hypertension regardless of its aetiology was evaluated by ICD-10 coded diseases.
F: Female; M: Male; eGFR, estimated glomerular filtration rate; G1, eGFR ≥90ml/min/1.73m²; G2, eGFR 60-89ml/min/1.73m²; G3, eGFR 30-59ml/min/1.73m²; G4, eGFR 15-29ml/min/1.73m²; G5, eGFR <15ml/min/1.73m².

Cohort	Cases [n]	Age	Sex	eGFR [mg/dl/1.73m²]	Proteinuria [mg/d]	Hypertension [y/n]
AC_B	320	51.78 (±16.06)	F: 129 (40.3%) M: 191 (59.7%)	39.75 (±31.49)	2,719.20 (±4,458.70)	y = 71.4% (n=202)
AC_N	38	60.55 (±17.21)	F: 19 (50.0%) M: 19 (50.0%)	-	-	y = 84.2% (n=32)
KPMP	36	20-29: 1 (2.8%) 30-39: 4 (11.1%) 40-49: 3 (8.3%) 50-59: 5 (13.9%) 60-69: 18 (50.0%) 70-79: 5 (13.9%)	F: 21 (58.3%) M: 15 (41.7%)	G1: 6 (17.7%) G2: 8 (23.5%) G3: 17 (50.0%) G4: 3 (8.8%)	<150: 4 (21.1%) 150-499: 5 (26.3%) 500-999: 3 (15.8%) ≥1,000: 7 (36.8%)	y = 79.4% (n=27)
HuBMAP	9	58.56 (±13.10)	F: 4 (44.4%) M: 5 (55.6%)	-	-	y = 55.6% (n=5)
VALIGA	648	36.07 (±15.17)	F: 182 (28.1%) M: 466 (71.9%)	72.39 (±29.95)	1,970.97 (±2,039.13)	-

Supplementary Table 2. Consensus criteria to annotate the six defined classes included in the structure segmentation.
FSGS: Focal segmental glomerulosclerosis.

Class	Annotation Criteria
Tubule	● Annotation along, but excluding the basement membrane ● Annotation along the inner part of split basement membrane segments
Glomerulus	● Annotation along the Bowman‘s capsule including the basement membrane ● Annotation along the inner part in cases with split Bowman’s capsule ● In cross-sections where urinary or vascular pole were visible, Annotations were continued in a straight line at the capsule border
Glomerular Tuft	● Subclass of the glomerulus class with annotation of the glomerular tuft only (including podocytes) ● In glomeruli with global glomerulosclerosis and no visible capillaries no tuft was annotated ● Extracapillary proliferates (=crescents) or FSGS tip lesions were not annotated as tuft
Non-Tissue Background	● Annotation of large „white“ areas including veins, section background and renal pelvis ● For veins only the lumen was annotated, endothelium was excluded ● Annotations were only performed for structures with a diameter of at least 30µm
Artery	● Annotation of intima and media of arteries excluding adventitia ● Minimum of one visible vascular smooth muscle cell layer was required for annotation
Arterial Lumen	● Subclass of the artery class ● Annotation of the lumen only (including intravascular cells), endothelium was excluded

Supplementary Table 3. Description of morphometric features extracted from predicted structures by the structure segmentation convolutional neural network.
WSI: Whole-slide image.

Feature	Description
Area A	Area of the segmented structure [µm²]
Diameter $dmax$	Diameter of the largest circle fully fitting inside the structure [µm]
Distance $distmin$	Closest distance between structures of the same class [µm]
Circularity C	Measures how circular the structure is using the ratio of its area (A) multiplied by 4*π to its squared perimeter (P). The circularity of a circle equals 1. $$C = \frac{4 * \pi * A}{P^2}$$
Elongation Elo	Function of the length of the structure's minor and major axis. The elongation of a circle is 0 and gets higher the more elongated the structure is. $$Elo = 1 - \frac{minor\_axis\_length}{major\_axis\_length}$$
Eccentricity Ecc	Ratio of the distance between the structure's focal points over the major axis' length. The eccentricity equals 0 for a circle and 1 for ellipses. $$Ecc = \frac{\sqrt{(major\_axis\_length^2 - minor\_axis\_length^2)}}{major\_axis\_length}$$
Solidity S	Measures the density of the structure by taking the ratio of its area to the area of its convex hull (H) $$S = \frac{A}{H}$$
Area Percentage	Proportion of a class' total area from the overall tissue area [%]
Count	Number of instances of a particular class present in one WSI or specimen

Supplementary Table 4. Quantitative analysis of glomerular and tubular morphometry in common native kidney diseases from our internal biopsy cohort (*AC_B*). Continuous feature distributions are reported as median with interquartile range in brackets.
IgAN: IgA nephropathy; MCD: Minimal change disease; Membranous: Membranous glomerulonephritis; Pauci: Pauci-immune glomerulonephritis, DN: Diabetic nephropathy; HTN: Hypertensive nephropathy.

Class	Glomerular Area [µm²]	Tuft Area [µm²]	Tuft Circularity	Tubular Diameter [µm]
Normal	16,897.73 (17,307.55)	13,425.64 (12,463.40)	0.45 (0.15)	30.79 (18.34)
IgAN	15,006.39 (16,601.41)	12,047.98 (12,848.73)	0.39 (0.18)	29.38 (20.41)
MCD	18,235.09 (19,933.58)	15,962.92 (15,162.28)	0.41 (0.16)	31.95 (19.78)
Lupus	18,149.40 (23,645.80)	16,071.97 (18,695.50)	0.38 (0.18)	29.28 (19.04)
Membranous	21,157.73 (25,670.11)	18,867.77 (18,320.57)	0.41 (0.18)	31.18 (20.43)
Pauci	16,900.64 (18,243.61)	11,778.88 (12,676.00)	0.36 (0.20)	29.67 (18.57)
DN	13,416.03 (17,739.20)	11,121.69 (14,969.37)	0.34 (0.19)	25.85 (17.62)
HTN	12,474.17 (21,559.04)	13,651.64 (20,009.59)	0.37 (0.19)	26.23 (20.69)

Supplementary Table 5. Quantitative analysis of glomerular morphometry based on proteinuria in our internal biopsy cohort (*AC_B*) and external *KPMP* cohort. Continuous feature distributions are reported as median with interquartile range in brackets.
MCD: Minimal change disease; Membranous: Membranous glomerulonephritis.

Class	Tuft Area [µm²]	Tuft Circularity
AC_B		
< 3.5g/d	13,026.12 (13.429.45)	0.40 (0.18)
> 3.5g/d	14,290.70 (18,988.11)	0.35 (0.18)
MCD		
< 3.5g/d	18,448.18 (13,204.64)	0.41 (0.13)
> 3.5g/d	17,942.00 (20,155.83)	0.40 (0.14)
Membranous		
< 3.5g/d	16,266.41 (11,756.31)	0.46 (0.14)
> 3.5g/d	24,563.42 (21,967.46)	0.37 (0.18)
KPMP		
< 1g/d	12,537.56 (15,241.31)	0.42 (0.18)
> 1g/d	14,882.10 (22,825.07)	0.35 (0.21)

Supplementary Table 6. Quantitative analysis of glomerular and tubular morphometry based on kidney function measured by estimated glomerular filtration rate of native biopsy cases from the *AC_B* cohort. Continuous feature distributions are reported as median with interquartile range in brackets.
eGFR: estimated glomerular filtration rate.

eGFR [ml/min/1.73m²]	Tuft Circularity	Tubular Diameter [µm]	Tubular Distance [µm]
> 60	0.42 (0.16)	30.47 (19.47)	3.24 (0.99)
30-60	0.39 (0.17)	29.67 (20.24)	4.53 (1.45)
< 30	0.36 (0.20)	28.69 (19.96)	6.04 (2.55)

Supplementary Table 7. Quantitative analysis of tubular morphometry based on scored histopathology of all cases from our internal biopsy cohort (*AC_B*). Structures from biopsy cases were grouped based on stratified reported scoring into four groups. Continuous feature distributions are reported as median with interquartile range in brackets.
IFTA: Interstitial fibrosis and tubular atrophy.

IFTA	Tubular Diameter [µm]	Tubular Distance [µm]
0-10% (none/marginal)	30.47 (19.63)	4.02 (2.10)
11-25% (mild)	27.31 (20.61)	5.91 (2.34)
26-50% (moderate)	26.98 (19.25)	5.91 (2.21)
>50% (severe)	23.83 (17.27)	8.24 (3.19)

Supplementary Table 8. Quantitative analysis of vascular morphometry based on scored arteriosclerosis in our internal biopsy cohort (*AC_B*) and clinically reported hypertension status in four cohorts (*AC_B*, *AC_N*, *KPMP*, *HuBMAP*). Structures from cases were grouped based on stratified reported scoring and/or hypertension status into groups. Continuous feature distributions are reported as median with interquartile range in brackets.

Group	Wall Diameter [µm]	Lumen Diameter [µm]
Arteriosclerosis		
none	22.0 (11.78)	28.18 (20.80)
moderate	24.02 (12.6)	25.34 (20.77)
severe	26.98 (15.33)	22.62 (19.51)
Hypertension		
AC_B		
Hypertension	23.83 (12.22)	24.54 (22.46)
Normotension	21.94 (11.21)	27.15 (21.29)
AC_N		
Hypertension	23.5 (14.74)	25.03 (23.28)
Normotension	21.22 (13.77)	27.1 (17.47)
KPMP		
Hypertension	22.61 (13.19)	23.8 (21.08)
Normotension	22.08 (9.59)	27.15 (16.36)
HuBMAP		
Hypertension	24.39 (14.92)	27.73 (29.91)
Normotension	20.27 (14.94)	37.78 (29.28)

Supplementary Table 9. Summary of uni- and multivariate Cox regression models for digital morphometric features. Multivariate models were adjusted for age (grouped by steps of 10 years), sex (male vs. female), estimated glomerular filtration rate (grouped by steps of 10ml/min/1.73m²) and M, E, S, T as well as C of the Oxford classification for IgA nephropathy. Features were reported as dichotomous based on maximally selected log-rank statistics. HR: Hazard ratio; CI: Confidence interval, eGFR: estimated glomerular filtration rate; M: Mesangial hypercellularity; E: Endocapillary hypercellularity; S: Segmental sclerosis; T: Tubular atrophy; C: Crescents.

	HR	95% CI	*p*-value	HR	95% CI	*p*-value	HR	95% CI	*p*-value	HR	95% CI	p-value	HR	95% CI	*p*-value
Feature:	**Tubular Distance**			**Tuft Circularity**			**Tuft Area**			**Bowman Area**			**Tubular Diameter**		
Univariate:	4.8	3.1-7.43	<0.0001	3.7	2.46-5.57	<0.0001	3.37	2.03-5.58	<0.0001	2.12	1.31-3.43	0.0021	3.01	1.96-4.64	<0.0001
Multivariate:															
- Feature	2.04	1.18-3.52	0.0107	1.99	1.29-3.07	0.0019	1.77	1.03-3.03	0.0371	1.29	0.79-2.11	0.3097	1.82	1.13-2.92	0.0137
- Age	0.96	0.81-1.13	0.6162	1.0	0.84-1.19	0.9958	0.99	0.84-1.18	0.9379	0.97	0.82-1.15	0.7375	1.01	0.85-1.2	0.8971
- Sex	1.36	0.83-2.21	0.2244	1.14	0.70-1.86	0.5877	1.21	0.74-1.98	0.4365	1.22	0.75-1.98	0.4249	1.28	0.79-2.07	0.3247
- eGFR	0.81	0.72-0.91	0.0006	0.82	0.72-0.92	0.0011	0.8	0.71-0.90	0.0002	0.79	0.7-0.89	<0.0001	0.8	0.71-0.9	0.0003
- M	1.8	1.17-2.77	0.007	1.79	1.16-2.75	0.0083	1.81	1.18-2.78	0.0068	1.9	1.24-2.91	0.0034	1.96	1.28-3.01	0.0021
- E	1.26	0.67-2.39	0.4731	1.32	0.69-2.54	0.3976	1.26	0.67-2.38	0.4718	1.26	0.66-2.38	0.4847	1.23	0.65-2.33	0.534
- S	2.68	1.14-6.29	0.0236	2.56	1.09-6.01	0.0303	2.43	1.03-5.73	0.0417	2.6	1.11-6.09	0.0285	2.62	1.12-6.18	0.0271
- T	1.45	0.97-2.16	0.0696	1.7	1.17-2.48	0.0052	1.62	1.11-2.36	0.0123	1.75	1.21-2.53	0.0028	1.59	1.08-2.33	0.0189
- C	1.2	0.66-2.19	0.5572	1.3	0.71-2.39	0.4009	1.21	0.66-2.22	0.5405	1.2	0.65-2.2	0.5599	1.28	0.7-2.35	0.4292

Supplementary Table 10. Number of cumulative events (i.e., reaching the combined endpoint) of the five fitted univariate Cox regression models for digital morphometric features in the *VALIGA* cohort.

Group	Year 0	Year 2	Year 4	Year 6	Year 8
Tubular Distance [µm]					
< 4.13µm	2	9	13	23	30
> 4.13µm	4	23	40	57	62
Tuft Circularity					
> 0.33	3	12	21	37	45
< 0.33	3	20	32	43	47
Tuft Area [µm²]					
> 9091µm²	1	5	8	17	19
< 9091µm²	5	27	45	63	73
Bowman Area [µm²]					
< 5674µm²	1	6	10	17	22
> 5674µm²	5	26	43	63	70
Tubular Diameter [µm]					
> 24.87µm	3	19	33	55	61
< 24.87µm	3	13	20	25	31

Supplementary Table 11. Summary of the two fitted multivariate Cox regression models for digital morphometric features (*Digital Biomarkers*) and MEST-C histopathology scoring (*MEST-C*) in the *VALIGA* cohort. Age and estimated glomerular filtration rate were grouped by steps of 10 years and ml/min/1.73m² respectively.
HR: Hazard ratio; CI: Confidence interval.

	HR	95% CI	*p*-value
Multivariate:	**Digital Biomarkers Model (Feature-based)**		
- Tubular Distance	1.71	0.97-3.0	0.0628
- Tuft Circularity	1.78	1.16-2.76	0.009
- Tuft Area	2.11	1.23-3.63	0.0067
- Bowman Area	1.7	1.01-2.87	0.0466
- Tubular Diameter	1.59	0.96-2.64	0.0747
- Age	0.98	0.83-1.16	0.7895
- Sex	1.37	0.84-2.25	0.2073
- eGFR	0.8	0.72-0.9	0.0001
Multivariate:	**MEST-C (Histopathology-based)**		
- M	1.87	1.22-2.87	0.0042
- E	1.24	0.66-2.35	0.5068
- S	2.65	1.13-6.22	0.025
- T	1.77	1.22-2.55	0.0026
- C	1.23	0.67-2.25	0.5071
- Age	0.98	0.83-1.16	0.7986
- Sex	1.22	0.75-1.99	0.4137
- eGFR	0.79	0.7-0.88	<0.0001

Supplementary Table 12. Data availability for internal and external cohorts (n plus relative frequency in brackets). Histopathology includes the scored amount of interstitial fibrosis and tubular atrophy (IFTA) and the MEST-C-score for each biopsy (in case of *VALIGA*).
eGFR: estimated glomerular filtration rate.

Cohort	Cases [n]	Age	Sex	Histo-pathology	eGFR	Proteinuria	Hypertension
AC_B	320	320 (100%)	320 (100%)	320 (100%)	236 (74%)	122 (38%)	283 (88%)
AC_N	38	38 (100%)	38 (100%)	0 (0%)	0 (0%)	0 (0%)	38 (100%)
KPMP	36	36 (100%)	36 (100%)	0 (0%)	34 (94%)	19 (53%)	34 (94%)
HuBMAP	9	9 (100%)	9 (100%)	0 (0%)	0 (0%)	0 (0%)	9 (100%)
VALIGA	648	648 (100%)	648 (100%)	648 (100%)	644 (99%)	621 (96%)	0 (0%)

Supplementary Table 13. Diagnoses present in the *AC_B* cohort based on the Banff-Classification for transplant pathology[8] and the Mayo-Classification for glomerulonephritides[9]. For cases with multiple diagnoses the amount of overlapping cases was further reported.
ABMR: Antibody mediated rejection; TCMR: T-cell mediated rejection; GN: Glomerulonephritis; GBM: Glomerular basement membrane; FSGS: Focal segmental glomerulosclerosis.

Class	Category	Diagnosis	Cases [n]	Overlap
Transplant	Banff Category 1	normal	32	-
	Banff Category 2	ABMR	7	-
	Banff Category 3	Borderline TCMR	13	-
	Banff Category 4	TCMR	9	-
	Banff Category 5	Other (e.g., Polyomavirus nephritis)	41	-
	Mixed	combined category 2 & 3/4 rejection	6	-
Native	Immune-complex GN	IgA nephropathy (IgAN)	47	3 (DN/HTN)
		Lupus nephritis (LN)	13	1 (Memb. GN)
		Membranous GN (Memb. GN)	9	1 (LN)
		Membranoproliferative GN (MPGN)	4	1 (Interst. Neph.)
	Pauci-immune GN	MPO/PR3 ANCA GN (Pauci)	31	1 (Interst. Neph.)
	C3 glomerulopathy	C3 glomerulopathy	3	-
	Anti-GBM GN	Anti-GBM GN	1	-
	Podocytopathy	Minimal change disease (MCD)	16	-
		primary FSGS	1	-
	Tubulointerstitial	Interstitial nephritis (Interst. Neph)	11	2 (Pauci/MPGN)
		Cast nephropathy	4	-
		Acute tubular injury (isolated)	5	-
		Light-chain proximal tubulopathy	1	-
	Vascular	Hypertensive nephropathy (HTN)	9	3 (IgAN/DN)
		Thrombotic microangiopathy	4	-
	Other	Diabetic nephropathy (DN)	8	3 (IgAN/HTN)
		Amyloidosis	7	-
		IgM nephropathy	1	-
	Normal	age-appropriate kidney biopsy	17	-
	Non-specific	descriptive biopsy report	27	-

Supplementary Table 14. Number of instances that were annotated as ground truth in the development and external validation of the structure segmentation convolutional neural network. Instances which were present in multiple patches were only counted once.
WSI: Whole-slide image.

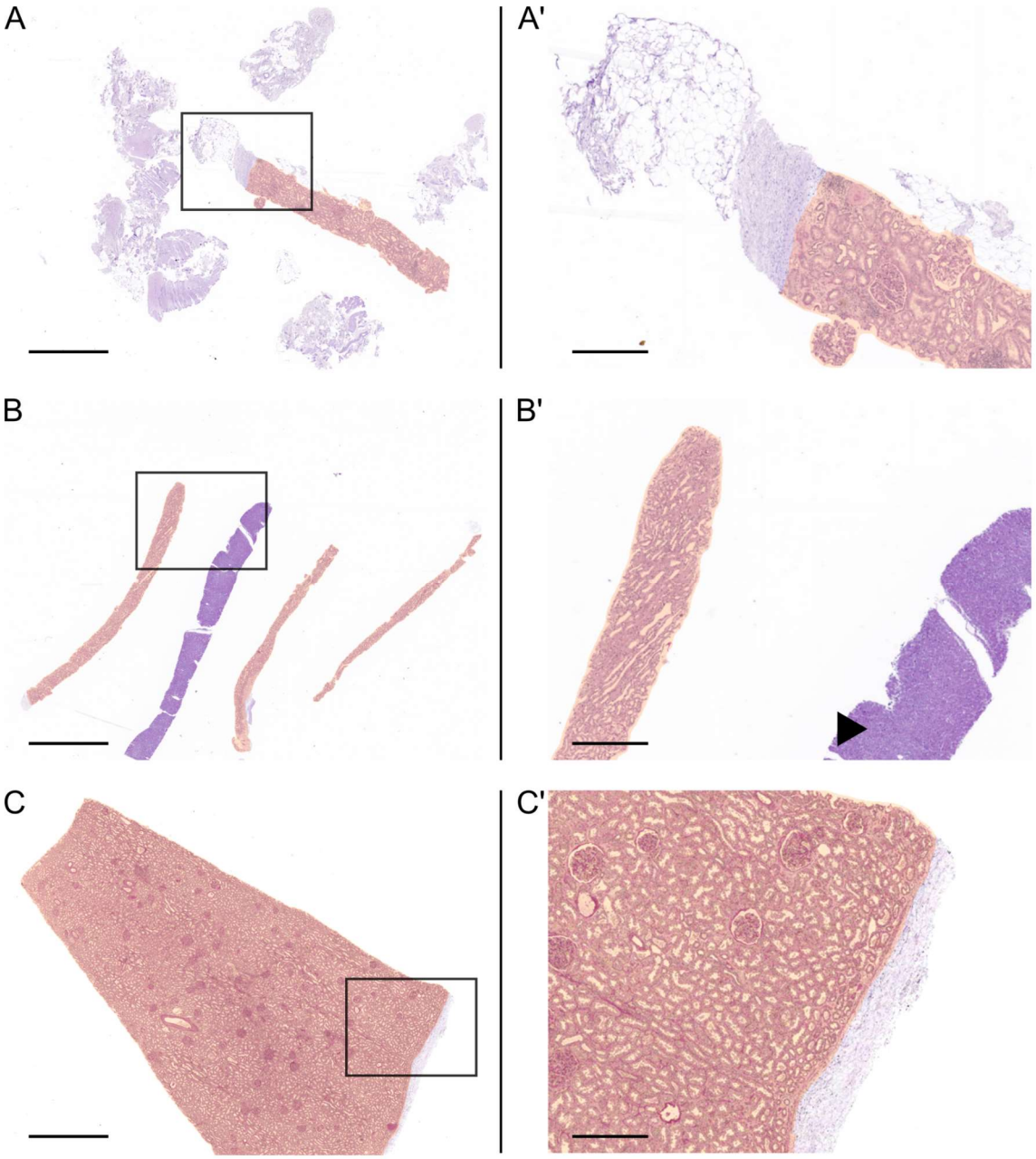
No. of Instances						
Cohort	Tubule	Glomerulus	Glomerular Tuft	Non-Tissue Background	Artery	Arterial Lumen
AC_B	14,492	761	627	605	1,887	2,041
AC_N	3,246	186	194	55	189	202
KPMP	1,070	50	46	38	177	189
HuBMAP	924	62	58	23	73	92
∑	19,732	1,059	925	721	2,326	2,524

Supplementary Table 15. Overview of the annotated ground truth data for the development and external validation of the tissue segmentation and structure segmentation convolutional neural network. We evaluated performance in our internal (*AC_B* & *AC_N*) cohort on a seperate test set. Unseen whole-slide images from the two external cohorts (*KPMP* & *HuBMAP*) were used for external validation.
WSI: Whole-slide image.

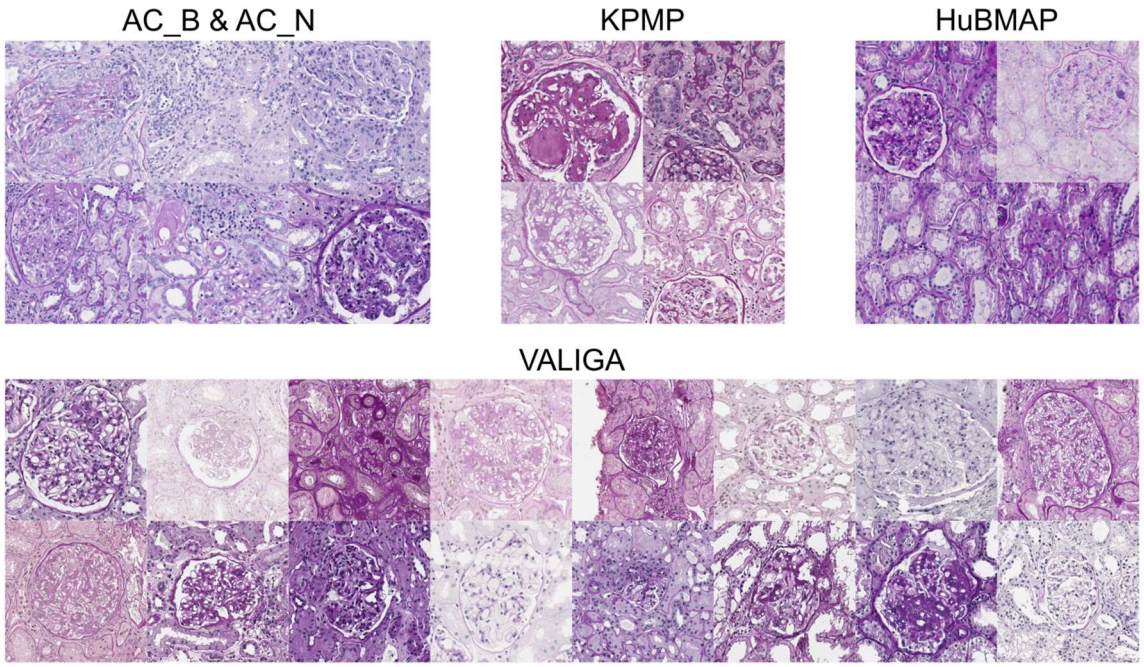
	Tissue Segmentation			Structure Segmentation					
	Number of WSIs			Number of WSIs			Number of Patches		
Cohort	Training	Testing/ Validation	∑	Training	Testing/ Validation	∑	Training	Testing/ Validation	∑
AC_B	746	191	937	54	14	68	2,621	541	3,162
AC_N	24	6	30	10	7	17	200	231	431
KPMP (external validation)	0	80	80	0	5	5	0	240	240
HuBMAP (external validation)	0	9	9	0	5	5	0	198	198
∑	770	286	1,056	64	31	95	2,821	1,210	4,031

Supplementary Table 16. Overview of cases and their respective diagnoses on which annotations were performed. The number of cases where diagnoses overlap (i.e., simultaneous presence of two or more kidney diseases) is provided in brackets. Diagnoses only overlap in the specific training or test set. Non-specific cases mostly were biopsies with considerable interstitial fibrosis and glomerulosclerosis. In the external *KPMP* cohort only the presence of either acute kidney injury or chronic kidney disease was reported.
ABMR: Antibody mediated rejection; TCMR: T-cell mediated rejection; GN: Glomerulonephritis; AKI: Acute kidney injury; CKD: Chronic kidney disease.

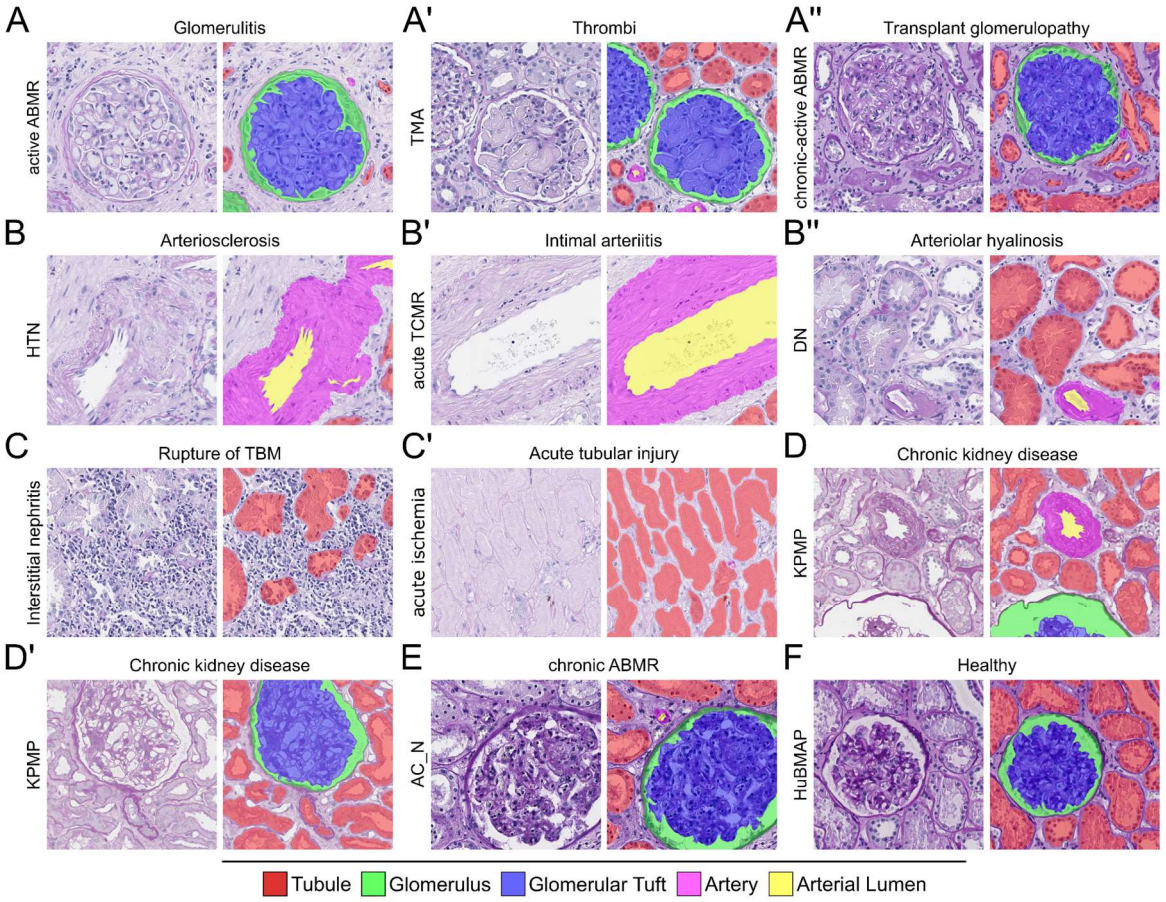
Cohort	Diagnosis	Cases Training (n)	Cases Test/Validation (n)
AC_B	ABMR	2	2 (1)
	TCMR	2	1 (1)
	Polyomavirus nephritis	1	-
	Immune-complex GN (not further differentiated)	1	-
	IgA nephropathy	18 (2)	3 (1)
	Lupus nephritis	5	2
	Membranous GN	-	1
	Membranoproliferative GN	1	-
	Pauci-immune GN	10	1
	C3 glomerulopathy	1	-
	Minimal change disease	2	-
	Hypertensive nephropathy	1 (1)	2 (2)
	Thrombotic microangiopathy	3	-
	Diabetic nephropathy	6 (3)	2 (2)
	Amyloidosis	1	-
	Non-specific (chronic damage)	5	2
AC_N	Transplant nephrectomy	-	2
	Tumour nephrectomy	9	3
KPMP (external validation)	AKI	-	1
	CKD	-	4
HuBMAP (external validation)	Healthy nephrectomy	-	5



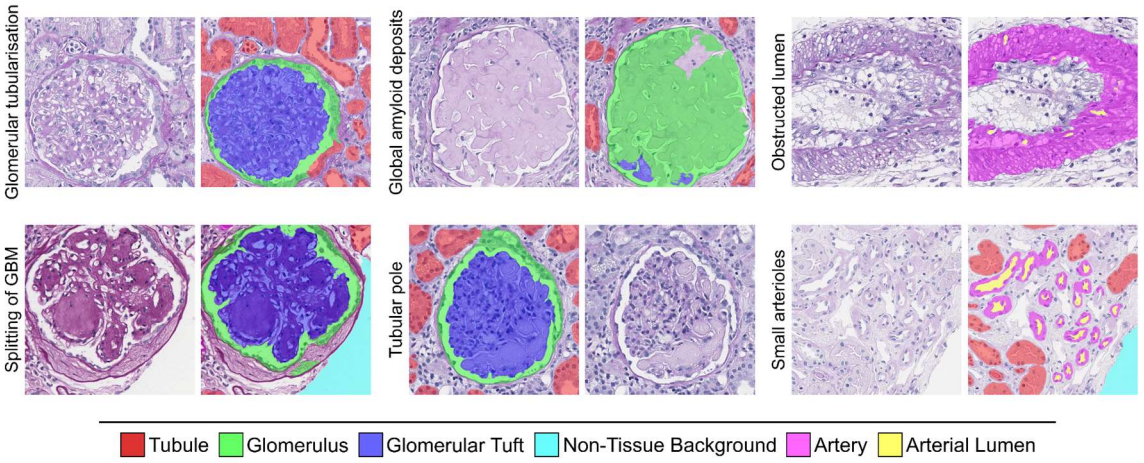
Supplementary Figure 1. Representative whole-slide images with prediction overlay (orange) of the tissue segmentation convolutional neural network. Kidney tissue from nephrectomy and biopsy specimens was precisely segmented. Non-kidney tissue of various origins such as connective tissue (A, A'), fat (A, A'), muscle (A, A'), fibrous capsule (C, C') and even liver tissue (B, arrowhead B') were not recognized as kidney tissue and therefore excluded.
Left (A-C) scale bar: 2mm; right (A'-C') scale bar: 0.5mm.



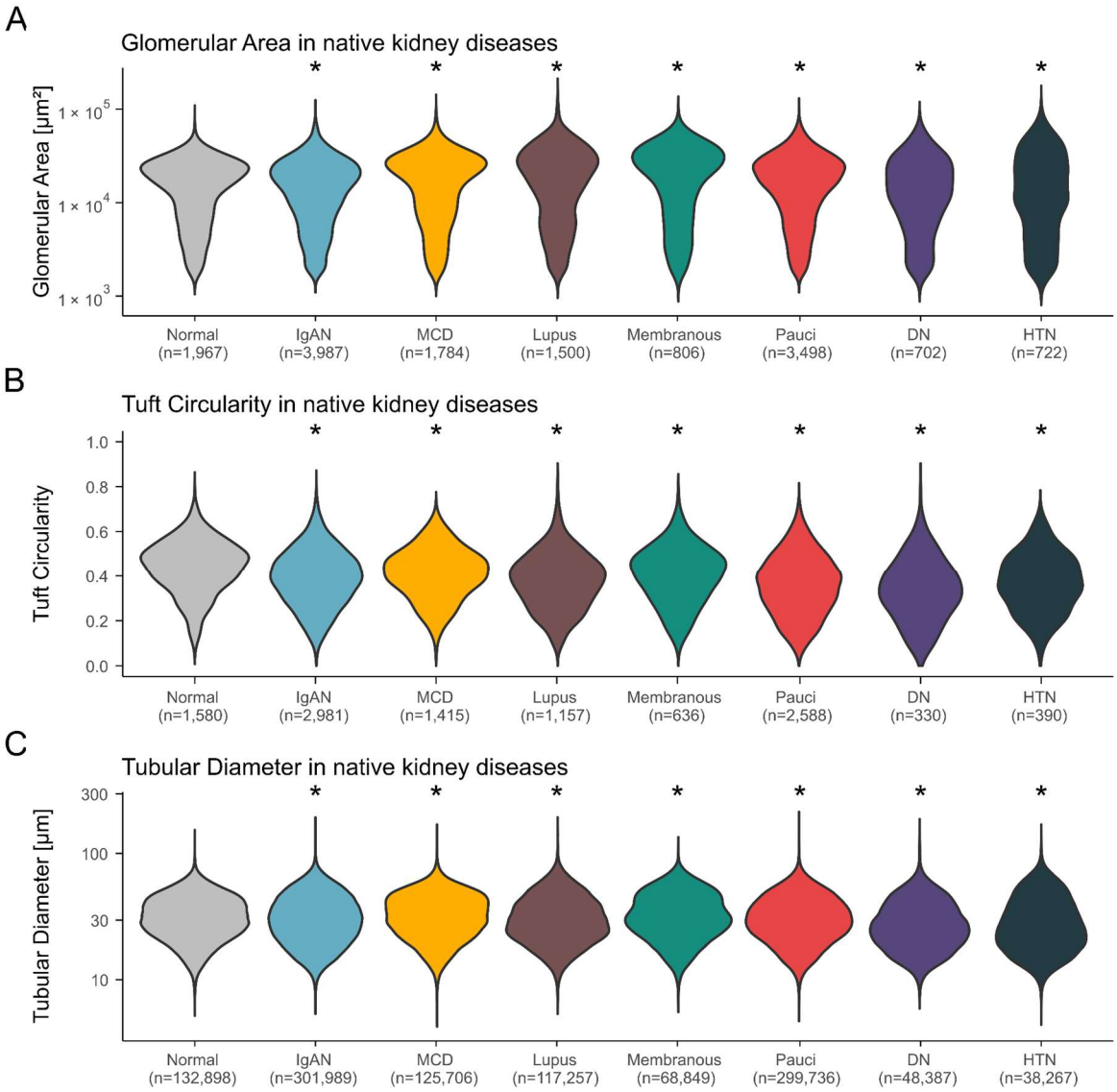
Supplementary Figure 2. Variation in Periodic Acid Schiff-staining between the five included cohorts. Internal cohorts (*AC_B* & *AC_N*) are combined due to similar staining and cutting routines. All patches have an edge length of 300µm.



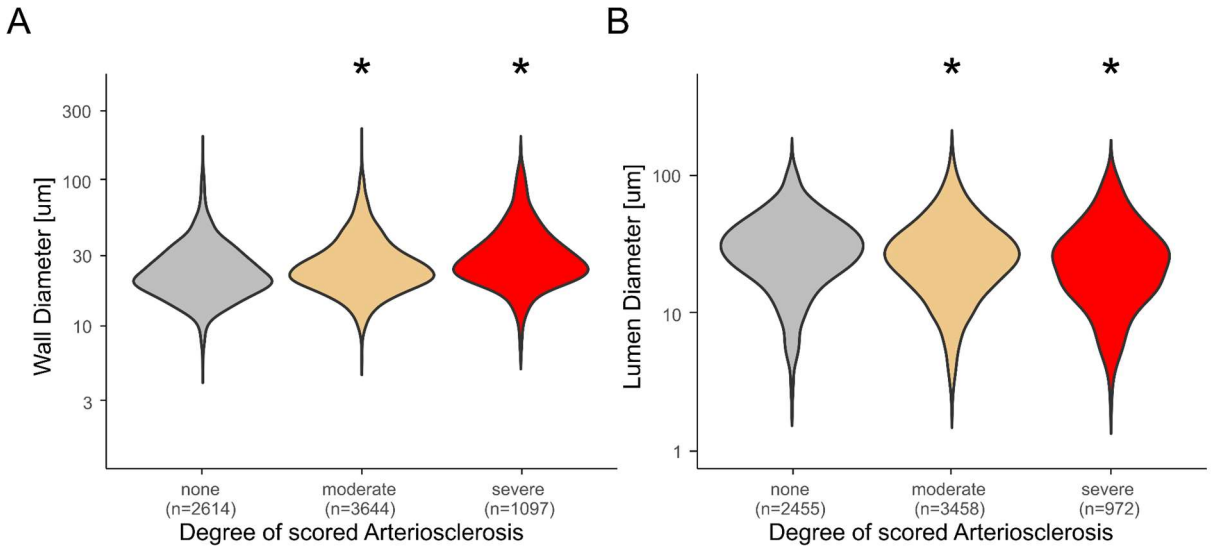
Supplementary Figure 3. Additional representative segmentation visualisations for glomeruli (A-A''), arteries (B-B'') and tubules (C-C') as well as segmentations from the internal nephrectomy cohort (*AC_N*) and external cohorts (*KPMP* & *HuBMAP*). For biopsies from the *KPMP* cohort (D and D') only chronic kidney disease or acute kidney injury were reported as diagnoses. Despite major differences in cutting and staining protocols, histological structures were precisely segmented. All patches have an edge length of 300µm.
ABMR: Antibody-mediated rejection; TMA: Thrombotic microangiopathy; HTN: Hypertensive nephropathy; TCMR: T-cell mediated rejection; DN: Diabetic nephropathy; TBM: Tubular basement membrane.



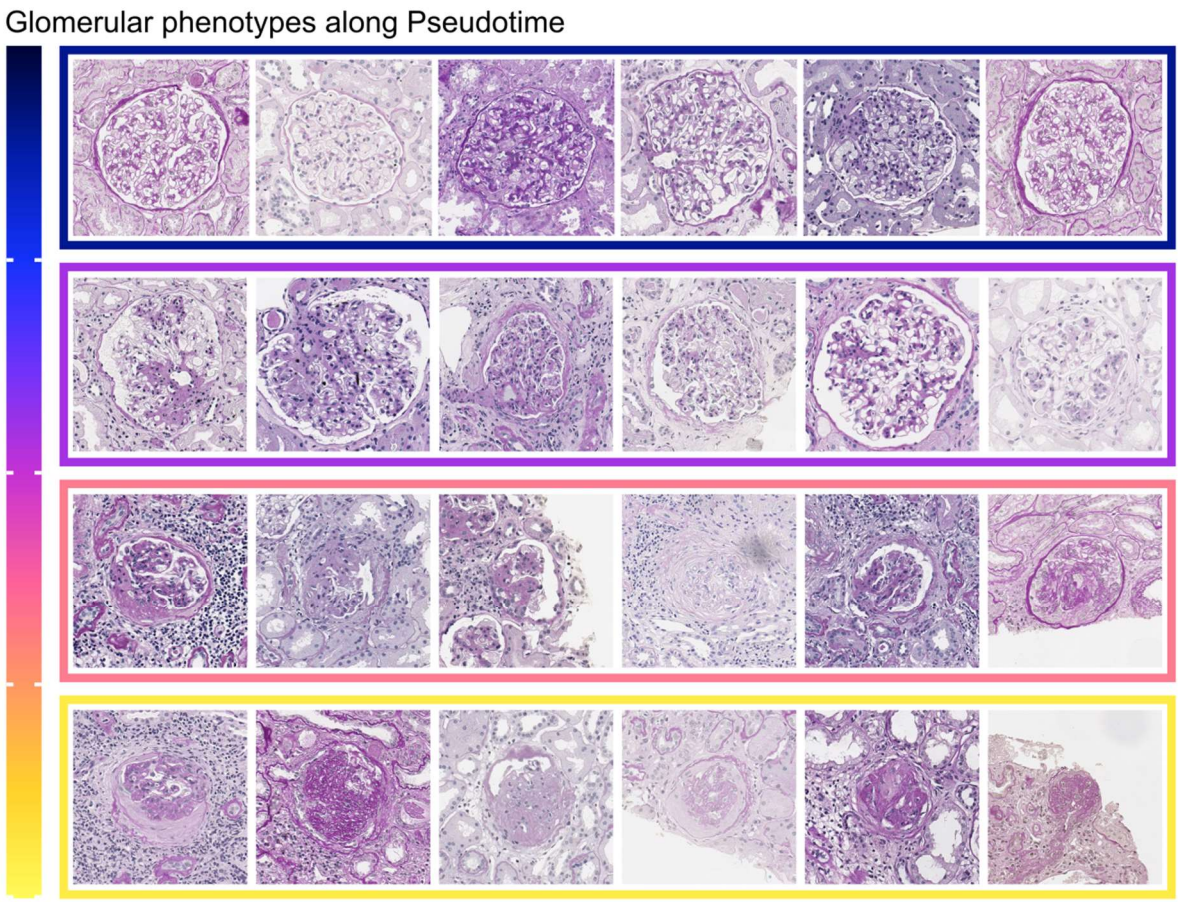
Supplementary Figure 4. Representative selection of imprecise segmentation results from the *AC_B*, *AC_N* and *KPMP* cohort of glomeruli, tubules and arteries. All patches have an edge length of 300µm.
GBM: Glomerular basement membrane.



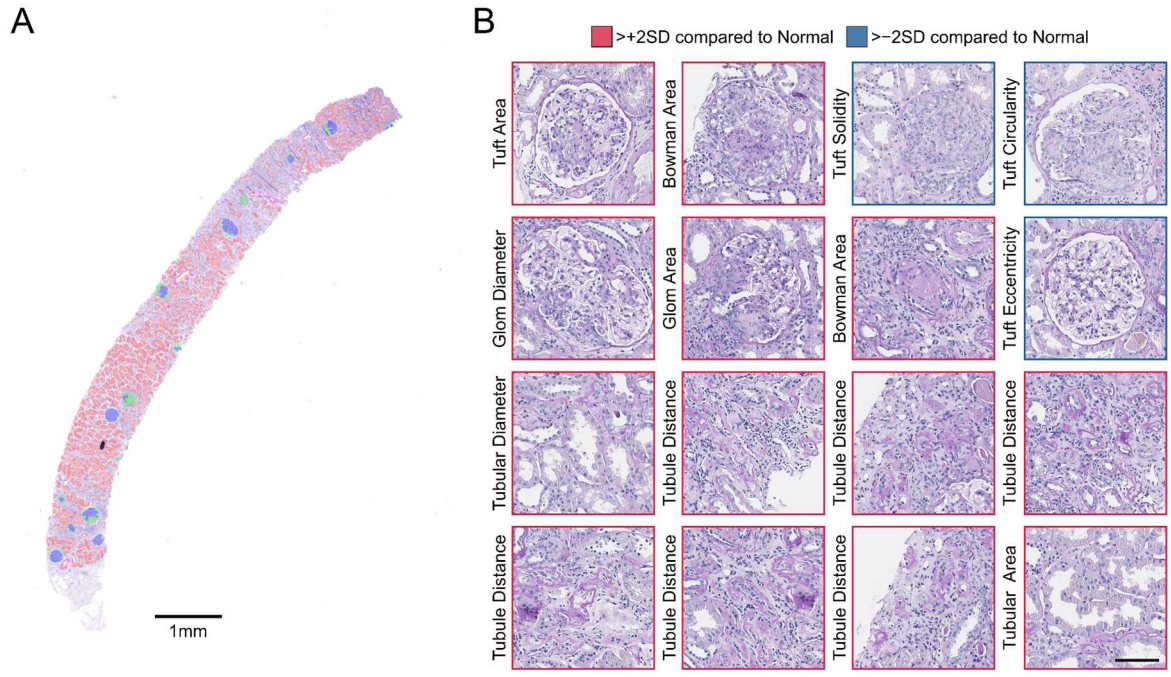
Supplementary Figure 5. Next-generation morphometry analysis for disease comparison. Analysis of (A) 14,966 glomeruli, (B) 11,077 glomerular tufts and (C) 1,133,089 tubules based on the reported diagnosis in the internal *AC_B* cohort. Different diseases display unique distributions for different features underlining the complex influence of native kidney diseases on its histomorphology.
IgAN: IgA nephropathy; MCD: Minimal change disease; Membranous: Membranous glomerulonephritis; Pauci: Pauci-immune glomerulonephritis, DN: Diabetic nephropathy; HTN: Hypertensive nephropathy; eGFR: estimated glomerular filtration rate. * = $p < 0.05$.



Supplementary Figure 6. Analysis of artery wall and lumen diameter based on the degree of arteriosclerosis scored by two nephropathologists. Wall and lumen diameters were assessed on instance-level. The reported degree of arteriosclerosis was assigned as a global label to all arteries and arterioles present in the whole-slide images of the associated biopsy. For the analysis of lumen diameter arteries without a lumen were excluded. * = $p < 0.05$.



Supplementary Figure 7. Additional examples of glomerular phenotypes from the pseudotime trajectory analysis (blue - healthy phenotype to yellow - very diseased phenotype) of the *VALIGA* trial. Colouring of borders displays the relative position of the displayed structures along the trajectory. All patches have an edge length of 300µm.



Supplementary Figure 8. Whole-slide segmentation and automated visualisation of feature outliers. (A) Segmentation of a representative IgA nephropathy case from the *AC_B* cohort. (B) Visualisation of patches which display at centre a glomerulus, glomerular tuft or tubule including additional context. Structures in the centre of the patch morphometrically diverge two standard deviations above (red) or below (blue) the feature specific distribution in normal biopsies. Scale bar for patches of extracted structures (B): 100µm

Additional literature

1. Levey, A. S. *et al.* A new equation to estimate glomerular filtration rate. *Ann. Intern. Med.* **150**, 604–612 (2009).
2. Bankhead, P. *et al.* QuPath: Open source software for digital pathology image analysis. *Sci. Rep.* **7**, 16878 (2017).
3. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).
4. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* 234–241 (Springer International Publishing, 2015).
5. Liu, L. *et al.* On the Variance of the Adaptive Learning Rate and Beyond. *arXiv [cs.LG]* (2019).
6. Mingqiang, Y., Kidiyo, K. & Joseph, R. A survey of shape feature extraction techniques. in *Pattern Recognition Techniques, Technology and Applications* (InTech, 2008).
7. Hothorn, T. & Lausen, B. Maximally selected rank statistics in R. *R News* (2002).
8. Roufosse, C. *et al.* A 2018 Reference Guide to the Banff Classification of Renal Allograft Pathology. *Transplantation* **102**, 1795–1814 (2018).
9. Sethi, S. *et al.* Mayo Clinic/Renal Pathology Society Consensus Report on Pathologic Classification, Diagnosis, and Reporting of GN. *J. Am. Soc. Nephrol.* **27**, 1278–1287 (2016).