

1 Measurement matters: A Metrological

Approach to Renal Pre-implantation Biopsy

Evaluation to address uncertainty in organ selection.

Authors: John OO Ayorinde¹ MD (UK), Xavier Loizeau² PhD, Victoria Bardsley³ MD (UK), Spencer Angus Thomas², Marina Romanchikova² PhD, Alex Samoshkin⁴ PhD, and Gavin J Pettigrew^{*1} MD (UK).

Affiliations:

1. Department of Surgery, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge, CB2 0QQ, United Kingdom.

2. National Physical Laboratory, Hampton Road, Teddington TW11 0LW, UK

3. Department of Histopathology, Addenbrooke's Hospital, Hills Road, Cambridge, CB2 0QQ, United Kingdom

4. Office for Translational Research, School of Clinical Medicine, University of Cambridge, Box 111, Cambridge Biomedical Campus, Cambridge CB2 0SP

Corresponding author: Dr John OO Ayorinde, Level 9 Department of Surgery, Addenbrooke's Hospital, Hills Road, Cambridge, CB2 0QQ, United Kingdom

Email: jtoa2@cam.ac.uk | Tel: 07723351242

Competing Interests: The authors declare no conflicts of interest.

2 Author contributions

JOA, GP, and AS conceptualized the study. JOA and VB were responsible for data curation. JOA, SAT, XL and MR were responsible for formal analysis. GJP, AS, and JOA were responsible for funding acquisition. JOA, VB, MR, XL, SAT, and AS were responsible for investigation. JOA, VB, MR, XL, SAT, and GJP were responsible for methodology. JOA, XL, and MR were responsible for visualization. GJP, AS, VB, and MR were responsible for supervision. JOA wrote the original draft. All authors reviewed and edited the manuscript.

Competing interests

The authors declare no conflicts of interest

Financial disclosures

XL & MR are funded by the UK government Department of Science, Innovation and Technology through the National Measurement System programme. This work was supported by awards from the Medical Research Council (Confidence in Concept A094757) and Addenbrooke's Charitable Trust in the UK.

3 Abstract

Background: Pre-implantation biopsy combines measurements of injury into a composite index to inform organ acceptance. The uncertainty in these measurements remains poorly characterised, raising concerns variability may contribute to inappropriate clinical decisions.

Methods: We adopted a metrological approach to evaluate biopsy score reliability. Variability was assessed by performing repeat biopsies (n=293) on discarded allografts (n=16), using three methods (core, punch, wedge). Uncertainty was quantified using a bootstrapping analysis. Observer effects were controlled by semi-blinded scoring, and the findings validated by comparison with standard glass evaluation.

Results: The surgical method strongly determined the size (core biopsy area 9.04mm^2 , wedge 37.9mm^2), and therefore yield (glomerular yield $r=0.94$, arterial $r=0.62$) of each biopsy. Core biopsies yielded inadequate slides most frequently. Repeat biopsy of a single kidney revealed marked variation. In 10/16 cases, scores were contradictory, crossing at least one decision boundary (i.e., to transplant or to discard). Bootstrapping demonstrated the uncertainty associated with single-slide assessment, however, for paired kidneys from one donor, the distribution of scores were similar.

Conclusions: Our investigation highlights the risks of relying on single slide assessment for quantifying organ injury. Biopsy evaluation is subject to uncertainty, meaning each slide is better conceptualised as providing an estimate of the kidney's condition, rather than a definitive result. Pooling multiples assessments could be used to improve the reliability of biopsy analysis, enhancing confidence. Where histological quantification is necessary, clinicians should seek to develop new protocols that utilise more tissue, and should consider automated methods to assist pathologists to deliver analysis within clinical timeframes.

Keywords: Kidney Biopsy, Metrology, Remuzzi Score, Transplantation, Histological Assessment, Measurement Variability

4 Introduction

Kidney transplantation is the best treatment for individuals with end-stage renal disease (ESRD). As the incidence of ESRD increases¹⁻³, the growing demand for kidneys poses a major challenge⁴. Over half of potential deceased donors in the UK are now over 60 years old⁵⁻⁸, and whilst indiscriminate use of kidneys from older donors may lead to poor transplant outcomes, the discard of potentially suitable organs from this group exacerbates constraints on organ supply⁷⁻¹⁰.

Current methods of assessing kidneys rely on donor history and visual inspection, yet even strongly predictive factors signify cohort, rather than individual risk, and there may be significant intra-cohort variability in outcomes^{6,9,11-15}. Similarly, the relationship between the macroscopic appearance and transplant outcomes remains ill-defined^{16,17}.

Quality assessment technologies, such as histopathological analysis or machine perfusion, may aid in organ selection by providing an additional, individualised data point. With histopathological analysis, the decision to implant or discard is informed by assessing the extent of chronic damage present on biopsy^{9,10}. Although validation of this approach through a prospective randomised study is awaited (e.g., PITHIA¹⁸), several studies have reported an association between pre-implantation or implantation biopsy assessment and transplant outcome, most notably in Remuzzi's seminal studies^{19,20}. Consequently, pre-implantation biopsy analysis has become widely adopted, with most retrieved organs undergoing biopsy in the USA, up to 78% in some areas^{5,9,21-29}.

Remuzzi quantifies each anatomical component of the kidney (glomeruli, tubules, arterioles and interstitium), providing a total score that determines the implantation strategy (implant each kidney singly into two recipients; implant both into one recipient; or discard both kidneys). This was devised from a theoretical consideration, of providing sufficient nephron mass to the recipient, but its introduction occurred without an in-depth consideration of the metrological aspects of the biopsy procedure, including variability associated with biopsy acquisition and subsequent assessment²⁰. Moreover, to be useful for decision making, clinicians need to assume that the biopsy score has negligible measurement uncertainty, and is representative of the overall quality of the kidney^{30–32}.

Awareness of the metrological attributes of an single biopsy result could potentially lead to different implantation decisions. Metrology provides established tools and methods to both quantify and reduce measurement uncertainty^{33–36}, yet these tools have been underutilised for histopathological assessment. Studies of chronic injury scoring to-date have mainly addressed agreement between pathologists^{37–42}, and different surgical approaches to performing the biopsy^{43–45}, but not the variance inherent to scoring, which can only be assessed by repeated measurement of biopsy scores. Given the widespread use of biopsy analysis and the potential harm from inappropriate use (or discard) of the kidney, this is a surprising oversight with real clinical relevance. In this work, we apply metrological concepts such as comparison and uncertainty to quantify the variabilities in Remuzzi scoring, using a series of discarded kidneys.

5 Methods

5.1.1 Biopsy technique and slide preparation

Two study cohorts were developed: experimental and historical. The experimental cohort involved repeat biopsies on 16 discarded renal allografts (Figure 1). Kidneys were discarded for clinical reasons: donor quality (n=7), high biopsy score (n=1), suspected malignancy (n=3), suboptimal perfusion (n=2), renal artery thrombosis from retrieval damage (n=2), and complex cystic lesions (n=1). Eight surgeons with retrieval experience performed 9 biopsies per kidney (3 core; 3 punch and 3 wedge) which were then processed into glass slides for digitally scanning and assessment.

Core biopsies were performed using an disposable spring-loaded guillotine soft tissue needle (18G 2BioPince[®] Full Core Biopsy Instrument), inserted at a 15-20-degree angle, yielding a 19 mm specimen length. Wedge biopsies involved resecting an ellipse of renal cortex using a scalpel. Punch biopsies used a 5mm Schuco Stiefel Biopsy Punch[®] which limited depth to 8mm.

Samples were processed according to clinical protocols, i.e., placed in 10% neutral buffered formalin and embedded in wax. This mirrors UK practice where biopsies are performed at retrieval by the National Organ Retrieval Service (NORS) team and processed in parallel with organ transport.¹⁸.

In this study, embedded tissue blocks were sectioned to maximise surface area for slide assessment. To match time-limited clinical practice, two histological sections from each biopsy were prepared: one stained with Hematoxylin & Eosin (HE) and the other with Periodic acid–Schiff (PAS). Slides were scanned using a 3DHIS-TECH

Panoramic DESK scanner at high resolution (0.12 $\mu\text{m}/\text{pixel}$, minimum 40x objective resolution), and annotated using Sysmex Caseviewer v2.3.

Out of 288 images from 144 biopsies (Table 1), 38 images were excluded due to inappropriate processing resulting in folded or incomplete specimens, or not displaying the largest cross-sectional area, leaving 250 images for final analysis. To compare the experimental biopsies with those taken for clinical purposes, we developed a comparator cohort of 1044 historical biopsies.

5.1.2 Biopsy assessment

We compared unblinded and semi-blinded methods of biopsy scoring. In the unblinded condition, images were assessed on-microscope by an experienced renal histopathologist, as in current practice. This method was considered unblinded, because the first assessment might cause anchoring bias on later assessments from the same kidney or patient ^{46,47}. To minimise bias, we developed an alternative method of semi-blinded assessment. Digital slides were evaluated by a non-specialist pathologist, supervised by an experienced renal pathologist at our centre. Each evaluation focussed on a single component (e.g., glomerulosclerosis) on a clean digital slide, with slides evaluated in random order (750 measuring events). Chronic injury features were directly measured on the digital slide, and data were entered into a database to generate the Remuzzi scores concurrently for all slides at the end of the annotation period.

The following measurements were made: Glomeruli: All glomeruli present were assessed, determining the proportion of globally sclerotic glomeruli from the total number.

Arteries: Up to two arteries per image were assessed, defined as having two or more layers of smooth muscle, or a diameter $>130\mu\text{m}$. If multiple arteries were present, the “worst” artery was used. Arteriosclerosis was determined by the ratio of wall and lumen diameter. Wall thickness was taken as the mean of opposing walls, correcting for non-transverse views. Score boundaries for the lumen-to-wall ratio were converted to definite thresholds to resolve ambiguities in Remuzzi’s original description (0 = (0% to 50%), 1 = (50, 80), 2 = (80, 120), 3 = (120, ∞)).

Interstitial fibrosis & Tubular atrophy (IFTA): Borders were drawn around detected areas of IFTA. We used the approach adopted during the PITHIA trial¹⁸ of censoring small areas (<5%) of fibrosis to return a score of 0, not 1. Adequacy: As in Remuzzi et al.'s original description, adequate images were defined as having at least 25 glomeruli and one scorable artery²⁰. Decision making used a modification of Remuzzi based on the Cambridge experience⁹: scores 1-4 for single transplantation, 5-6 for dual transplantation, and 7 or higher for discard. This methodology aligns with national protocols, and aims to reduce unnecessary discard and maximise organ utility¹⁸.

5.1.3 Statistical methods

To quantify the strength of relation between continuous variables from repeated measurements, we report Pearson's correlation coefficient (r). For numerical measures of agreement for continuous variables, we report the Concordance Correlation Coefficient (CCC) proposed by Lin, which accounts for both covariance and deviation from the 45° line of perfect agreement⁴⁸. For categorical variables we use weighted kappa coefficients⁴⁹. Strength of agreement is interpreted as: 0-0.2 'slight', 0.2-0.4 'fair', 0.4-0.6 'moderate' 0.6-0.8 'substantial', 0.8-1 'perfect'³⁰.

For comparing the sample means of two groups, we use Student's t-test, and chi-squared test for categorical variables. For comparing sample means for three or more groups, we use Analysis of variance (ANOVA)⁵⁰.

To quantify uncertainty and capture the variability in the Remuzzi score, we conducted a bootstrap resampling analysis. This non-parametric approach creates empirical distributions by sampling the experimental population with replacement^{51,52}. Each kidney's Remuzzi scores were resampled with replacement to create subsamples. We performed 1000 bootstrap iterations, with each subsample containing the same number of scores as the original data. This allowed us to compute statistics such as the mean and standard deviation for each subsample and derive distributions for these without assuming normality.

All tests were considered as significant when the p-value was less than 0.05. Tests were performed using Python v3.6 with Numpy v1.16, Scipy v1.13, and Scikitlearn v1.0.2⁵³⁻⁵⁵.

6 Results

6.1.1 Biopsy evaluation and comparison of surgical approaches

We made measurements from the semi-blinded approach to evaluate the physical characteristics of biopsies obtained through different surgical techniques. The choice of biopsy significantly affected the cross-sectional area available for assessment (Table 2 & Figure 2). Core biopsies yielded the smallest samples, while wedge biopsies were largest on average (punch vs wedge, $p=0.001$, Table 2 & Figure 2a). Glomerular yield correlated with cross-sectional area (Figure 2b, $r=0.94$), with core biopsy yielding the fewest and wedge biopsy the most glomeruli. The number of arteries yielded also correlated with biopsy size (Figure 2c, $r=0.62$), though some large biopsies yielded no arteries. The extent of glomerulosclerosis was only weakly correlated with biopsy size (Figure 2d).

Next, we generated Remuzzi scores for each biopsy image, as detailed in the methods section. We grouped and compared the Remuzzi scores for each kidney according to biopsy technique. This comparison assessed whether the biopsy technique affected Remuzzi scores and, consequently, the transplant recommendation. Weighted Kappa coefficients from paired samples confirmed substantial agreement in assessment across all surgical techniques (Table 3). There was no evidence of a relationship between surgical method and the resulting transplant recommendation (chi-squared test, $p=0.41$).

6.1.2 Biopsy quality

Using Remuzzi's criteria for adequacy^{19,20}, 40% of images in our series were inadequate: 24% failed to meet the glomerular threshold, and 42% had no scoreable artery. Table 2 shows that adequacy varied by biopsy technique, with wedge and punch biopsies achieving adequacy of 76%, while only 27.5% of core biopsies were deemed adequate. This is largely due to core biopsies providing insufficient glomerular yield (glomerular adequacy 30%; Figure 2b); arterial adequacy was 52%. There was no discernible pattern of a patient (kidney) or surgeon effect on adequacy rates (Figure S1).

To ensure that the quality and technique of the biopsies in our experimental cohort were consistent with clinical settings (particularly for wedge biopsies), we compared the physical characteristics of experimental biopsies against a historical cohort of approximately 1,000 clinical biopsies taken over a 10-year period (Figure S2). The physical characteristics and adequacy rates were similar between the experimental and clinical cohorts.

6.1.3 Assessment reliability and quantifying variability with bootstrapping

To maximize the reliability of Remuzzi scoring, we employed a semi-blinded approach (Figure 3). Each kidney provided up to 9 samples (stained H&E and PAS), with both kidneys from a single donor being available in 4 instances (donors 1, 2, 10, and 11).

We observed significant variation in Remuzzi scores among repeat samples from the same kidney, despite uniform processing and only a single assessor (Figure 3). Of note, for 10 out of 16 kidneys, these variations would have led to different clinical decisions regarding implantation or discard. This unexpected and potentially clinically

relevant finding emphasised the need to better understand, and robustly quantify the observed variability in Remuzzi scores.

In our analysis, we used bootstrapping to quantify variability. Bootstrapping is a powerful statistical technique where random samples are drawn with replacement from the original data. “With replacement” means that an individual datapoint can be included more than once in each subsample, ultimately to generate thousands of simulated datasets. This helps overcome the limited number of repeat samples in our study, allowing for estimation of population statistics without assuming normal data distribution.

Through bootstrapping, we determined that the variability in Remuzzi scores likely reflects inherent sampling variability as opposed to underlying differences in chronic injury. This was substantiated by statistically similar distributions of scores between left and right kidneys from the same donor (e.g., Mann-Whitney test PL10 vs PR10; $p=0.37$), and similar means and variances in bootstrapped distributions. The range of agreement on transplant outcome within pairs for bootstrapped samples was between 62-81% (Table 4, left section). However, statistical dissimilarity was observed when comparing kidneys from different donors (e.g., PL/R10 vs PL/R11; $p=0.004$).

Furthermore, bootstrapping demonstrated that incorporating multiple measurements can reduce measurement uncertainty and enhance confidence. Each kidney exhibited a unique probability distribution of Remuzzi scores for each sampled set, consistent between surgical methods (Figures S3-6). For some kidneys (e.g., PR7), this distribution was narrow, whilst for others (e.g., PR5), sample scores were more widely spread across decision thresholds. We propose that this may represent differing

levels of ‘confidence’ in determining the most appropriate outcome. Using the mean of each sampled set as the estimate (Figure 4) reduced the spread of estimates and the proportion of assessments crossing decision-making thresholds (Table 4, left section vs right), leading to increased agreement on clinical recommendations. The exception was pair P11, which demonstrates the rule: assessment means and variances for each kidney were similar, but because PR11's distribution lies in close proximity to the upper Remuzzi threshold, there were a substantial proportion of scores crossing categories, and causing within pair disagreement (Table 4 right section, pair match 51%).

6.1.4 Pathologist validation of reliability findings

Semi-blinded assessment was implemented to mitigate anchoring effects and to serve as a bridge to automated slide analysis⁵⁶. However, this approach raised concerns about the introduction of new biases, particularly in the assessment of IFTA – where normally visual approximation is used instead of measurement of bounded pixel area - and arteriosclerosis, where the boundaries between classes in Remuzzi's framework are ambiguous.

To evaluate the validity of semi-blinded assessments, a renal histopathologist independently assessed the original glass slides (Figure 5). Notably the pathologist's assessments were conducted using both PAS and H&E stains, in line with clinical practice, effectively halving the number of assessments compared to the semi-blinded approach. Despite every advantage given the lack of blinding, and making fewer assessments, we found significant assessment variability for individual kidneys, which crossed decision boundaries in 6 out of 16 cases. Since increasing the number of assessments increase the chance of any two assessments disagreeing, the reduced number of instances where decision boundaries were crossed might simply be attributable to the fact that the pathologist assessed fewer cases. Importantly we observed a similar pattern of scoring for each kidney, with similar median scores to the semi-blinded cohort.

6.1.5 Visualisation and analysis of consistency between semi-blinded and unblinded methods

To examine the consistency between the unblinded and semi-blinded assessment methods, we categorized each digital slide based on the injury category assigned by

the pathologist (0-3) and visualised the measurements made for that slide (Figure 6). For this analysis, we focused on glomerulosclerosis and IFTA, as these are continuous measurements that apply to the entire slide, rather than a specific object as in arteriosclerosis. The Remuzzi score for IFTA comprises two separate assessments – interstitial fibrosis and tubular atrophy. Although we derived separate values for interstitial fibrosis and tubular atrophy these scores did not diverge in either the experimental or clinical cohorts. Hence here they are considered as a single measure of injury.

In the case of perfect agreement between the two assessment methods, one would have expected data points to cluster within the Remuzzi classes of injury delineating each decision boundary. However, for both glomerulosclerosis and IFTA, we observed that whilst average injury measurements did increase with the assigned category of injury, there was significant overlap between classes.

7 Discussion

7.1 Impact of surgical technique on biopsy quality and assessment

Surgical technique may influence biopsy assessment because of differences between the physical profiles and sampling locations of each method^{25,31,38–41,44,57–60}. Core biopsy, the standard approach, has two main disadvantages: it provides a small volume of tissue and skill is required to avoid damage to hilar structures, or sampling medulla rather than cortex. The main alternative is wedge biopsy, where a scalpel is used to resect an ellipse of cortex. However, without standard measurements or anatomical landmarks, samples vary greatly. A minority are ‘safe’, but small, subcortical resections that are not useful for quality assessment, while extremely large, deep biopsies risk the safety of the transplant. In 2012 Bargo-Horvath proposed the punch biopsy, which provides a consistent barrel of renal cortex, and limits insertion safe depth to 8 mm⁴⁵. Despite these advantages, punch biopsy is not widely used, and current knowledge is limited to retrospective comparisons with historical cohorts. Differences in sampling location may also introduce bias. For instance, Muruve et al. found that small wedge biopsies overestimate glomerulosclerosis by sampling the subcapsular region where sclerosed glomeruli are more prevalent⁴³, a finding confirmed by Rule et al⁶¹. Researchers are yet to reach consensus on the best surgical method. Core and wedge biopsies have been extensively studied, with some groups advocating core biopsies for evaluating vascular lesions^{40,58,62} while others have favoured wedge biopsies for predicting clinical outcomes^{44,63}.

To our knowledge this is the first prospective comparison of all three biopsy methods. Despite theoretical concerns, our analysis found substantial agreement between each method (Table 3), aligning with Mazzucco et al.'s comparison of 'core' and 'wedge' biopsies to whole kidney sectioning ($k=0.57$ wedge vs. section; $k=0.73$ core vs. section score)⁴¹. However the choice of technique did affect the quality of sample provided. Core biopsies consistently produced the smallest samples (Figure 2a), and in our study that led to a slide adequacy of only 26% because of the correlation between biopsy size and glomerular yield (Figure 2b). Inadequate sampling is a concern because of the risk of repeat biopsy causing unnecessary complications and prolonging cold ischaemia.

7.2 Reproducibility

Our study revealed significant variation in chronic injury on repeated assessment, and this persisted whether the slide was assessed as standard, or by direct measurement of chronic injury features on the slide. The magnitude of variability was clinically significant, and would have led to contradictory recommendations regarding utilisation or discard. Signs of this phenomenon have been recognised previously (e.g. left vs right discrepancy), but the cause has not been reliably identified⁶⁴. Numerous studies have remarked on inconsistent agreement between pathologists³⁰. Indeed, pathologist experience has been repeatedly identified as a factor influencing the degree of inter-observer agreement, suggesting that variability can be reduced by training or experience^{37,39}.

In contrast, the role of measurement in variability has been underexplored. When scoring, discretion is used in selecting and in measuring which can introduce variation.

In the previous example, measurement style might vary between specialist renal and general pathologists. Additional variation could arise from heterogeneity of disease expression in the kidney cortex, fixation effects, staining impacts on injury visibility, time constraints, or subjective understanding of assessment criteria.

Our analysis suggests that each slide contributes an estimate of chronic injury which lies within a probability distribution. The quality of the kidney is therefore represented by the distribution of scores it receives, rather than by any single result. This better reflects reality, and usefully, provides a way of quantifying assessment confidence. We propose that confidence is inversely related to the percentage of estimates that cross decision boundaries; such that high levels of confidence are seen with a narrow distribution of scores which lie within a single injury category. Using this framework, we could demonstrate the value of combining multiple measurements (pooling) to reduce variance (Table 4), thus increasing confidence and reducing the frequency of contradictory recommendations within kidney pairs. Confidence could be used in future studies to guide interventions aimed at minimising variability. Additionally, our approach addresses variability in clinical settings, since our data supports clinicians to avoid overly dogmatic interpretations of chronic injury assessment, as well as to avoid attributing differences in Remuzzi scores (particularly within pairs) solely to biological factors.

7.3 Limitations

Discarded kidneys were necessary to allow repeated biopsy, but their use may limit the generalisability of our findings. Although we assessed a comparable number of biopsy slides to other major studies in this area (n=250), they were derived from only 16 kidneys, potentially introducing bias. Future studies should systematically assess hundreds or thousands of slides, likely using supervised automated assessment methods^{56,65,66}. Moreover, the 18-gauge needles used in this study reflected NORS practice per the QUOD protocol, and may have reduced sample quality compared to 16-gauge needles which provide better tissue sampling⁶⁷⁻⁷⁰. Future work should employ 16-gauge needles to improve sample quality.

As the kidneys were not transplanted, we could not assess the impact of large wedge biopsies on transplantability, or bleeding, which are important clinical concerns.

Additionally, our methods may not fully align with clinical practices. For example, in hours, our centre processes and stains biopsies at a minimum of 10 levels, allowing pathologists to search for an artery within the stack if one is not found on the initial slide. By reporting each slide individually, we may overestimate the clinical likelihood of an inadequate sample for the whole biopsy, potentially explaining the disparity in adequacy rates between our clinical and experimental cohorts (approximately 95% vs. 75%, Table S1).

Lastly, our analysis relied solely on the Remuzzi assessment score. Whilst other scoring systems, such as Banff, are used in transplant pathology, we selected Remuzzi for its focus on chronic injury features (vs rejection) and organ evaluation. Remuzzi offers clear thresholds, simplicity, and potential compatibility with other

donor risk assessments. Nonetheless, our metrological insights should to apply to any method of histopathological quantification.

7.4 Implications for future practice

Unlocking the full potential of biopsy assessment requires a deeper understanding of measurement error. Systematic investigations into biopsy metrology are needed to identify and reduce the major contributors to variation. Remuzzi's original description contains obvious candidates for study and optimisation, including:

- that the analysed tissue safely represents the entire organ,
- the chosen measurands appropriately predict organ quality,
- the thresholds dividing scores effectively separate organs with differing quality,
- the measured quantities are combined and weighted optimally,
- measurement uncertainties do not affect decision making.

The impact of metrology on interpretation is clear. If biopsy scores are highly reproducible, then kidneys with conflicting scores from the same donor ought to have different implantation strategies. However, given the observed low reproducibility as in this study, differences could merely reflect variance. Clinicians should use both sources of information and average contradictory scores, unless there are clinical indicators to suspect unilateral dysfunction (e.g., large size discrepancy).

Pooling samples is a promising strategy to address the deficiencies of single slide assessment, yet to achieve high levels of reliability (i.e. >95% chance repeat testing gives an equivalent score, or >99% chance the same decision) means analysing much more tissue than currently. However, implementing multi-slide and multi-biopsy protocols using traditional methods may not be practical. Advancements in digital

pathology and automation offer an opportunity to reconsider the time and effort available for urgent biopsy analysis^{56,66,71–78}. Interdisciplinary collaboration among clinicians, metrologists, and data scientists could facilitate the introduction of technologies to greatly improve decision-making. Automated slide analysis could transform the gold standard of pathology review, reduce variability and improve predictive power through exhaustive whole-biopsy analysis⁶⁵; working within time constraints by performing assessments in parallel^{56,66,71}.

7 Conclusion

Our study highlights the limitations of relying on single slide assessments for organ selection. We observed that biopsy quality is significantly influenced by surgical technique, and demonstrated the need to identify and reduce other sources of variation. Whilst acknowledging the limitations of this study, our findings suggest that there are opportunities to improve histological assessment as a quantification tool, through a comprehensive evaluation of histopathological practices, optimisation of current assessment systems, and rigorous metrological analysis. Advances in computer vision could enable analysis of much more renal tissue within clinical timeframes.

Improving measurement standards to the level seen in other areas of routine medical care (e.g. blood tests) is likely to be cost-effective and greatly improve patient outcomes.

8 Acknowledgements

This work would not have taken place without the determined support of the Office for Translational Research, University of Cambridge. This study was also supported by the National Institute for Health and Care Research (NIHR) Blood and Transplant Research Unit in Organ Donation and Transplantation (NIHR203332), a partnership between NHS Blood and Transplant, University of Cambridge and Newcastle University. The views expressed are those of the author(s) and not necessarily those of the NIHR, NHS Blood and Transplant, or the Department of Health and Social Care

9 Ethics Approval and Consent to Participate

Ethics Approval and Consent to Participate: This study received ethics approval from the Cambridge Research Ethics Committee (REC), under the reference numbers NRES 19/EE/0022 for the study's main ethical considerations and NRES 12/EE/0202 specifically for the use of discarded renal grafts. All protocols were conducted in strict adherence to the 2013 Helsinki Declaration. Written, informed consent was obtained from all study participants, or from their next of kin in the case of deceased individuals. Where individual data, including details, images, or videos, were used, additional written consent for publication was secured. All reasonable measures have been taken to ensure patient anonymity in the publication.

10 Data availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

11 References

- 1 Hill CJ, Fogarty DG. Changing Trends in End-Stage Renal Disease Due to Diabetes in the United Kingdom. *J Ren Care* 2012; **38**: 12–22.
- 2 Patzer RE, McClellan WM. Influence of race, ethnicity and socioeconomic status on kidney disease. *Nat Rev Nephrol* 2012; **8**: 533–41.
- 3 van Walraven C, Manuel DG, Knoll G. Survival trends in ESRD patients compared with the general population in the United States. *Am J Kidney Dis Off J Natl Kidney Found* 2014; **63**: 491–9.
- 4 Ayorinde JOO, Saeb-Parsy K, Hossain A. Opportunities and Challenges in Using Social Media in Organ Donation. *JAMA Surg* 2020; published online June 17. DOI:10.1001/jamasurg.2020.0791.
- 5 Summers DM, Johnson RJ, Allen J, *et al.* Analysis of factors that affect outcome after transplantation of kidneys donated after cardiac death in the UK: a cohort study. *The Lancet* 2010; **376**: 1303–11.
- 6 Mallon DH, Riddiough GE, Summers DM, *et al.* Successful Transplantation of Kidneys From Elderly Circulatory Death Donors by Using Microscopic and Macroscopic Characteristics to Guide Single or Dual Implantation. *Am J Transplant* 2015; **15**: 2931–9.
- 7 Callaghan CJ, Harper SJF, Saeb-Parsy K, *et al.* The discard of deceased donor kidneys in the UK. *Clin Transp* 2014; **28**: 345–53.
- 8 Callaghan CJ, Mumford L, Pankhurst L, Baker RJ, Bradley JA, Watson CJE. Early Outcomes of the New UK Deceased Donor Kidney Fast-Track Offering Scheme. *Transplantation* 2017; **101**: 2888–97.
- 9 Kosmoliaptsis V, Salji M, Bardsley V, *et al.* Baseline donor chronic renal injury confers the same transplant survival disadvantage for DCD and DBD kidneys. *Am J Transplant* 2015; **15**: 754–63.
- 10 Mittal S, Adamusiak A, Horsfield C, *et al.* A Re-evaluation of Discarded Deceased Donor Kidneys in the UK: Are Usable Organs Still Being Discarded? *Transplantation* 2017; **101**: 1698–703.
- 11 Watson CJE, Johnson RJ, Mumford L. Overview of the Evolution of the UK Kidney Allocation Schemes. *Curr Transplant Rep* 2020; **7**: 140–4.
- 12 Rao PS, Schaubel DE, Guidinger MK, *et al.* A Comprehensive Risk Quantification Score for Deceased Donor Kidneys: The Kidney Donor Risk Index. *Transplantation* 2009; **88**: 231–6.

- 13 Pérez-Sáez MJ, Montero N, Redondo-Pachón D, Crespo M, Pascual J. Strategies for an Expanded Use of Kidneys from Elderly Donors. *Transplantation* 2017; **101**: 727–45.
- 14 Sutherland AI, IJzermans JNM, Forsythe JLR, Dor FJMF. Kidney and liver transplantation in the elderly. *Br J Surg* 2016; **103**: e62–72.
- 15 Fehr T. Elderly kidney donors: better than their reputation-but where are the limits? *Transpl Int* 2018; **31**: 700–2.
- 16 Dare AJ, Pettigrew GJ, Saeb-Parsy K. Preoperative assessment of the deceased-donor kidney: From macroscopic appearance to molecular biomarkers. *Transplantation* 2014; **97**: 797–807.
- 17 Ayorinde JOO, Hamed M, Goh MA, *et al*. Development of an objective, standardized tool for surgical assessment of deceased donor kidneys: The Cambridge Kidney Assessment Tool. *Clin Transplant* 2020; **34**: e13782.
- 18 Ayorinde JOOO, Summers DM, Pankhurst L, *et al*. PreImplantation Trial of Histopathology in renal Allografts (PITHIA): A steppedwedge cluster randomised controlled trial protocol. *BMJ Open* 2019; **9**: 1–10.
- 19 Remuzzi G, Cravedi P, Perna A, *et al*. Long-Term Outcome of Renal Transplantation from Older Donors. *N Engl J Med* 2006; **354**: 343–52.
- 20 Remuzzi G, Grinyò J, Ruggenenti P, *et al*. Early experience with dual kidney transplantation in adults using expanded donor criteria. Double Kidney Transplant Group (DKG). *J Am Soc Nephrol* 1999; **10**: 2591–8.
- 21 Wang CJ, Wetmore JB, Crary GS, Kasiske BL. The Donor Kidney Biopsy and Its Implications in Predicting Graft Outcomes: A Systematic Review. *Am J Transplant* 2015; **15**: 1903–14.
- 22 Fernández-Lorente L, Riera L, Bestard O, *et al*. Long-Term Results of Biopsy-Guided Selection and Allocation of Kidneys From Older Donors in Older Recipients. *Am J Transplant* 2012; **12**: 2781–8.
- 23 Randhawa P. ROLE OF DONOR KIDNEY BIOPSIES IN RENAL TRANSPLANTATION. *Transplantation* 2001; **71**: 1361–5.
- 24 Karpinski J, Lajoie G, Cattran D, *et al*. OUTCOME OF KIDNEY TRANSPLANTATION FROM HIGH-RISK DONORS IS DETERMINED BY BOTH STRUCTURE AND FUNCTION. *Transplantation* 1999; **67**: 1162–7.
- 25 Pokorná E, Vítko S, Chadimová M, Schück O, Ekberg H. PROPORTION OF GLOMERULOSCLEROSIS IN PROCUREMENT WEDGE RENAL BIOPSY CANNOT ALONE DISCRIMINATE FOR ACCEPTANCE OF MARGINAL DONORS. *Transplantation* 2000; **69**: 36.

- 26 Escofet X, Osman H, Griffiths DFR, Woydag S, Adam Jurewicz W. The presence of glomerular sclerosis at time zero has a significant impact on function after cadaveric renal transplantation. *Transplantation* 2003; **75**: 344–6.
- 27 Lu AD, Desai D, Myers BD, Dafoe DC, Alfrey EJ. Severe glomerular sclerosis is not associated with poor outcome after kidney transplantation. *Am J Surg* 2000; **180**: 470–4.
- 28 Lopes JA, Moreso F, Riera L, *et al.* Evaluation of pre-implantation kidney biopsies: Comparison of Banff criteria to a morphometric approach. *Kidney Int* 2005; **67**: 1595–600.
- 29 Variation in use of procurement biopsies and its implications for discard of deceased donor kidneys recovered for transplantation - Lentine - 2019 - American Journal of Transplantation - Wiley Online Library.
<https://onlinelibrary.wiley.com/doi/full/10.1111/ajt.15325> (accessed Oct 4, 2021).
- 30 Furness PN, Taub N, Assmann KJM, *et al.* International variation in histologic grading is large, and persistent feedback does not improve reproducibility. *Am J Surg Pathol* 2003; **27**: 805–10.
- 31 Snoeijis MGJ, Boonstra LA, Buurman WA, *et al.* Histological assessment of pre-transplant kidney biopsies is reproducible and representative. *Histopathology* 2010; **56**: 198–202.
- 32 Wang HJ, Kjellstrand CM, Cockfield SM, Solez K. On the influence of sample size on the prognostic accuracy and reproducibility of renal transplant biopsy. *Nephrol Dial Transplant Off Publ Eur Dial Transpl Assoc - Eur Ren Assoc* 1998; **13**: 165–72.
- 33 Brown RJC. Measuring measurement – What is metrology and why does it matter? *Measurement* 2021; **168**: 108408.
- 34 Bièvre by PD. The 2012 International Vocabulary of Metrology: “VIM”. *Chem Int -- Newsmag IUPAC* 2012; **34**: 26–7.
- 35 Iso I, OIML B. Guide to the Expression of Uncertainty in Measurement. *Geneva Switz* 1995; **122**: 16–7.
- 36 Organization IS. ISO 5725-1: Accuracy (trueness and precision) of measurement methods and results-Parts 1, 2, 3, 4 and 6. 1994.
- 37 Haas M. Donor kidney biopsies: pathology matters, and so does the pathologist. *Kidney Int* 2014; **85**: 1016–9.
- 38 Azancot MA, Moreso F, Salcedo M, *et al.* The reproducibility and predictive value on outcome of renal biopsies from expanded criteria donors. *Kidney Int* 2014; **85**: 1161–8.

- 39 Girolami I, Gambaro G, Ghimenton C, *et al*. Pre-implantation kidney biopsy: value of the expertise in determining histological score and comparison with the whole organ on a series of discarded kidneys. *J Nephrol* 2020; **33**: 167–76.
- 40 Husain SA, Shah V, Alvarado Verduzco H, *et al*. Impact of Deceased Donor Kidney Procurement Biopsy Technique on Histologic Accuracy. *Kidney Int Rep* 2020; **5**: 1906–13.
- 41 Mazzucco G, Magnani C, Fortunato M, Todesco A, Monga G. The reliability of pre-transplant donor renal biopsies (PTDB) in predicting the kidney state. A comparative single-centre study on 154 untransplanted kidneys. *Nephrol Dial Transplant Off Publ Eur Dial Transpl Assoc - Eur Ren Assoc* 2010; **25**: 3401–8.
- 42 Singh P, Farber JL, Doria C, *et al*. Peritransplant kidney biopsies: comparison of pathologic interpretations and practice patterns of organ procurement organizations. *Clin Transplant* 2012; **26**: E191-199.
- 43 Muruve NA, Steinbecker KM, Luger AM. Are wedge biopsies of cadaveric kidneys obtained at procurement reliable? *Transplantation* 2000; **69**: 2384–8.
- 44 Yong ZZ, Aitken EL, Khan KH, Kingsmore DB. Wedge Versus Core Biopsy at Time Zero: Which Provides Better Predictive Value for Delayed Graft Function With the Remuzzi Histological Scoring System? *Transplant Proc* 2015; **47**: 1605–9.
- 45 Bago-Horvath Z, Kozakowski N, Soleiman A, Bodingbauer M, Mühlbacher F, Regele H. The cutting (w)edgecomparative evaluation of renal baseline biopsies obtained by two different methods. *Nephrol Dial Transplant* 2012; **27**: 3241–8.
- 46 Jacowitz KE, Kahneman D. Measures of anchoring in estimation tasks. *Pers Soc Psychol Bull* 1995; **21**: 1161–6.
- 47 Featherston R, Downie LE, Vogel AP, Galvin KL. Decision making biases in the allied health professions: A systematic scoping review. *PLOS ONE* 2020; **15**: e0240716.
- 48 Lawrence I, Lin K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989; : 255–68.
- 49 Cantor AB, Lee H. Sample-Size Calculations for Cohen’s Kappa. *Psychol Methods* 1996; : 150–3.
- 50 Kim TK. Understanding one-way ANOVA using conceptual figures. *Korean J Anesthesiol* 2017; **70**: 22–6.
- 51 Bland JM, Altman DG. Statistics Notes: Bootstrap resampling methods. *BMJ* 2015; **350**: h2622.
- 52 Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat Med* 2000; **19**: 1141–64.

- 53 Harris CR, Millman KJ, Van Der Walt SJ, *et al*. Array programming with NumPy. *Nature* 2020; **585**: 357–62.
- 54 Virtanen P, Gommers R, Oliphant TE, *et al*. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020; **17**: 261–72.
- 55 About us —scikit-learn 0.11-git documentation.
<https://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/about.html> (accessed May 24, 2023).
- 56 Ayorinde JOO, Citterio F, Landrò M, *et al*. Artificial Intelligence You Can Trust: What Matters Beyond Performance When Applying Artificial Intelligence to Renal Histopathology? *J Am Soc Nephrol* 2022; **33**: 2133.
- 57 Muruve NA, Steinbecker KM, Luger AM. ARE WEDGE BIOPSIES OF CADAVERIC KIDNEYS OBTAINED AT PROCUREMENT RELIABLE?1. *Transplantation* 2000; **69**: 2384–8.
- 58 Haas M, Segev DL, Racusen LC, *et al*. Arteriosclerosis in kidneys from healthy live donors: comparison of wedge and needle core perioperative biopsies. *Arch Pathol Lab Med* 2008; **132**: 37–42.
- 59 Mohan S, Campenot E, Chiles MC, *et al*. Association between Reperfusion Renal Allograft Biopsy Findings and Transplant Outcomes. *J Am Soc Nephrol* 2017; **28**: 3109–17.
- 60 Sagasta A, Sánchez-Escuredo A, Oppenheimer F, *et al*. Pre-implantation analysis of kidney biopsies from expanded criteria donors: testing the accuracy of frozen section technique and the adequacy of their assessment by on-call pathologists. *Transpl Int* 2016; **29**: 234–40.
- 61 Denic A, Ricaurte L, Lopez CL, *et al*. Glomerular Volume and Glomerulosclerosis at Different Depths within the Human Kidney. *J Am Soc Nephrol JASN* 2019; **30**: 1471–80.
- 62 Yushkov Y, Selck FW. An Approach to Needle Biopsy Technique to Improve Glomerulus Yield. *Transplant Proc* 2008; **40**: 1051–3.
- 63 Naesens M. Zero-Time Renal Transplant Biopsies: A Comprehensive Review. *Transplantation* 2016; **100**: 1425–39.
- 64 Husain SA, Chiles MC, Lee S, *et al*. Characteristics and Performance of Unilateral Kidney Transplants from Deceased Donors. *Clin J Am Soc Nephrol* 2018; **13**: 118–27.
- 65 Yi Z, Xi C, Menon MC, *et al*. A large-scale retrospective study enabled deep-learning based pathological assessment of frozen procurement kidney biopsies to predict graft loss and guide organ utilization. *Kidney Int* 2024; **105**: 281–92.

- 66 Marsh JN, Liu T-C, Wilson PC, Swamidass SJ, Gaut JP. Development and Validation of a Deep Learning Model to Quantify Glomerulosclerosis in Kidney Biopsy Specimens. *JAMA Netw Open* 2021; **4**: e2030939.
- 67 Comparison of automated percutaneous kidney biopsy using needles of different gauge - LYNN - 1997 - Nephrology - Wiley Online Library. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1440-1797.1997.tb00217.x?casa_token=ugOyGiaOyHsAAAAA:BqfRyeHMBemPbAm7jWWwV3aJ4ESxYPcTntDmDKKoVKpwItaYdZDWHZleJ6goXvfEcQQGIr5MEsLauw (accessed May 28, 2024).
- 68 Nicholson ML, Wheatley TJ, Doughman TM, *et al*. A prospective randomized trial of three different sizes of core-cutting needle for renal transplant biopsy. *Kidney Int* 2000; **58**: 390–5.
- 69 Peters B, Mölne J, Hadimeri H, Hadimeri U, Stegmayr B. Sixteen Gauge biopsy needles are better and safer than 18 Gauge in native and transplant kidney biopsies. *Acta Radiol* 2017; **58**: 240–8.
- 70 Sousanieh G, Whittier WL, Rodby RA, Peev V, Korbet SM. Percutaneous Renal Biopsy Using an 18-Gauge Automated Needle Is Not Optimal. *Am J Nephrol* 2021; **51**: 982–7.
- 71 Marsh JN, Matlock MK, Kudose S, *et al*. Deep Learning Global Glomerulosclerosis in Transplant Kidney Frozen Sections. *IEEE Trans Med Imaging* 2018; : 1–1.
- 72 Ginley B, Jen K-Y, Rosenberg A, *et al*. Neural Network Segmentation of Interstitial Fibrosis, Tubular Atrophy, and Glomerulosclerosis in Renal Biopsies. ; : 31.
- 73 Bukowy JD, Dayton A, Cloutier D, *et al*. Region-Based Convolutional Neural Nets for Localization of Glomeruli in Trichrome-Stained Whole Kidney Sections. *J Am Soc Nephrol* 2018; **29**: 2081–8.
- 74 Marechal E, Jaugey A, Tarris G, *et al*. Automatic Evaluation of Histological Prognostic Factors Using Two Consecutive Convolutional Neural Networks on Kidney Samples. *Clin J Am Soc Nephrol CJASN* 2022; **17**: 260–70.
- 75 Gadermayr M, Dombrowski A-K, Klinkhammer BM, Boor P, Merhof D. CNN cascades for segmenting sparse objects in gigapixel whole slide images. *Comput Med Imaging Graph* 2019; **71**: 40–8.
- 76 Cui M, Zhang DY. Artificial intelligence and computational pathology. *Lab Invest* 2021; **101**: 412–22.
- 77 Jiang L, Chen W, Dong B, *et al*. A Deep Learning-Based Approach for Glomeruli Instance Segmentation from Multistained Renal Biopsy Pathologic Images. *Am J Pathol* 2021; **191**: 1431–41.

- 78 Sharma K, Rupperecht C, Caroli A, Aparicio MC, Baust M, Navab N. Automatic Segmentation of Kidneys using Deep Learning for Total Kidney Volume Quantification in Autosomal Dominant Polycystic Kidney Disease. *Sci Rep* 2017; : 1–10.

12 Legends

Figures

1. Figure 1: Sample processing and analysis in unblinded and semi-Blinded conditions. This flowchart depicts the progression of samples from collection to analysis. For both methods, 12 patients provided 16 kidneys, from which 144 biopsies were produced and two HE and PAS glass slides generated. In the unblinded scenario, slides were examined jointly and a single Remuzzi score reported for each biopsy. In the semi-blinded scenario slides were assessed individually.
2. Figure 2: Scatter plots analysing core (green cross), punch (orange circle) and wedge (purple star) biopsies. 2a) Physical profile (n=250), core biopsies are long and thin with wedge biopsies having the reverse profile. Punch biopsies are the most consistent. 2b) & 2c) Glomerular yield (n=250, $r=0.94$) & Arterial yield (n=249, $r=0.62$) retrieved and the size of each sample. 2d) Relationship between the number of glomeruli in the sample and the percentage of glomerulosclerosis. The background colour reflects the Remuzzi score (0=blue, 1=green, 2=orange, 3=red). Glomerulosclerosis is weakly correlated with size of the sample (Spearman's rank coefficient $r=0.13$, $p=0.03$).
3. Figure 3: Distribution of Remuzzi scores for each kidney, with each column representing a single kidney (e.g., PL1 = Donor 1, Left kidney). Shading illustrates the decision to transplant singly (1-4, white), as a dual (>4-6, grey) or discard (≥ 7 , red). Markers indicate the type of biopsy performed: green cross for core, orange circle for punch, and purple star for wedge biopsy.

Noticeable variation is present within kidney groups, and pairs of kidneys (e.g., PL2 & PR2) exhibit similar score distributions.

4. Figure 4: Violin plot representing the bootstrapped ($n=1000$) distribution of Remuzzi score sample means for each kidney. The plot presents the spread and density estimation of the simulated means. Shading illustrates the decision to transplant singly (1-4, white), as a dual ($>4-6$, grey) or discard (≥ 7 , red). Bootstrapped means distributions show a tighter range as well as high similarity distribution within pairs, whilst distributions from other donors are generally less alike.
5. Figure 5: Distribution of Unblinded (Glass slide) Remuzzi scores for each kidney, with each column representing a single kidney (e.g., PL1 = Patient 1, Left kidney). Shading describes the decision to transplant singly (1-4, white), as a dual ($>4-6$, grey) or discard (≥ 7 , red).
6. Figure 6: ($n=265$) Box plots which show the measured degree of injury on the y axis, for each slide scored by the pathologist. Glomerulosclerosis =upper, IFTA=lower). Perfect agreement would show increasing, non-overlapping groups which are separable at the Remuzzi threshold for each component (grey boundaries 0-3).
7. Figure S1: Scatter plot of dimensions of the biopsies returned by each surgeon (colour). (upper) wedge biopsies, very small (inadequate) & very large (dangerous) biopsies possible. (lower) punch biopsy variability largely confined to depth of the sample, returning more consistent biopsy size.
8. Figure S2: Scatter plots analysing core (green cross), punch (orange circle) and wedge (purple star) biopsies in the clinical cohort. 7a) Physical profile ($n=1044$),

similar physical profile allowing for inconsistencies in labelling dimensions. 7b) & 7c) Glomerular yield & Arterial yield retrieved and the size of each sample, few reports contained specific count for arteries. 7d) Glomerulosclerosis pattern is similar as for the experimental cohort. Background shading demonstrates each Remuzzi score assigned.

9. Figure S3: Violin plot representing the raw bootstrapped (n=1000) Remuzzi scores sample means for each kidney. The plot presents the spread and density estimation of the simulated means. Shaded areas denote the transplant strategy (white=single, grey=dual, red=discard). Bootstrapped means distributions show a tighter range as well as high similarity distribution within pairs, whilst distributions from other donors are generally less alike.
10. Figure S4: Violin plot representing the raw bootstrapped (n=1000) Remuzzi scores sample means for core biopsies of each kidney. The plot presents the spread and density estimation of the simulated means. Shaded areas denote the transplant strategy (white=single, grey=dual, red=discard). Bootstrapped means distributions show a tighter range as well as high similarity distribution within pairs, whilst distributions from other donors are generally less alike
11. Figure S5: Violin plot representing the raw bootstrapped (n=1000) Remuzzi scores sample means for punch biopsies of each kidney. The plot presents the spread and density estimation of the simulated means. Shaded areas denote the transplant strategy (white=single, grey=dual, red=discard). Bootstrapped means distributions show a tighter range as well as high similarity distribution within pairs, whilst distributions from other donors are generally less alike

12. Figure S6: Violin plot representing the raw bootstrapped ($n=1000$) Remuzzi scores sample means for wedge biopsies of each kidney. The plot presents the spread and density estimation of the simulated means. Shaded areas denote the transplant strategy (white=single, grey=dual, red=discard). Bootstrapped means distributions show a tighter range as well as high similarity distribution within pairs, whilst distributions from other donors are generally less alike

13 Tables

Table 1. Demographics table for donor (n=12) and slide (n=250) characteristics in this study

	Proportion (%)	Average (range)
Kidney image library		
DBD	16.7	--
DCD	83.3	--
Cause of death		
Hypoxic brain injury	53	--
Intracranial event	47	--
Sex		
Female	54	--
Comorbidity		
Age (years)	--	65.4 (52-76)
Diabetes	8	--
Hypertension	68	--
Smoking	49	--
BMI	--	28.7 (16.5-30.5)
Terminal creatinine (umol/l)	--	73.2 (46-138)

Table 2. Characteristics of surgical biopsies included in image library (n= 250)

	Core	Punch	Wedge	ANOVA
	(mean (sd))	(mean (sd))	(mean (sd))	p-value
Semi-blind measurements				
Width (mm)	0.8 (0.2)	2.6 (0.4)	10.7 (4.03)	<0.001*
Depth (mm)	13.0 (4.3)	5.4 (1.2)	4.5 (2.1)	<0.001*
Cortex area	9.04 (5.1)	13.7 (3.9)	37.9 (24.8)	<0.001*
Vessel count ¹	2.86 (2.5)	3.5 (2.7)	7.5 (7.4)	<0.001*
Glomeruli	20.2 (12.6)	32.0 (11.2)	86.4 (62.9)	<0.001*
Glomerulosclerosis (%)	9.4 (11.8)	11.4 (13.2)	9.0 (9.2)	0.34
IFTA (%)	18.1 (22.3)	21.9 (22.3)	23.2 (20.8)	0.13
Adequacy rate ²	26%	76%	76%	<0.001*

¹ Vessels include arteries and arterioles, ²Adequate samples/Total samples (%)

* "Following significant ANOVA results, pairwise comparisons showed significant differences in depth, width, total cortical area, and glomerular count between biopsy types. Wedge biopsies were larger, with increased depth, width, and glomerular count. However, no significant difference was found in the mean total cortical area between core and punch biopsies.

Table 3. Weighted statistics for agreement of Remuzzi scores following each type of biopsy

	Core (weighted kappa)	Punch (weighted kappa)	Wedge (weighted kappa)
Core	--	0.75	0.62
Punch	0.75	--	0.78

Table 4. This table presents the result of two bootstrapping approaches applied to the dataset. The tan rows (left), represent the proportions of individual bootstrapped samples that are distributed across the three decision outcome categories. The grey rows (right) display key statistics derived from the bootstrapped means of Remuzzi scores for each kidney (mean, standard deviation), and the percentage of these means that fall within the three decision outcome categories.

	Single tx (%)	Dual tx (%)	Discard tx (%)	Pair Match ¹ (%)	Mean Remuzzi score (sd)	Single tx (%)	Dual tx (%)	Discard tx (%)	Pair match ¹ (%)
Kidney									
PL1	72	23	1	-	3.3 (0.3)	98	2	0	-
PR1	100	0	0	72	1.7 (0.3)	100	0	0	98
PL2	12	20	68	-	7.3 (0.6)	0	1	99	-
PR2	7	6	87	81	8.0 (0.4)	0	0	100	99
PL3	94	6	0	-	1.9 (0.3)	100	0	0	-
PR4	62	19	19	-	3.4 (0.6)	85	15	0	-
PR5	34	35	31	-	5.6 (0.8)	3	69	28	-
PR6	0	35	65	-	6.7 (0.4)	0	0	100	-
PR7	100	0	0	-	0.6 (0.4)	100	0	0	-
PR8	48	39	13	-	3.6 (0.8)	73	27	0	-
PR9	71	17	12	-	2.7 (0.4)	100	0	0	-
PL10	31	42	27	-	4.7 (0.4)	5	95	0	-
PR 10	14	62	24	80	5.2 (0.2)	0	100	0	95
PL11	0	21	79	-	7.3 (0.2)	0	0	100	-
PR11	10	43	47	68	6.0 (0.3)	0	49	51	51
PR12	53	40	7	-	3.8 (0.5)	65	35	0	-