Received: 15 May 2018 | Revised: 20 October 2018 | Accepted: 25 October 2018

DOI: 10.1111/ctr.13441

ORIGINAL ARTICLE

WILEY Clinical TRANSPLANTATION
The Journal of Clinical and Translational Research

Reliability of deceased-donor procurement kidney biopsy images uploaded in United Network for Organ Sharing

Sherry G. Mansour[1,2] | Isaac E. Hall[3] | Peter P. Reese[4] | Yaqi Jia[1,5] | Heather Thiessen-Philbrook[1,5] | Gilbert Moeckel[6] | Francis L. Weng[7] | Monica P. Revelo[8] | Mazdak A. Khalighi[8] | Anshu Trivedi[9] | Mona D. Doshi[10] | Bernd Schröppel[11] | Chirag R. Parikh[5]

[1]Program of Applied Translational Research, Department of Medicine, Yale University School of Medicine, New Haven, Connecticut

[2]Division of Nephrology, Yale University School of Medicine, New Haven, Connecticut

[3]Division of Nephrology, Hypertension and Renal Transplantation, Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, Utah

[4]Renal-Electrolyte and Hypertension Division, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania

[5]Division of Nephrology, School of Medicine, Johns Hopkins University, Baltimore, Maryland

[6]Division of Pathology, Yale University School of Medicine, New Haven, Connecticut

[7]Saint Barnabas Medical Center, Livingston, New Jersey

[8]Department of Pathology and Laboratory Medicine, University of Utah School of Medicine, Salt Lake City, Utah

[9]Division of Pathology, Hartford Hospital, Hartford, Connecticut

[10]University of Michigan, Ann Arbor, Michigan

[11]Section of Nephrology, University Hospital, Ulm, Germany

Correspondence
Chirag R. Parikh, MD, PhD, Director, Division of Nephrology, Ronald Peterson Professor of Medicine, Baltimore, MD.
Email: Chirag.Parikh@jhmi.edu

Funding information
Our work was sustained and supported by 1) the National Institutes of Health, grant R01DK-93770, grant K24DK090203, P30 DK 079310-07, 3) a career development award from the American Heart Association to Dr. Hall, and 4) the Health Resources and Services Administration contract 234-2005-37011C. Funded by the American Heart Association Career Development, grant 18CDA34110151 (Dr. Sherry Mansour).

Abstract

Prior studies demonstrate poor agreement among pathologists' interpretation of kidney biopsy slides. Reliability of representative images of these slides uploaded to the United Network of Organ Sharing (UNOS) web portal for clinician review has not been studied. We hypothesized high agreement among pathologists' image interpretation, since static images eliminate variation induced by viewing different areas of movable slides. To test our hypothesis, we compared the assessments of UNOS-uploaded images recorded in standardized forms by three pathologists. We selected 100 image sets, each having at least two images from kidneys of deceased donors. Weighted Cohen's kappa was used for inter-rater agreement. Mean (SD) donor age was 50 (13). Acute tubular injury had kappas of 0.12, 0.14, and 0.19; arteriolar hyalinosis 0.16, 0.27, and 0.38; interstitial inflammation 0.30, 0.33, and 0.49; interstitial fibrosis 0.28, 0.32, and 0.67; arterial intimal fibrosis 0.34, 0.42, and 0.59; tubular atrophy 0.35, 0.41, and 0.52; glomeruli thrombi 0.32, 0.53, and 0.85; and global glomerulosclerosis 0.68, 0.70, and 0.77. Pathologists' agreement demonstrated kappas of 0.12 to 0.77. The lower values raise concern about the reliability of using images. Although further research is needed to understand how uploaded images are used clinically, the field may consider higher-quality standards for biopsy photomicrographs.

KEYWORDS
agreement, biopsy, deceased donor, images, kappa, pathologists, renal, transplant

1 | INTRODUCTION

For the 100 000 patients on the kidney transplant waiting list, only 20% will actually receive transplants and 5% (about 22 patients every day) will die before they can receive a transplant this year.[1,2] Despite the growing disparity between the number of kidney transplants needed versus performed, kidney discard rates nearly quadrupled between 1988 and 2009, from about 5% to 20% of procured kidneys.[3] Researchers have extensively investigated reasons for discard in hopes of closing the gap and salvaging all viable organs.[3,4] Based on national registry data from the Organ Procurement and Transplantation Network (OPTN), the most frequently documented reason for kidney discard remains "biopsy findings."[5]

There is an association between procurement kidney biopsy findings and organ discard rates, but when biopsy findings are reassuring in kidneys from marginal donors, biopsies may also be associated with organ acceptance. Hence, it is important to understand the reliability of the reporting of histological findings from these biopsies generated in the organ procurement setting.[6] Azancot et al demonstrated considerable variability in pathologists' reports, with minimal agreement between less experienced pathologists and only moderate agreement among more experienced and expert pathologists.[7] Furthermore, the authors found no significant associations between donor histological findings and recipient graft function when biopsies were assessed by less experienced pathologists, but histology was significantly and independently associated with recipient graft function when reported by an experienced and expert renal pathologist.[7] Liapis et al found good reproducibility in only four out of 12 histological findings when assessed by 32 expert renal pathologists.[8]

Despite the variability in biopsy slide interpretations, the reading pathologists may decide which sections to highlight by taking representative images of the biopsy slide. These images are available for review in the United Network of Organ Sharing (UNOS) web-accessible database. Organ procurement organizations (OPOs) upload representative photomicrographs of pathology slides for online review during organ offers, as a way for transplant centers to assess the histology for themselves or to verify elements of on-call reports. The process of creating high-resolution digital images of histological material is gaining wide use in the field of pathology, with whole slide imaging currently being used both clinically and in research.[9] In contrast to whole slide imaging, a single image may or may not be representative of the entire slide, and its interpretability may vary depending on the viewer.

The value of donor biopsy images uploaded to UNOS may also depend on the quality of pathology review. One important dimension of quality is reliability, that is, the similarity in interpretation of pathology findings between independent pathologists. It is unknown whether images uploaded to UNOS would be consistently interpreted even under optimal circumstances with experienced renal pathologists using standardized reporting methods. We hypothesized that agreement among pathologists when interpreting static images would be high. The rationale for this hypothesis was that static image interpretation is likely more reproducible as it eliminates variation induced by viewing different areas of a freely movable slide. Therefore, we aimed to evaluate the agreement across standardized histological findings in UNOS-uploaded deceased-donor procurement biopsy images between three experienced renal transplant pathologists.

2 | METHODS

2.1 | Study design

One hundred kidney procurement biopsy image sets from 85 distinct deceased donors were obtained from UNOS and included in the study to evaluate inter-rater agreement. Ninety-one percent of image sets used in this study consisted of photomicrographs of frozen wedge biopsies, while 9% were needle biopsies. Donors were selected from the preexisting prospective multicenter Deceased-Donor Cohort Study (DDS), which has been described in detail elsewhere.[10] For inclusion in the current analysis, kidney biopsies of donors had to have at least two images uploaded in the web-accessible UNOS system known as DonorNet. Out of 425 UNOS image sets available for this study, we selected all those with moderate and severe findings for glomerulosclerosis, interstitial fibrosis, and acute tubular injury as described on the UNOS and OPO biopsy reports. For the remaining image sets, we utilized random disproportionate stratified sampling, which involved dividing the image sets into two smaller strata of image sets with acute tubular injury as reported by UNOS and image sets without acute tubular injury. We then disproportionately sampled image sets from each stratum to ensure a reasonable distribution of pathology among the 100 image sets. De-identified image sets were securely distributed to three experienced academic renal pathologists with 17, 4, and 18 years of experience for pathologists 1, 2, and 3, respectively, since completion of renal pathology fellowship training at different academic institutions. The pathologists were blinded to the OPO and UNOS biopsy reports and to each other's findings. Pathologists 1 and 2 were from the same institution, and the third pathologist was from a separate institution. Representative images are shown in Figure 1. Each pathologist was asked to complete a standardized scoring sheet adapted from Liapis et al with the following eight histological characteristics: percent glomerulosclerosis, glomeruli thrombi, interstitial fibrosis, tubular atrophy, interstitial inflammation, arterial intimal fibrosis, arteriolar hyalinosis, and acute tubular injury as shown in Figure S1.[8] Each histological characteristic was given an ordinal definition as none, mild, moderate, or severe along with a corresponding percentage. Pathologists were instructed to follow the provided ordinal definitions on the scoring sheet. Each donor image set was evaluated using one scoring sheet. Thus, each scoring sheet was representative of a unique donor. In addition, each pathologist evaluated a set of 10 random image sets more than once to evaluate intra-rater agreement. These samples were selected via simple

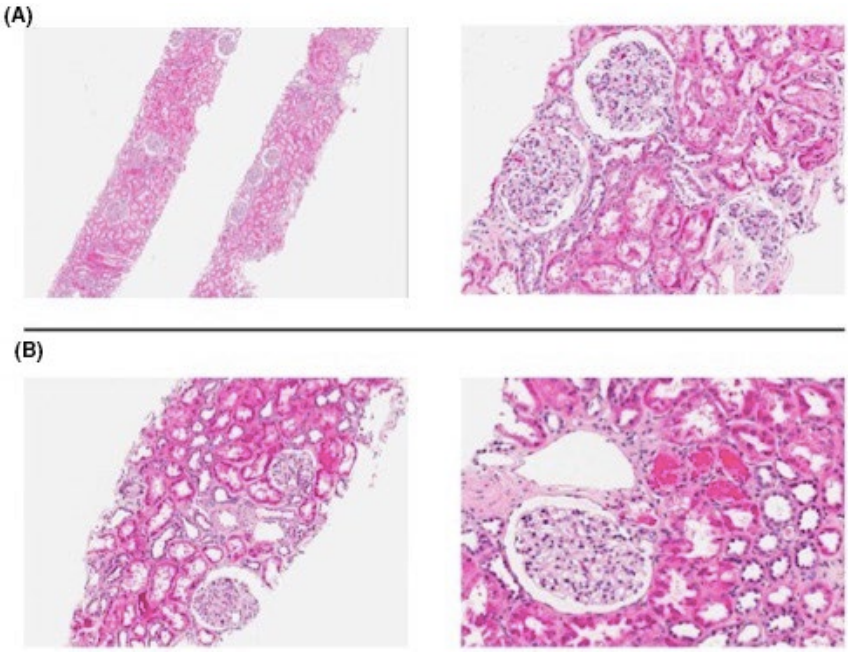

FIGURE 1 This figure represents two sets of images from different donors that were given to pathologists for intra- and inter-rater agreement analysis. Figure A shows an image of a core needle biopsy, while figure B shows an image of a frozen wedge biopsy

random sampling, where each image set had an equal probability of being chosen.

This study used data from the OPTN. The OPTN data system includes data on all donors, wait-listed candidates, and transplant recipients in the US, submitted by the members of the OPTN. The Health Resources and Services Administration (HRSA), US Department of Health and Human Services provides oversight to the activities of the OPTN contractor. The analyses are based on OPTN data and biopsy images as of May 6, 2016. The institutional review boards of all participating centers approved this study (Human Investigation Committee Protocol Number: 1206010465). All clinical investigators abided by the ethical principles for medical research involving human subjects as outlined in the Declaration of Helsinki.

2.2 | Statistical analysis

Baseline characteristics are presented as mean ± standard deviation if continuous and as frequencies (%) if categorical. Kappa was used to evaluate rater agreement not due to chance. More specifically, we used weighted Cohen's kappa statistic to assess intra-rater agreement as well as pairwise inter-rater agreement between two pathologists at a time, and we used Fleiss kappa to assess inter-rater agreement among all three pathologists.[11] Given the ordinal nature of the data, a weighted kappa was used to account for the degree of disagreement.[12] As such, two-level disagreements were weighted as a higher degree of disagreement compared to one-level disagreements. For example, a difference between pathologists of no fibrosis versus moderate fibrosis was weighted as more important than a difference of no fibrosis vs mild fibrosis. For each histological finding, we reported overall Fleiss kappa, weighted pairwise Cohen's kappa, prevalence-adjusted bias-adjusted kappa (PABAK),[13] which assumes that the bias of prevalence is absent and that prevalence is fixed at 50%, and pairwise percent agreement. The interpretations of the Kappa coefficients are as follows: none (0-0.20), minimal (0.21-0.39), mild (0.40-0.59), moderate (0.60-0.79), strong (0.80-0.90), and almost perfect (>0.90).[14] An acceptable kappa statistic is usually ≥0.60, which corresponds with moderate or higher agreement.[14]

For all histological characteristics, we combined moderate and severe terms to generate three categories of none, mild, and greater than mild (ie, moderate or severe) because <5% of scores were moderate and <5% of scores were severe. Indeterminate and missing data (<5% of scores) were excluded from the analysis. For glomerulosclerosis, 1%-20% and >20% were the two categories used to calculate kappa among the three pathologists, as the cutoff of >20% has been shown to be associated with discard. We also calculated the overall kappa among the three pathologists when histological findings were scored as only two categories of none/mild and moderate/severe. Inference testing for the kappa statistic was done using the *Z*-test. *P*-values < 0.05 were considered statistically significant. With 100 image sets, the study had sufficient statistical power of at least 85% to detect mild agreement between two pathologists (kappa of at least 0.55 with a null hypothesis of 0.30).[15] For each histological finding, we reported overall Fleiss kappa and pairwise percent agreement.

To evaluate whether the presence of UNOS images affected clinical decision-making as compared to biopsies without images, we assessed the distribution of discarded kidneys, and cold ischemia time (hours) between kidneys with biopsies without UNOS images (n = 1326) and kidneys with biopsies plus UNOS images (n = 425). Inference testing was done using the Z-test for the dichotomous outcome of discarded kidneys. For the continuous outcome of cold ischemia time (hours), we used Wilcoxon signed-rank test.

3 | RESULTS

Baseline characteristics of all donors are shown in Table 1. The mean age was 50 ± 13%, and 57% of donors were male. Forty percent of kidneys were discarded. The average number of images per donor kidney image set was two. Thirty-five (41%) of the 85 donors had images representative of the left kidney, 35 (41%) had images representative of the right kidney, and 15 (18%) donors had images representative of both left and right kidneys, which yielded a total of 100 image sets of either left or right kidneys from 85 distinct donors. Sixty (60%) kidneys were transplanted and 31 (52%) of the transplanted kidneys developed delayed graft function (DGF, defined as any dialysis within the first week of transplant), with an average 6-month estimated glomerular filtration rate (eGFR) of about 47 mL/min/1.73m^2, 6-month kidney graft failure of 3%, and 6-month mortality of 5% as shown in Table 1. The median image file size was 427 kB, with a range of 70-3654 kB. Out of all 221 images, 3% had 100× magnification, 36% had 200× magnification, and 62% of images had 400× magnification. Out of the 100 image sets, 2% had all images with 100× magnification, 20% had all images with 200x magnification, 43% had all images with 400× magnification, and 35% had images with different magnifications within each set. When shown duplicated image sets in blinded fashion, pathologists demonstrated a high level of agreement with their prior interpretations of the biopsy as shown in Table S1. Pathologist 1, 2, and 3 had median (range) weighted Cohen's kappa of 0.63 (0.32, 1.00), 1.00 (0.74, 1.00), and 0.69 (0.31, 1.00), respectively. Percent agreement and weighted Cohen's kappa between pairs of pathologists for each histological finding per image set are shown in Figure 2. Agreement across the eight histological characteristics is described below:

TABLE 1 Baseline characteristics of deceased donors

Variables	All kidneys (n = 100)[a] Mean (SD) or n (%)
Age (years)	50 (13)
Males	57 (57%)
Black race	18 (18%)
Height (cm)	169 (12)
Weight (kg)	87 (26)
History of hypertension	54 (54%)
History of diabetes	10 (10%)
Donor cause of death	
Anoxia	29 (29%)
Stroke	52 (52%)
Head trauma	19 (19%)
Negative hepatitis C antibody	100 (100%)
Admission serum creatinine (mg/dL)	1.17 (0.54)
Terminal serum creatinine (mg/dL)	1.48 (1.19)
Expanded criteria donor	41 (41%)
Donation after cardiac death	10 (10%)
Kidney donor risk index	1.28 (0.35)
Kidney donor profile index (%) relative to 2010 median donor	67 (23)
Number of kidneys discarded	40 (40%)
Transplanted kidneys (n = 60)	
Delayed graft function	31 (52%)
6-month eGFR (mL/min/m^2)	47 (20)
6-month graft failure	2 (3%)
6-month mortality	3 (5%)

eGFR, estimated glomerular filtration rate.
[a]The 100 kidney image sets came from 85 distinct donors.

3.1 | Acute tubular injury

When initially assessed using whole slides, 40% had moderate acute tubular injury based on OPO biopsy reports as shown in Table 2. Review of image sets by study pathologists revealed a range of scores from 5% to 40% as having moderate acute tubular injury depending on the interpreting pathologist.

Overall, the pathologists had no agreement in regard to acute tubular injury (weighted Cohen's kappa was 0.12, 0.14, and 0.19; PABAK was −0.07, 0.30, and 0.31 when comparing pathologists 1&3, 2&3, and 1&2, respectively; and Fleiss kappa [95% CI] was 0.07, [−0.01,0.16]), but this was not statistically significant as shown in Figure 2 and Table 3.

3.2 | Glomerulosclerosis

Global glomerulosclerosis >20% was noted in 20% of OPO biopsy reports of whole kidney slides as shown in Table 2. Glomerulosclerosis >20% was found in 14% to 25% of the image sets reviewed by pathologists. Biopsy image sets had an average of 5 ± 3 glomeruli per image set among all three pathologists. Overall, the pathologists had mild to moderate agreement with regard to glomerulosclerosis (weighted Cohen's kappa was 0.68, 0.70, and 0.77; PABAK was 0.62, 0.76, and 0.82 when comparing pathologists 1&3, 2&3, and 1&2, respectively; and Fleiss kappa [95% CI] was 0.57 [0.45, 0.68]) as shown in Figure 2 and Table 3.

3.3 | Interstitial fibrosis

Organ procurement organization whole slide biopsy reports noted interstitial fibrosis >25% in 11% of reports as shown in Table 2. image sets reviewed by pathologists revealed interstitial fibrosis >25% in 6% to 13% of image sets, with minimal overall agreement (weighted Cohen's kappa was 0.28, 0.32, and 0.67; PABAK was 0.24, 0.25, and 0.73 when comparing pathologists 1&3, 2&3, and 1&2, respectively; and Fleiss kappa [95% CI] was 0.29 [0.20, 0.38]) as shown in Figure 2 and Table 3.

3.4 | Interstitial inflammation

Interstitial inflammation > 25% was assessed by the pathologists in 2% to 15% of the image sets as shown in Table 2, with minimal agreement overall (weighted Cohen's kappa was 0.30, 0.33, and 0.49; PABAK was 0.41, 0.47, and 0.72 when comparing pathologists

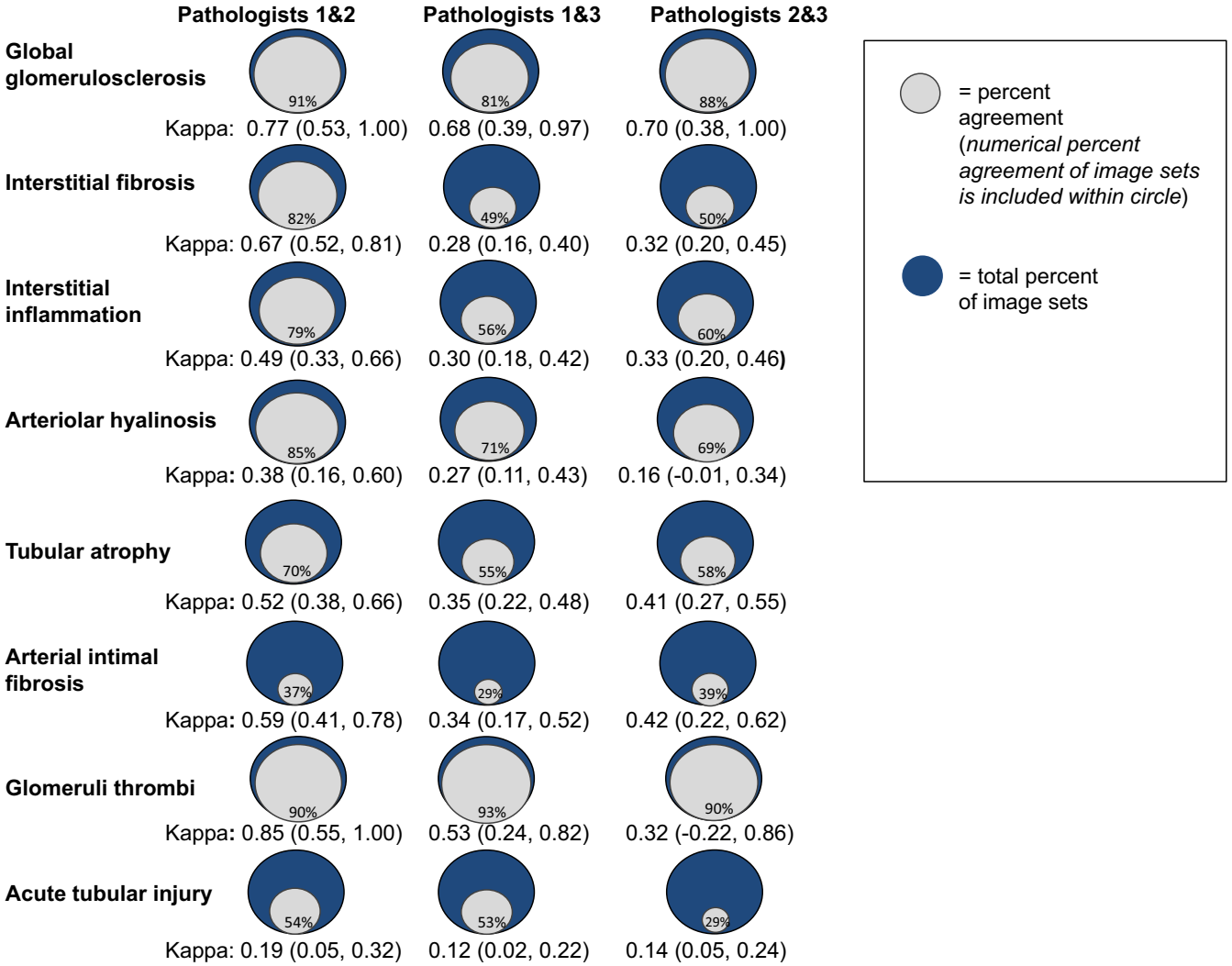

FIGURE 2 This figure shows percent agreement between Pathologists 1&2, 1&3, and 2&3 for each histological finding as well as the weighted Cohen's kappa and 95% confidence intervals. The dark blue circles represent 100% of the images (total of 100 images, except for arterial intimal fibrosis since only images with arteries were included.) The numbers within the light grey circles represent percent agreement between each pair of pathologists %

1&3, 2&3, and 1&2, respectively; and Fleiss kappa [95% CI] was 0.27 [0.18, 0.36]) as shown in Figure 2 and Table 3.

3.5 | Arteriolar hyalinosis

The majority of image sets were found to have no arteriolar hyalinosis based on the pathologists' scoring as shown in Table 2, with minimal overall agreement (weighted Cohen's kappa was 0.16, 0.27, and 0.38; PABAK was 0.54, 0.57, and 0.78 when comparing pathologists 2&3, 1&3, and 1&2, respectively; and Fleiss kappa [95% CI] was 0.21 [0.11, 0.31]) as shown in Figure 2 and Table 3.

3.6 | Tubular atrophy

Tubular atrophy > 25% was noted by the pathologists in 6% to 12% of the image sets as shown in Table 2, with minimal to mild overall agreement (weighted Cohen's kappa was 0.35, 0.41, and 0.52; PABAK was 0.40, 0.44, and 0.60 when comparing pathologists 1&3, 2&3, and 1&2, respectively; and Fleiss kappa [95% CI] was 0.33 [0.24, 0.42]) as shown in Figure 2 and Table 3.

3.7 | Arterial intimal fibrosis

Arterial intimal fibrosis > 25% was noted by the pathologists in 3% to 8% of the image sets as shown in Table 2, with minimal to mild overall agreement (weighted Cohen's kappa was 0.34, 0.42, and 0.59; PABAK was 0.05, 0.16, and 0.19 when comparing pathologists 1&3, 2&3, and 1&2, respectively; and Fleiss kappa [95% CI] was 0.33 [0.30, 0.36]) as shown in Figure 2 and Table 3.

3.8 | Glomeruli thrombi

The pathologists scored 90% to 95% of image sets as "none" with regard to thrombi within glomeruli as shown in Table 2, and overall

TABLE 2 Distribution of histological findings per pathologist with corresponding readings from OPO biopsy reports

Histological findings	Ordinal scale	Pathologist 1, n (%)	Pathologist 2, n (%)	Pathologist 3, n (%)	OPO[a], n (%)
Acute tubular injury	None	9 (9%)	40 (40%)	6 (6%)	52 (52%)
	Mild (epithelial flattening, tubule dilation, nuclear dropout, loss of brush border)	86 (86%)	48 (48%)	54 (54%)	8 (8%)
	Moderate (focal necrosis) and severe (infarction)	5 (5%)	12 (12%)	40 (40%)	40 (40%)
Glomerulosclerosis	0%-20%	75 (75%)	82 (82%)	86 (86%)	80 (80%)
	>20%	25 (2 5%)	18 (18%)	14 (14%)	20 (20%)
Interstitial fibrosis	None (<5% of cortex)	71 (71%)	67 (67%)	29 (29%)	12 (12%)
	Mild (6%-25%)	23 (23%)	24 (24%)	58 (58%)	77 (77%)
	Moderate (26%-50%) or Severe (>50%)	6 (6%)	9 (9%)	13 (13%)	11 (11%)
Interstitial inflammation	None (<10% of cortex)	75 (75%)	79 (79%)	47 (47%)	NA
	Mild (10%-25%)	22 (22%)	17 (17%)	38 (38%)	
	Moderate (26%-50%) or Severe (>50%)	3 (3%)	2 (2%)	15 (15%)	
	Indeterminate		2 (2%)		
Arteriolar hyalinosis	None	86 (86%)	87 (87%)	70 (70%)	NA
	Mild (at least one arteriole)	10 (10%)	13 (13%)	27 (27%)	
	Moderate (more than one arteriole) or Severe (multiple arterioles affected)	4 (4%)		3 (3%)	
Tubular atrophy	None (0% of cortical tubules)	61 (61%)	55 (55%)	30 (30%)	NA
	Mild (<25%)	32 (32%)	36 (36%)	58 (58%)	
	Moderate (26%-50%) or Severe (>50%)	6 (6%)	9 (9%)	12 (12%)	
	Indeterminate	1 (1%)			
Arterial intimal fibrosis[b]	None (0% vascular narrowing)	33 (61%)	24 (38%)	35 (41%)	NA
	Mild (<25%)	17 (31%)	33 (52%)	43 (50%)	
	Moderate (26%-50%) or Severe (>50%)	3 (6%)	3 (5%)	8 (9%)	
	Indeterminate	1 (2%)	3 (5%)		
Glomeruli thrombi	None	93 (93%)	90 (91%)	95 (95%)	NA
	Mild (<10% of capillaries occluded)	3 (3%)	1 (1%)	3 (3%)	
	Moderate (10%-25% occlusion) or Severe (>25% occlusion)	1 (1%)	1 (1%)	2 (2%)	
	Indeterminate	3 (3%)	7 (7%)		

OPO, organ procurement organization

All categorical values are presented as frequencies n (%).

[a]OPO readings are based on whole slide review, whereas pathologists' readings are based on biopsy images.

[b]Percentages are calculated from the total number of images with identified arteries. For pathologist 1 n = 54, pathologist 2 n = 63, and pathologist 3 n = 86.

agreement was mild (weighted Cohen's kappa was 0.32, 0.53, and 0.85; PABAK was 0.87, 0.87, and 0.91 when comparing pathologists 2&3, 1&3, and 1&2, respectively; and Fleiss kappa [95% CI] was 0.41 [0.32, 0.50]) as shown in Figure 2 and Table 3.

3.9 | Clinical outcomes and kidney biopsies

Forty percent of donor kidneys were discarded. Fleiss kappa statistics for the three pathologists when assessing image sets of discarded kidneys are shown in Table S2. Among statistically significant kappa coefficients for discarded kidneys, agreement was similar (<0.10 difference in kappa coefficients) between discarded kidneys and the entire 100 image sets except for glomeruli thrombi and arterial intimal fibrosis, which had less agreement among discarded kidneys.

When we evaluated the distribution of discarded kidneys and cold ischemia time between kidneys with biopsies plus images (n = 425) and kidneys with biopsies alone (n = 1326), there was no statistically significant difference in discard rates [131 (31%) vs 456 (34%), $P = 0.175$, respectively]; however, cold ischemia time (hours) was significantly longer in kidneys with biopsies and images compared to biopsies alone (median [IQR] 8 [14, 23] vs 16 [12, 21], $P < 0.001$, respectively).

TABLE 3 Weighted Cohen's kappa, Prevalence-adjusted bias-adjusted kappa, and Fleiss kappa for all three pathologists per histological finding

	Weighted Cohen's kappa			Prevalence-adjusted bias-adjusted kappa				
	Pathologists			Pathologists				
Histological findings	1/2	1/3	2/3	1/2	1/3	2/3	Fleiss kappa (95% CI)	*P*-value*
Glomerulosclerosis	0.77	0.68	0.70	0.82	0.62	0.76	0.57 (0.45, 0.68)	<0.001
Glomeruli thrombi	0.85	0.53	0.32	0.91	0.87	0.87	0.41 (0.32, 0.50)	<0.001
Tubular atrophy	0.52	0.35	0.41	0.60	0.40	0.44	0.33 (0.24, 0.42)	<0.001
Arterial intimal fibrosis	0.59	0.34	0.42	0.19	0.05	0.16	0.33 (0.30, 0.36)	<0.001
Interstitial fibrosis	0.67	0.28	0.32	0.73	0.24	0.25	0.29 (0.20, 0.38)	<0.001
Interstitial inflammation	0.49	0.30	0.33	0.72	0.41	0.47	0.27 (0.18, 0.36)	<0.001
Arteriolar hyalinosis	0.38	0.27	0.16	0.78	0.57	0.54	0.21 (0.11, 0.31)	<0.001
Acute tubular injury	0.19	0.12	0.14	0.31	-0.07	0.30	0.07 (−0.01, 0.16)	0.065

*Z-test was used to calculate *P*-values for Fleiss kappa. No *P*-values were calculated for PABAK or weighted Cohen's kappa as only three values were available (kappa per pathologist).

Lastly, we assessed overall kappa among the three pathologists when the histological findings were scored as two categories of none/mild and moderate/severe as shown in Table S3. There were no significant changes in kappa values as compared to the 3-level categories except for arterial hyalinosis with reduction in agreement ($P = 0.015$) and interstitial fibrosis with improvement in agreement ($P = 0.028$).

4 | DISCUSSION

In this study of UNOS-uploaded deceased-donor photomicrographs of procurement kidney biopsies, we identified minimal to moderate inter-rater agreement among three experienced renal transplant pathologists using a standardized evaluation form with defined histological categories. The selection process for the 100 image sets was based on histological findings that have been identified in the literature to affect clinical decisions and discard rates.[16,17] Out of the eight histological findings assessed, global glomerulosclerosis had the highest inter-rater reliability but still had only mild to moderate agreement. Given the variability in interpreting these image sets, the field should consider investing in efforts to optimize the quality of biopsy specimens and the display of these specimens via photomicrographs to improve reliability across readers.

While other studies have identified pathologist experience and lack of standardized reporting as possible reasons for poor inter-rater agreement when utilizing physical biopsy slides,[7,18] we attempted to account for these factors in this novel evaluation of clinical biopsy images via a panel of experienced renal transplant pathologists using standardized evaluation forms. This is, in fact, the first study to evaluate biopsy images uploaded to the UNOS database and available for review by transplant centers during organ allocation offers. Nonetheless, we noted only minimal to mild agreement for the histological findings in images among the pathologists. Our results show that the pathologist who most recently completed fellowship had the best intra-rater agreement across most histological fields. This shows that number of years may not correlate with level of expertise, as it depends on training and exposure to cases in the years of experience. Furthermore, pathologists who are trained more recently are likely to navigate the medical records with more ease and hence could be more frequently exposed to evaluating biopsy images as compared to pathologists who completed fellowship years earlier.

Besides experience and reporting standards, several other reasons have been postulated to explain variability in reports of histological findings. It is important to recognize that while these potential drawbacks are inherent to current practice and apply to pathologists interpreting physical biopsy slides, they likely also apply to our study of clinical photomicrographs. First, even experienced renal pathologists may bypass standardized percentile definitions on evaluation forms in favor of their own interpretations for mild, moderate, or severe histopathology. Some of our definitions may have also been simplistic without specifications regarding the histological findings, which may have led to pathologists forgoing the standardized definitions provided and applying their own. With regard to global glomerulosclerosis in the current study, pathologists were asked to calculate the actual percentage, which could have contributed to variability because of the limited numbers of visible glomeruli (about five per image set). Glomerulosclerosis in particular requires substantial sample size, as noted in the Banff recommendations for at least seven glomeruli and the Pirani score recommendations for at least 25 glomeruli.[19,20] We calculated PABAK kappa, which assumes that prevalence bias is absent, and it did improve kappa values; however this is not reflective of the actual distribution of our dataset nor real life because prevalence is not fixed at 50% and its variation introduces bias. While frozen sections are typically used for procurement kidney biopsies (including the majority of the image sets for the current study) because of

significant time constraints around organ allocation, subtle histological findings can be more difficult to assess and thereby increase reporting variability compared with formalin-fixed tissue. Individual center practices may also influence pathologists, as similar patterns in reporting are more likely to exist between pathologists in the same center.[21] This was evident in our study as pathologists 1 and 2 were from the same institution and demonstrated the highest pairwise agreement.

It is important to recognize that when interpreting physical slides, pathologists can freely adjust the magnification, focus, and field of view for the specimen, but the interpretation of a static image is likely more reproducible as it eliminates variation induced by viewing different areas of a freely movable slide.[21] As such, the static nature of images may lead to overestimation of the agreement between pathologists compared to usual practice. We acknowledge that some limitations of our study may have affected the agreement among pathologists when reviewing the biopsy images. Reviewing photomicrographs of frozen wedge biopsies limits the interpretation of vascular structures as it has been shown that core needle biopsies are superior for evaluating renal vascular histology.[22] Image quality could have also been poor and limited the interpretation of some of the histological findings. All of the images used in this study were hematoxylin and eosin stained and hence could have limited the interpretation of findings such as fibrosis, which require special staining to be accurately described. In addition to low-quality images and staining, many other factors affect the quality of biopsy specimens, which in turn would impact the quality of the photomicrographs uploaded in UNOS. The thickness of the specimen, the quality of staining, and presence of artifacts secondary to freezing are just some of the factors that impact specimen quality. If biopsy processing improves then photomicrographs of these biopsies may also improve and lead to more accurate interpretations by experienced pathologists as well as clinicians reviewing these images. Given that interpretations of histological findings on biopsies can be associated with both kidney discard and organ acceptance, addressing the quality of specimen processing is crucial. Images taken from good quality specimens may enhance and build on the knowledge gained from a biopsy interpretation alone. An image may also aid physicians in further assessing the severity of findings since visualizing a photomicrograph can clarify if "moderate" is closer to 26% or 50% in severity. Although low-quality images of standard specimen processing in our study did not appear to influence decisions regarding kidney discard, they did have significantly longer cold ischemia time compared to biopsies without images. As there are no guidelines for obtaining photomicrographs from procurement kidney biopsies, it is difficult to ascertain why biopsies with photomicrographs had significantly longer cold ischemia time compared to biopsies without photomicrographs. However, we can postulate that kidneys with significant histological findings that prompt the pathologist to obtain photomicrographs will also have longer cold ischemia time likely due to an instinctive reluctance to accept kidneys with abnormal histology. Higher-quality images of better processed kidney specimens may further impact clinical decisions posttransplantation as they are used routinely in recipient protocol and for-cause biopsies.

Another limitation in our study is that two pathologists were from the same institution, which may have led to additional overestimation of agreement. As a result, we believed it was important to recruit the third expert pathologist from a different institution to account for some degree of variability in renal pathology practices between institutions.

There are also statistical limitations regarding the use of kappa. The kappa statistic is most useful for testing agreement for binary outcomes that are not due to chance, and most histological findings were scored on an ordinal scale. To accommodate for ordinal histological definitions, we used weighted Cohen's kappa when appropriate. Also, kappa is influenced by trait prevalence.[23] Thus, the generalizability of our results would be limited if the distributions of histological findings in our cohort do not resemble that of the general population of deceased-donor kidney biopsies.

In conclusion, we found moderate to almost perfect intrarater agreement but minimal to moderate inter-rater agreement among the three expert pathologists for important histopathological findings on clinical photomicrographs of procurement kidney biopsies. These results raise concerns about the reliability of uploaded biopsy images, and it may be that replacing static images with whole slide imaging would increase the clinical value of donor biopsy. Pending future studies to assess how uploaded biopsy images are used clinically, the field may consider seeking higher-quality standards for biopsy processing and display via photomicrographs.

ACKNOWLEDGEMENTS

We are tremendously grateful for the study participation of our collaborators at the following organ procurement organizations: Gift of Life Philadelphia, New York Organ Donor Network, Michigan Organ and Tissue Donation Program, New Jersey Sharing Network, and New England Organ Bank.

The content is the responsibility of the authors alone and does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. These organizations were not involved in study design, analysis, interpretation, or manuscript creation.

The data reported here have been supplied by the United Network for Organ Sharing (UNOS) as the contractor for the Organ Procurement and Transplantation Network (OPTN). The interpretation and reporting of these data are the responsibility of the author(s) and in no way should be seen as an official policy of or interpretation by the OPTN or the US Government.

CONFLICT OF INTEREST

None.

ORCID

Isaac E. Hall http://orcid.org/0000-0003-0885-8450

Peter P. Reese http://orcid.org/0000-0003-1440-069X

Chirag R. Parikh http://orcid.org/0000-0001-9051-7385

REFERENCES

1. Organ Procurement and Transplantation Network online data reports. Secondary Organ Procurement and Transplantation Network online data reports. https://optn.transplant.hrsa.gov/data/view-data-reports/national-data.
2. United Network Organ Sharing Data Reports: Transplant Trends. Secondary United Network Organ Sharing Data Reports: Transplant Trends. https://www.unos.org/data/transplant-trends/-transplants_by_organ_type+year+2016/.
3. Stewart DE, Garcia VC, Rosendale JD, Klassen DK, Carrico BJ. Diagnosing the decades-long rise in the deceased donor kidney discard rate in the United States. *Transplantation*. 2017;101(3):575-587.
4. Cho YW, Shah T, Cho ES, et al. Factors associated with discard of recovered kidneys. *Transplant Proc*. 2008;40(4):1032-1034.
5. Reese PP, Harhay MN, Abt PL, Levine MH, Halpern SD. New solutions to reduce discard of kidneys donated for transplantation. *J Am Soc Nephrol*. 2016;27(4):973-980.
6. 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*. 1998;13(1):165-172.
7. 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(5):1161-1168.
8. Liapis H, Gaut JP, Klein C, et al. Banff histopathological consensus criteria for preimplantation kidney biopsies. *Am J Transplant*. 2017;17(1):140-150.
9. Elmore JG, Longton GM, Pepe MS, et al. A randomized study comparing digital imaging to traditional glass slide microscopy for breast biopsy and cancer diagnosis. *J Pathol Inform*. 2017;8:12.
10. Reese PP, Hall IE, Weng FL, et al. Associations between deceased-donor urine injury biomarkers and kidney transplant outcomes. *J Am Soc Nephrol*. 2015;27(5):1534-1543.
11. Wood JM. Understanding and Computing Cohen's Kappa: A Tutorial. Secondary Understanding and Computing Cohen's Kappa: A Tutorial; 2007. https://wpe.info/papers_table.html.
12. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull*. 1968;70(4):213-220.
13. Chen G, Faris P, Hemmelgarn B, Walker RL, Quan H. Measuring agreement of administrative data with chart data using prevalence unadjusted and adjusted kappa. *BMC Med Res Methodol*. 2009;9:5.
14. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med*. 2012;22(3):276-282.
15. Flack V, Lachenbruch PA, Schouten H, et al. Sample size determinations for the two rater kappa statistic. *Psychometrika*. 1988;53(3):321-325.
16. Hall IE, Schroppel B, Doshi MD, et al. Associations of deceased donor kidney injury with kidney discard and function after transplantation. *Am J Transplant*. 2015;15(6):1623-1631.
17. Sung RS, Christensen LL, Leichtman AB, et al. Determinants of discard of expanded criteria donor kidneys: impact of biopsy and machine perfusion. *Am J Transplant*. 2008;8(4):783-792.
18. Lewis JS Jr, Tarabishy Y, Luo J, et al. Inter- and intra-observer variability in the classification of extracapsular extension in p16 positive oropharyngeal squamous cell carcinoma nodal metastases. *Oral Oncol*. 2015;51(11):985-990.
19. Corwin HL, Schwartz MM, Lewis EJ. The importance of sample size in the interpretation of the renal biopsy. *Am J Nephrol*. 1988;8(2):85-89.
20. Remuzzi G, Grinyo 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(12):2591-2598.
21. Furness PN, Taub N, Assmann KJ, et al. International variation in histologic grading is large, and persistent feedback does not improve reproducibility. *Am J Surg Pathol*. 2003;27(6):805-810.
22. 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(1):37-42.
23. Thompson WD, Walter SD. A reappraisal of the kappa coefficient. *J Clin Epidemiol*. 1988;41(10):949-958.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Mansour SG, Hall IE, Reese PP, et al. Reliability of deceased-donor procurement kidney biopsy images uploaded in United Network for Organ Sharing. *Clin Transplant*. 2018;32:e13441. https://doi.org/10.1111/ctr.13441