Published in final edited form as:
Arch Pathol Lab Med. 2018 May ; 142(5): 613–625. doi:10.5858/arpa.2017-0181-OA.

Reproducibility and Feasibility of Strategies for Morphologic Assessment of Renal Biopsies Using the Nephrotic Syndrome Study Network Digital Pathology Scoring System

Jarcy Zee, PhD, **Jeffrey B. Hodgin, MD, PhD**, **Laura H. Mariani, MD, MS**, **Joseph P. Gaut, MD, PhD**, **Matthew B. Palmer, MD, PhD**, **Serena M. Bagnasco, MD**, **Avi Z. Rosenberg, MD, PhD**, **Stephen M. Hewitt, MD, PhD**, **Lawrence B. Holzman, MD**, **Brenda W. Gillespie, PhD**, and **Laura Barisoni, MD**

Biostatistics, Arbor Research Collaborative for Health, Ann Arbor, Michigan (Dr Zee); the Departments of Pathology (Dr Hodgin), Internal Medicine (Dr Mariani), and Biostatistics (Dr Gillespie), University of Michigan, Ann Arbor; Arbor Research Collaborative for Health, Ann Arbor, Michigan (Dr Mariani); the Department of Pathology & Immunology, Washington University, St Louis, Missouri (Dr Gaut); the Departments of Pathology and Laboratory Medicine (Dr Palmer) and Medicine (Dr. Holzman), University of Pennsylvania, Philadelphia; the Department of Pathology, Johns Hopkins University, Baltimore, Maryland (Drs Bagnasco and Rosenberg); the Kidney Diseases Branch, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, Maryland (Dr Rosenberg); the Laboratory of Pathology, National Cancer Institute, Bethesda, Maryland (Dr Hewitt); and the Department of Pathology, University of Miami, Miami, Florida (Dr Barisoni)

Abstract

Context—Testing reproducibility is critical for the development of methodologies for morphologic assessment. Our previous study using the descriptor-based Nephrotic Syndrome Study Network Digital Pathology Scoring System (NDPSS) on glomerular images revealed variable reproducibility.

Objective—To test reproducibility and feasibility of alternative scoring strategies for digital morphologic assessment of glomeruli and explore use of alternative agreement statistics.

Design—The original NDPSS was modified (NDPSS1 and NDPSS2) to evaluate (1) independent scoring of each individual biopsy level, (2) use of continuous measures, (3) groupings of individual descriptors into classes and subclasses prior to scoring, and (4) indication of pathologists' confidence/uncertainty for any given score. Three and 5 pathologists scored 157 and 79 glomeruli using the NDPSS1 and NDPSS2, respectively. Agreement was tested using

Reprints: Laura Barisoni, MD, Department of Pathology, University of Miami, Miller School of Medicine, 1400 NW 12th Ave, Room 4079, Miami, FL 33136 (lbarisoni@med.miami.edu).

Dr Zee and Dr Barisoni have equally contributed to the study design and writing of the manuscript.

Supplemental digital content is available for this article. See text for hyperlink.

Dr Laura Barisoni was previously a consultant for Amicus, Protalix, and Shire and is currently a consultant for Moderna for Fabry disease clinical trials using digital pathology. The other authors have no relevant financial interest in the products or companies described in this article.

conventional (Cohen κ) and alternative (Gwet agreement coefficient 1 [AC_1]) agreement statistics and compared with previously published data (original NDPSS).

Results—Overall, pathologists' uncertainty was low, favoring application of the Gwet AC_1. Greater agreement was achieved using the Gwet AC_1 compared with the Cohen κ across all scoring methodologies. Mean (standard deviation) differences in agreement estimates using the NDPSS1 and NDPSS2 compared with the single-level original NDPSS were −0.09 (0.17) and −0.17 (0.17), respectively. Using the Gwet AC_1, 79% of the original NDPSS descriptors had good or excellent agreement. Pathologist feedback indicated the NDPSS1 and NDPSS2 were time-consuming.

Conclusions—The NDPSS1 and NDPSS2 increased pathologists' scoring burden without improving reproducibility. Use of alternative agreement statistics was strongly supported. We suggest using the original NDPSS on whole slide images for glomerular morphology assessment and for guiding future automated technologies.

In the setting of clinical trials and translational research, the morphologic evaluation of renal biopsies has progressively transitioned from use of conventional light microscopy to digital pathology on whole slide images (WSIs).[1–3] Previous studies have revealed that interreader and intrareader reproducibility of morphology scoring or diagnoses are generally higher when using WSIs and enhanced by annotation.[1,4–9] The establishment of digital pathology repositories also facilitates the testing of different scoring systems and metrics, simultaneously or at different times, using the same set of WSIs.[2]

The Nephrotic Syndrome Study Network (NEPTUNE) pathology working group pioneered the establishment of the NEPTUNE digital pathology protocol to enable standardized morphologic assessment of digital renal biopsies from children and adults with minimal change disease (MCD), focal segmental glomerulosclerosis (FSGS), and membranous nephropathy (MN).[2,10] The NEPTUNE digital pathology protocol includes protocols to populate a digital pathology repository, to annotate (enumerate) individual glomeruli across biopsy levels, to morphologically assess renal biopsies using the descriptor-based NEPTUNE Digital Pathology Scoring System (NDPSS), and for digital morphometry.[2,11] This multicenter effort has served as a model for other international consortia such as the International Digital Nephropathology Network.[12]

A critical element in establishing new scoring systems, besides their clinical significance, is their reproducibility. Reproducibility can be modulated by several factors, including pathologists' training, the type of lesions being scored, the metrics, or the statistical approach applied.[12] For example, cross-training of pathologists prior to scoring and grouping of individual descriptors that share common features into categories can increase reproducibility.[3,4] Furthermore, morphologic features captured as dichotomous measures (ie, present versus absent) may be better represented by continuous measures. Lastly, the agreement statistic used to evaluate reproducibility needs to be carefully chosen. For example, the Cohen κ is conventionally used in pathology partly because it makes a correction for agreement by chance, but it also inherently assumes that all ratings may be rated randomly.[13,14] However, when the scoring process is performed by experts and preceded by rigorous cross-training processes, it is plausible that only a portion of observations is subject to random ratings, in which case the Cohen κ may overestimate and therefore overcorrect for chance agreement. An alternative agreement statistic that tends to be more liberal by assuming a lower proportion of random ratings may be more suitable in this case, such as the Gwet agreement coefficient 1 (AC_1).[14,15]

Although the NDPSS was designed to include all biopsy levels available for assessment, our first reproducibility test was conducted on single static (JPEG) images of glomeruli.[3] With the current study, we aim to explore reproducibility and feasibility of alternative scoring strategies, metrics, and statistical approaches for optimizing the original NDPSS, with the goal of establishing a robust methodology for morphologic assessment of digital renal biopsies in the settings of clinical research, clinical trials, and ultimately routine practice.

MATERIALS AND METHODS

Study Cohort

The WSIs included in this study are part of the set of cases enrolled in the multicenter and multiethnic prospective cohort study NEPTUNE.[10] As previously described, renal biopsy material was collected according to the NEPTUNE digital pathology protocol and made available to study pathologists through password-protected access to the NEPTUNE digital pathology repository.[2]

Overall Study Design

Our study was designed to address 3 goals: (1) to test whether alternative scoring strategies and metrics improve interpathologist reproducibility, we modified the original NDPSS to create the NDPSS1 and NDPSS2; (2) to determine the statistical approach that would most accurately measure the agreement (or disagreement) among pathologists, we compared Cohen κ and Gwet AC_1 statistics across all scoring strategies; and (3) to evaluate the feasibility of each scoring strategy, we collected pathologists' feedback on the use of the different approaches.

Scoring Systems

Original NDPSS—Previously published scoring data using the original NDPSS were retrieved from the NEPTUNE database and reanalyzed in the current study. Data were previously obtained by 12 pathologists, who reviewed 315 JPEG images of individual glomeruli (equivalent to assessing the glomerular profile on a single biopsy level) and recorded the presence or absence of 51 glomerular descriptors using an electronic scoring matrix (Figure 1, A).[3] In the current study, we used scores from 39 of 51 descriptors pertinent to MCD, FSGS, and MN; we also generated classes and subclasses of descriptors by applying postscoring grouping strategies mimicking those used in NDPSS1 and NDPSS2 described below (Tables 1 and 2; Figure 2, A through D).

The NDPSS1

Scoring Strategies: An electronic scoring matrix specifically designed for NDPSS1 (Figure 1, B) was used to test alternative scoring strategies (Table 3), including the use of all biopsy levels (Figure 3, A through F), grouping descriptors prior to scoring (Tables 1 and 2), and the application of ordinal-scale and continuos-scale scoring (Tables 1 and 2).

Pathologist Training: Three NEPTUNE pathologists received 2 hours of training using an online webinar to review the NDPSS1 scoring protocol and the corresponding electronic scoring matrix (Figure 1, B). Understandability of the NDPSS1 protocol was then tested by having each pathologist score 4 example glomeruli, which was then followed by an additional 2 hours of webinar discussion and cross-training to increase reproducibility.

Case Selection and Distribution: The NEPTUNE database contains cases previously scored using the original NDPSS. From these data, we identified glomeruli with high numbers of structural features present to maximize the information gained from each glomerulus. 157 glomeruli from 60 FSGS/MCD and 2 MN cases were selected to test NDPSS1. Each case contributed between 1 and 5 glomeruli and had at least 1 WSI of a biopsy section stained with hematoxylin-eosin, periodic acid–Schiff, trichrome, or silver. Cases were randomly assigned to each of the 3 scoring pathologists such that each pathologist scored about 100 glomeruli, with overlap such that each glomerulus would have 2 sets of scores.

The NDPSS2

Scoring Strategies: Based on initial reproducibility estimates using the Cohen κ statistic and pathologists' feedback from the NDPSS1 study (see Results), a second set of scoring strategies, the NDPSS2, was implemented (Table 3) and was recorded on an electronic scoring matrix specifically designed for the NDPSS2 (Figure 1, C). The scoring strategies included scoring of individual biopsy sections/levels independently (Figure 3), different groupings of individual descriptors (Tables 1 and 2; Figure 2), continuous-scale scoring (Tables 1 and 2), a gestalt overall damage score, and indication of poor image quality and stain type.

Pathologist Training: Two additional NEPTUNE pathologists were added to the study, and all 5 pathologists collectively reviewed the results of the NDPSS1 data using case examples and discussed disagreements. The 5 pathologists received a 2-hour webinar training to review the NDPSS2 scoring protocol and the corresponding electronic scoring matrix (Figure 1, C). All pathologists participated in a practice round by scoring every level of 2 glomeruli to ensure understandability of the scoring protocol, followed by an additional 2 hours of cross-training to improve reproducibility.

Case Selection and Distribution: A total of 79 annotated glomeruli on WSIs from the same 60 FSGS/MCD and 2 MN NEPTUNE cases were scored. Each case contributed between 1 and 5 glomeruli. Cases were randomly assigned to each of the 5 scoring pathologists, with overlap such that each glomerulus would have at least 2 sets of scores to evaluate interpathologist reproducibility.

Statistical Analysis Strategies

Cohen κ Statistic and Intraclass Correlation Coefficient Across All Scoring Strategies

Original NDPSS: Interpathologist agreement estimates for 39 MCD, FSGS, and MN glomerular descriptors recorded as present or absent were extracted from the set of data previously published for comparison with NDPSS1 and NDPSS2 (see Table 2 in Barisoni et al,[3] 12 NEPTUNE & Non-NEPTUNE pathologists II Kappa column). Using the same data set, descriptors were grouped, postscoring, into classes and subclasses that mimicked those scored in NDPSS1 and NDPSS2, and Cohen κ was calculated for each grouping. Average prevalence across pathologists was estimated for individual descriptors and subclasses and classes of descriptors to aid interpretation of agreement estimates.

The NDPSS1: To calculate interpathologist agreement using the Cohen κ and to determine average prevalence across pathologists, we dichotomized each ordinal glomerular descriptor score using different cut points (eg, 0 versus 0.25–1, 0–0.25 versus 0.50–1). To evaluate whether indicating probabilities of presence or absence of individual descriptors or classes or subclasses of descriptors improved reproducibility, the intraclass correlation coefficient (ICC) was also calculated for both dichotomized and ordinal measures. Reproducibility for subclasses of descriptors scored as a percentage was assessed using the ICC.

The NDPSS2: We calculated interpathologist reproducibility of glomerulus-level scores for each glomerular descriptor using the Cohen κ and the average prevalence across pathologists. The ICC was calculated for subclasses of descriptors scored as a percentage and case-level gestalt overall damage scores. The analysis was repeated after exclusion of any poor-quality images and was performed separately by stain types in NDPSS2.

Gwet AC_1 Statistic Across All Scoring Strategies—Although the Cohen κ is useful to compare with historical studies, it can be sensitive to prevalence and assumes that all observations may be rated randomly and thus that all are susceptible to chance agreement.[13,14] We hypothesized that this assumption could be violated in nephropathology. Our hypothesis was generated by observing pathologists' behavior and the low rate of uncertainty. Thus, we explored a different statistical approach using the Gwet AC_1, which is less sensitive to descriptor prevalence and assumes only that an unknown proportion of observations are subject to chance agreement. The Gwet AC_1 statistic was applied to estimate interpathologist reproducibility of each dichotomous WSI glomerular descriptor, class, and subclass of descriptors across all scoring strategies. Reproducibility was compared between Cohen κ and Gwet AC_1 estimates across all scoring strategies.

To assess the suitability of these agreement statistics, we used the results from strategy 3 of NDPSS1 to estimate the proportion of glomeruli with random ratings. Any class, subclass, or individual descriptor scored as 0.25, 0.50, or 0.75 was considered to have some uncertainty. Thus, if pathologists were instructed to score presence or absence dichotomously for these glomeruli, they would have had to use some degree of randomness to determine scores. Because each glomerulus had 2 sets of scores, glomeruli scored with uncertainty by either pathologist were considered to be subject to random ratings. All statistical analyses were conducted using R version 3.2.3 (R Foundation for Statistical Computing, Vienna, Austria).

Pathologists' Feedback

At the end of the NDPSS1 scoring, we asked pathologists for feedback on the scoring strategies applied compared with the original NDPSS, focusing specifically on understandability, effectiveness of the training prior to scoring, ease of use, and time spent. Feedback from the NDPSS1 scoring strategies was used to design the NDPSS2. At the end of the NDPSS2 scoring, feedback from the scoring pathologists was again recorded.

RESULTS

Original NDPSS

The prevalence of most individual descriptors was low, with only 8 individual descriptors having prevalence of 10% or greater (Figure 4, a). Agreement for descriptors with such low prevalence must be interpreted with caution, that is, only as negative agreement or agreement that the descriptor is absent. Cohen κ agreement estimates for the NDPSS individual descriptors have been previously reported[3] and are illustrated here in Figure 4, a. Six of 39 individual descriptors had $\kappa \geq 0.6$, and only 1 of 8 individual descriptors with a prevalence of 10% or greater had $\kappa \geq 0.6$. Agreement generally improved after grouping of descriptors postscoring into subclasses and further after grouping of subclasses into classes, similarly to that previously reported (Figure 4, a).[3] Six of 11 subclasses and 4 of 5 classes with at least 10% prevalence had $\kappa \geq 0.6$.

Gwet AC_1 agreement estimates for individual descriptors and postscoring groupings into classes and subclasses of descriptors are illustrated in Figure 4, b. In general, agreement estimates using the Gwet AC_1 were slightly higher than the corresponding κ estimates for the 8 individual descriptors with prevalence of 10% or greater, but overall much higher. Specifically, using the Gwet AC_1 statistic, agreement was 0.60 or greater in 6 of 8 individual descriptors, 9 of 11 subclasses, and 4 of 5 classes of descriptors with a prevalence of 10% or greater.

The NDPSS1

The NDPSS1 used all biopsy levels for scoring rather than a single JPEG image as in the published NDPSS study. Across the 57 individual descriptors and subclasses and classes of descriptors, the percentage of glomeruli for which pathologists indicated uncertainty had a median of 8.2 and interquartile range of 1.9 to 16.6 (Supplemental Figure 1; see supplemental digital content). The subclass other segmental lesions had the greatest uncertainty at 48% of glomeruli, followed by glomerular foam cells at 37% uncertainty, synechia at 32%, and mesangiopathic changes at 26%. All other descriptors were rated with uncertainty for less than 25% of glomeruli. Overall, the 0.25, 0.50, and 0.75 scores were rarely used by pathologists.

We hypothesized that the use of ordinal probabilities may improve agreement compared with the dichotomous scores used in the original NDPSS. Given that pathologists had low levels of uncertainty, it was not surprising that dichotomization of the ordinal probabilities made little difference in agreement. Specifically, reproducibility of the probability of presence for each descriptor was almost identical to reproducibility when probabilities were dichotomized, no matter which cut point for dichotomization was used (Supplemental Figure 2). Similarly, ICCs calculated on probabilities of presence of descriptors were close to those calculated on dichotomized versions of descriptors. Subsequent results are shown for dichotomized descriptors comparing 0 to 0.25 (no or probably no) with 0.5 to 1 (maybe, probably yes, or yes).

Forty-five of 57 individual and groups of dichotomized descriptors had prevalence of 10% or greater. Five of the 45 with prevalence of 10% or greater had agreement estimates greater than 0.6 using the Cohen κ (Figure 5, a), including 1 individual descriptor (segmental hyaline droplets in epithelial cells), 3 subclasses (tip lesion, segmental hyalinosis, and podocyte hyaline droplets), and 1 class (podocyte injury).

Thirty-eight of the 45 individual and groups of dichotomized descriptors with prevalence of 10% or greater had agreement greater than 0.6 when the Gwet AC_1 statistic was used (Figure 5, b), including 24 individual descriptors, 11 subclasses, and 3 classes of descriptors. Compared with the Cohen κ, the Gwet AC_1 agreement estimates tended to be higher, particularly for descriptors with lower prevalence (Figure 6). The Gwet AC_1 was also able to show that descriptors with the highest amounts of uncertainty (ie, other segmental lesions, glomerular foam cells, synechia, and mesangiopathic changes) had lower agreement relative to other descriptors. The grouping of individual descriptors in classes and subclasses did not always result in higher agreement estimates using the Cohen κ or the Gwet AC_1.

The NDPSS1 also tested whether scoring on a continuous scale rather than having to specify segmental or global proliferation of lesion would improve agreement. The ICC estimates for the 8 subclasses and classes scored as a percentage ranged from 0.04 (collapse) to 0.95 (spikes) (Supplemental Figure 3). Only the ICC for percentage spikes was greater than 0.6.

The pathologists' feedback from NDPSS1 was instrumental in developing alternative scoring strategies for the NDPSS2. First, the hierarchy of classes and subclasses was sometimes unclear. Thus, an alternative hierarchy was generated for the NDPSS2. Second, pathologists expressed that the process of calculating the percentage of the glomerulus involved by a specific descriptor across multiple biopsy levels was time consuming and inefficient. Doing so involved memorization of lesions seen on many sections and mental formulation of the 3-dimensional glomerulus. Thus, the NDPSS2 scoring matrix allowed for scoring of the percentage of the glomerular section affected by a specific lesion using separate columns for each of the levels. Third, pathologists indicated that scoring the uncertainty of the presence or absence of a specific lesion was an "unnatural" process compared with the original dichotomous approach. Especially given that pathologists rarely used the probability option, it was removed from NDPSS2. Additional training was provided to assure complete understanding of the scoring process and matrix.

The NDPSS2

Nineteen of the 21 individual or groups of descriptors scored as dichotomous measures had an average prevalence of 10% or greater. Nine of these 19 had an interrater agreement greater than 0.6 using the Cohen κ (Figure 7, a) versus 12 using the Gwet AC_1 (Figure 7, b). For 10 of 19, the Gwet AC_1 was 0.8 or greater. Similarly to what we observed in the original NDPSS and in NDPSS1, Gwet AC_1 estimates were generally higher compared with Cohen κ estimates. In particular, the Cohen κ was less than 0.6 whereas the Gwet AC_1 was greater than 0.6 for the global wrinkling and segmental sclerosis subclasses and the global obliteration class.

The ICC estimates for the 8 descriptors scored as a percentage varied between 0.2 for any deflation and 0.8 for any sclerosis or tip lesion (Supplemental Figure 4). Only 3 of these had good agreement (ICC ≥ 0.60). The ICC estimate for the gestalt overall damage score was 0.78. All results were similar when excluding the 0% to 27% of images that pathologists indicated had poor quality. Some differences in reproducibility estimates were noted when stratified by stain type. Although silver- and trichrome-stained sections yielded similar results compared with overall results, periodic acid–Schiff stains gave slightly worse and hematoxylin-eosin stains moderately worse reproducibility estimates.

Pathologists reported that percentages were more easily determined for each biopsy section individually as compared with the entire glomerulus in NDPSS1. However, reporting percentages in deciles was much more time-consuming than using the original dichotomous approach. In addition, the conclusion was that recording scores for each section of each glomerulus independently, although possible for a small pilot study, would not be feasible for a much larger study. Notably, aggregating section-level scores to the glomerulus level did not result in substantially different reproducibility estimates compared with scoring done directly at the glomerulus level.

Comparison Between the Original NDPSS and Modified Versions

We focus our comparison between the original and modified versions of the NDPSS only on individual descriptors and groups of descriptors with moderate prevalence (≥10%). With only a few exceptions, agreement estimates using either the Cohen κ (Figures 5, a, and 7, a) or the Gwet AC_1 (Figures 5, b, and 7, b) from the NDPSS1 or NDPSS2 were lower than or similar to those from the single-level original NDPSS. Mean (standard deviation) differences in agreement estimates using NDPSS1 and NDPSS2 compared with the single-level original NDPSS were −0.14 (0.19) and −0.07 (0.18), respectively, using the Cohen κ, and −0.09 (0.17) and −0.17 (0.17), respectively, using the Gwet AC_1. Additionally, the reproducibility of classes and subclasses was independent of whether the grouping was done prescoring (as in NDPSS1 and NDPSS2) or postscoring (as in the original NDPSS).

DISCUSSION

Robust scoring and classification systems for diseases are based on several critical elements, including standardization, comprehensiveness, objectivity, accuracy, and reproducibility.[12] Historically, most scoring systems to evaluate various organs, including the renal parenchyma, were generated prior to testing of their clinical relevance or reproducibility.[16] Only recently, with the Oxford scoring system for immunoglobulin A nephropathy, was reproducibility of measures first assessed to determine which parameters should be considered for further validation— for example, by assessing associations with clinical outcomes—and subsequently included in the test score.[17] This approach ensures that only parameters that are both reproducible and clinically relevant are included in the classification system. The 2-step procedure can dramatically reduce the number of parameters tested for clinical relevance, which not only saves time and resources but also mitigates statistical concerns about spurious findings resulting from multiple comparisons. In implementing such an approach, however, investigators must be wary of inadvertently discarding important features because of seemingly poor reproducibility.

In our previous studies we tested the reproducibility of the NDPSS across multiple pathologists using single biopsy images and conventional statistical measures (Cohen κ). Our initial results, although promising, indicated that not all parameters were reproducible.[3] In an effort to optimize the reproducibility of the NDPSS, we tested 2 sets of alternative scoring strategies using a common set of WSIs and evaluated alternative statistical approaches for estimating agreement.

The current study has elucidated some of the aspects that modulate reproducibility and ultimately the choice of one scoring strategy versus others. For example, we hypothesized that ordinal or continuous versions of descriptors would better capture pathologists' scores than their dichotomized versions and thus would result in higher reproducibility. However, pathologists rarely indicated probabilities of presence of descriptors. This is not entirely surprising, because it is in the nature of their training and routine operation to use a dichotomous approach (eg, presence or absence of a lesion), minimize uncertainty, and commit to a diagnosis. Reproducibility estimates using the ordinal or continuous measures were not substantially improved, and were sometimes worse, across all alternative strategies. One exception was with the estimation of overall damage score. However, the damage score is a subjective measure that does not contain any qualitative (eg, type of lesion) or quantitative (eg, amount of lesions present in renal tissue) information. Thus, it has limited practical use besides its potential predictive value. Last, estimating percentages of the glomerular tuft with a specific lesion was also considered challenging, time consuming, and tedious, whether achieved using all biopsy levels available for a comprehensive estimate or each level separately.

Because our previous study demonstrated increased reproducibility when individual descriptors were grouped,[3,4] we tested different strategies for descriptor grouping. We compared reproducibility from directly scoring the groups with creating groups of individual descriptors after scoring. The former strategy did not result in improved reproducibility. Poor reproducibility of individual descriptors and higher reproducibility of classes and subclasses could be an argument in favor of eliminating many individual descriptors from the scoring process. However, 2 counterarguments can be made: first, some of the nonreproducible descriptors may be clinically relevant, and second, reproducibility may increase with training.[3] Thus, discarding these granular descriptors is probably premature. Additionally, scoring of some classes and subclasses proved to be impractical because individual descriptors did not always fit neatly into a hierarchy. Thus, because individual descriptors can always be grouped after scoring and in various configurations, the best approach is to score descriptors individually as per the original NDPSS. Overall, the dichotomous scoring of individual descriptors in the original NDPSS was considered the most feasible and reproducible strategy.

With this study, we also tested alternative statistical approaches for estimating agreement other than the commonly used Cohen κ. The Cohen κ statistic is based on the assumption that all observations may be rated with randomness. However, this assumption is likely to be too stringent given that pathologists rarely indicated uncertainty during the scoring tests. Low uncertainty suggests that although each individual pathologist's evaluation process may be slightly different, it is likely that the primary reason for agreement or disagreement is not random. The Gwet AC_1 statistic is more liberal than the Cohen κ in that its correction for chance agreement is not as high and its assumption that only an unknown proportion of observations is rated with randomness appears to be more plausible. Our results also showed that the Gwet AC_1 could better identify unreliable descriptors in terms of pathologists' indications of uncertainty, as descriptors with more uncertainty had lower Gwet AC_1 estimates and vice versa. The Gwet AC_1 may thus provide a more accurate estimation of agreement for the descriptor-based NDPSS. Although in some studies, for example those designed to flag discordant raters, a more conservative agreement statistic like the Cohen κ may be more prudent, a liberal agreement statistic like the Gwet AC_1 would be less likely to incorrectly discard important descriptors prior to subsequent validation studies. Statistical methodology research is also currently in progress to develop agreement statistics that may be more accurate by empirically estimating the probability of chance agreement.

Our study has 2 limitations worth noting. First, we had a relatively small sample size of glomeruli to test new scoring strategies. However, a larger study would not have been feasible to test the many strategies under consideration on numerous cases. As a pilot study, however, the current study was able to identify the most feasible scoring strategies based on the small sample size. Second, although glomeruli were specifically chosen among those with multiple structural features, we still had low prevalence of some descriptors, thus limiting interpretation of agreement estimates. Low prevalence may result in high negative agreement masking low positive agreement, for example when pathologists agree on absence of a descriptor but disagree on its presence. This may be solved by reporting both positive and negative agreement estimates. However, low prevalence implies there would be few observations to calculate positive agreement, and currently available chance-corrected positive and negative agreement statistics do not offer additional information beyond the Cohen κ.[18] Therefore, in this study, we report prevalence with all agreement estimates and advise caution in interpretation. Larger reproducibility studies using the original NDPSS are in progress and will also be helpful for evaluating these rare descriptors.

Despite these limitations, we demonstrated that the alternative strategies (NDPSS1 and NDPSS2) increased pathologists' scoring burden without improving reproducibility. We also found empirical evidence to support the use of the Gwet AC_1 statistic for estimating agreement rather than the Cohen κ. Based on the Gwet AC_1 statistic, the NDPSS had a large proportion of descriptors with good to excellent reproducibility. The NDPSS using WSIs is currently being tested for clinical and biological relevance by assessing associations with outcomes and gene expression data. The NDPSS will further be used for developing novel classification systems for proteinuric glomerular diseases. Such integration of quantitative pathology with clinical and molecular studies has been identified as a critical component to the understanding of disease pathogenesis and categorization and for the development of targeted therapy and precision medicine.[19] Furthermore, by improving and expanding our understanding of structural changes that differentiate glomerular diseases, we can inform machine learning efforts to establish computer-automated methodologies for renal biopsy evaluation. Thus, the NEPTUNE digital pathology protocol and NDPSS provide an excellent platform for nephropathology research to inform morphologic profiling of renal biopsies in clinical practice.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The Nephrotic Syndrome Study Network Consortium (NEPTUNE), U54-DK-083912, is a part of the National Institutes of Health (NIH) Rare Disease Clinical Research Network (RDCRN), supported through a collaboration between the Office of Rare Diseases Research (ORDR), the National Center for Advancing Translational Sciences (NCATS), and the National Institute of Diabetes, Digestive, and Kidney Diseases (NIDDK). Additional funding and/or programmatic support for this project has also been provided by the University of Michigan, NephCure Kidney International, the Halpin Foundation, and Arbor Research Collaborative for Health. The views expressed in written materials or publications do not necessarily reflect the official policies of the Department of Health and Human Services, nor does mention by trade names, commercial practices, or organizations imply endorsement by the US Government.

References

1. Barisoni L, Jennette JC, Colvin R, et al. Novel quantitative method to evaluate globotriaosylceramide inclusions in renal peritubular capillaries by virtual microscopy in patients with Fabry disease. Arch Pathol Lab Med. 2012; 136(7):816–824. [PubMed: 22742555]
2. Barisoni L, Nast CC, Jennette JC, et al. Digital pathology evaluation in the multicenter Nephrotic Syndrome Study Network (NEPTUNE). Clin J Am Soc Nephrol. 2013; 8(8):1449–1459. [PubMed: 23393107]
3. Barisoni L, Troost JP, Nast C, et al. Reproducibility of the NEPTUNE descriptor-based scoring system on whole-slide images and histologic and ultrastructural digital images. Mod Pathol. 2016; 29(7):671–684. [PubMed: 27102348]
4. Wang H, Sima CS, Beasley MB, et al. Classification of thymic epithelial neoplasms is still a challenge to thoracic pathologists: a reproducibility study using digital microscopy. Arch Pathol Lab Med. 2014; 138(5):58–63.
5. Ozluk Y, Blanco PL, Mengel M, Solez K, Halloran PF, Sis B. Superiority of virtual microscopy versus light microscopy in transplantation pathology. Clin Transplant. 2012; 26(2):336–344. [PubMed: 21955102]
6. Jen KY, Olson JL, Brodsky S, Zhou XJ, Nadasdy T, Laszik ZG. Reliability of whole slide images as a diagnostic modality for renal allograft biopsies. Hum Pathol. 2013; 44(5):888–894. [PubMed: 23199528]
7. Reyes C, Ikpatt OF, Nadji M, Cote RJ. Intra-observer reproducibility of whole slide imaging for the primary diagnosis of breast needle biopsies. J Pathol Inform. 2014; 5(1):5. [PubMed: 24741464]
8. Furness P. A randomized controlled trial of the diagnostic accuracy of internet-based telepathology compared with conventional microscopy. Histopathology. 2007; 50(2):266–273. [PubMed: 17222256]

9. Rosenberg AZ, Palmer M, Merlino L, et al. The application of digital pathology to improve accuracy in glomerular enumeration in renal biopsies. PLoS One. 2016; 11(6):e0156441. [PubMed: 27310011]

10. Gadegbeku CA, Gipson DS, Holzman LB, et al. Design of the Nephrotic Syndrome Study Network (NEPTUNE) to evaluate primary glomerular nephropathy by a multidisciplinary approach. Kidney Int. 2013; 83(4):749–756. [PubMed: 23325076]

11. Nast CC, Lemley KV, Hodgin JB, et al. Morphology in the digital age: integrating high-resolution description of structural alterations with phenotypes and genotypes. Semin Nephrol. 2015; 35(3): 266–278. [PubMed: 26215864]

12. Barisoni L, Gimpel C, Kain R, et al. Digital pathology imaging as a novel platform for standardization and globalization of quantitative nephropathology. Clin Kidney J. 2017; 10(2): 176–187. [PubMed: 28584625]

13. Gwet K. Kappa statistic is not satisfactory for assessing the extent of agreement between raters. Stat Methods Interrater Reliability Assess. 2002; 1:1–5.

14. Zhao X, Liu JS, Deng K. Assumptions behind intercoder reliability indices. Ann Int Commun Assoc. 2013; 36(1):419–480.

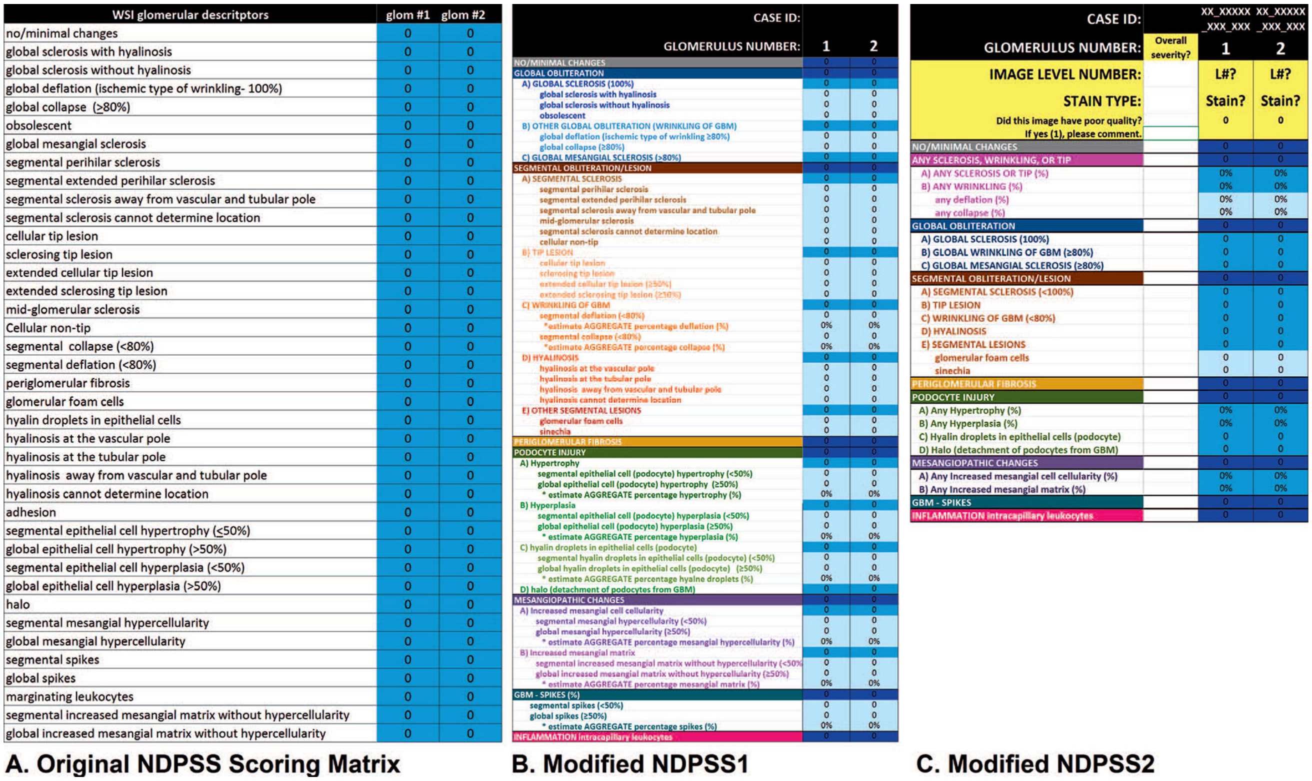
15. Gwet K. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. 2008; 61(pt 1):29–48. [PubMed: 18482474]

16. Haas M, Rastaldi MP, Fervenza C. Histologic classification of glomerular diseases: clinicopathologic correlations, limitations exposed by validation studies, and suggestions for modification. Kidney Int. 2014; 86(3):648. [PubMed: 25168500]

17. Roberts IS, Cook HT, Troyanov S, et al. The Oxford classification of IgA nephropathy: pathology definitions, correlations, and reproducibility. Kidney Int. 2009; 76(5):546–556. [PubMed: 19571790]

18. Cicchetti DV, Feinstein AR. High agreement but low kappa, II: resolving the paradoxes. J Clin Epidemiol. 1990; 43(6):551–558. [PubMed: 2189948]

19. Rushing EJ, Wesseling P. Towards an integrated morphological and molecular WHO diagnosis of central nervous system tumors: a paradigm shift. Curr Opin Neurol. 2015; 28(6):628–632. [PubMed: 26402407]



A. Original NDPSS Scoring Matrix

WSI glomerular descritptors	glom #1	glom #2
no/minimal changes	0	0
global sclerosis with hyalinosis	0	0
global sclerosis without hyalinosis	0	0
global deflation (ischemic type of wrinkling- 100%)	0	0
global collapse (≥80%)	0	0
obsolescent	0	0
global mesangial sclerosis	0	0
segmental perihilar sclerosis	0	0
segmental extended perihilar sclerosis	0	0
segmental sclerosis away from vascular and tubular pole	0	0
segmental sclerosis cannot determine location	0	0
cellular tip lesion	0	0
sclerosing tip lesion	0	0
extended cellular tip lesion	0	0
extended sclerosing tip lesion	0	0
mid-glomerular sclerosis	0	0
Cellular non-tip	0	0
segmental collapse (<80%)	0	0
segmental deflation (<80%)	0	0
periglomerular fibrosis	0	0
glomerular foam cells	0	0
hyalin droplets in epithelial cells	0	0
hyalinosis at the vascular pole	0	0
hyalinosis at the tubular pole	0	0
hyalinosis away from vascular and tubular pole	0	0
hyalinosis cannot determine location	0	0
adhesion	0	0
segmental epithelial cell hypertrophy (≤50%)	0	0
global epithelial cell hypertrophy (>50%)	0	0
segmental epithelial cell hyperplasia (<50%)	0	0
global epithelial cell hyperplasia (>50%)	0	0
halo	0	0
segmental mesangial hypercellularity	0	0
global mesangial hypercellularity	0	0
segmental spikes	0	0
global spikes	0	0
marginating leukocytes	0	0
segmental increased mesangial matrix without hypercellularity	0	0
global increased mesangial matrix without hypercellularity	0	0

B. Modified NDPSS1

CASE ID:		
GLOMERULUS NUMBER:	1	2
NO/MINIMAL CHANGES	0	0
GLOBAL OBLITERATION	0	0
A) GLOBAL SCLEROSIS (100%)	0	0
global sclerosis with hyalinosis	0	0
global sclerosis without hyalinosis	0	0
obsolescent	0	0
B) OTHER GLOBAL OBLITERATION (WRINKLING OF GBM)	0	0
global deflation (ischemic type of wrinkling ≥80%)	0	0
global collapse (≥80%)	0	0
C) GLOBAL MESANGIAL SCLEROSIS (>80%)	0	0
SEGMENTAL OBLITERATION/LESION	0	0
A) SEGMENTAL SCLEROSIS	0	0
segmental perihilar sclerosis	0	0
segmental extended perihilar sclerosis	0	0
segmental sclerosis away from vascular and tubular pole	0	0
mid-glomerular sclerosis	0	0
segmental sclerosis cannot determine location	0	0
cellular non-tip	0	0
B) TIP LESION	0	0
cellular tip lesion	0	0
sclerosing tip lesion	0	0
extended cellular tip lesion (≥50%)	0	0
extended sclerosing tip lesion (≥50%)	0	0
C) WRINKLING OF GBM	0	0
segmental deflation (<80%)	0	0
*estimate AGGREGATE percentage deflation (%)	0%	0%
segmental collapse (<80%)	0	0
*estimate AGGREGATE percentage collapse (%)	0%	0%
D) HYALINOSIS	0	0
hyalinosis at the vascular pole	0	0
hyalinosis at the tubular pole	0	0
hyalinosis away from vascular and tubular pole	0	0
hyalinosis cannot determine location	0	0
E) OTHER SEGMENTAL LESIONS	0	0
glomerular foam cells	0	0
sinechia	0	0
PERIGLOMERULAR FIBROSIS	0	0
PODOCYTE INJURY	0	0
A) Hypertrophy	0	0
segmental epithelial cell (podocyte) hypertrophy (<50%)	0	0
global epithelial cell (podocyte) hypertrophy (≥50%)	0	0
* estimate AGGREGATE percentage hypertrophy (%)	0%	0%
B) Hyperplasia	0	0
segmental epithelial cell (podocyte) hyperplasia (<50%)	0	0
global epithelial cell (podocyte) hyperplasia (≥50%)	0	0
* estimate AGGREGATE percentage hyperplasia (%)	0%	0%
C) hyalin droplets in epithelial cells (podocyte)	0	0
segmental hyalin droplets in epithelial cells (podocyte) (<50%)	0	0
global hyalin droplets in epithelial cells (podocyte) (≥50%)	0	0
* estimate AGGREGATE percentage hyalne droplets (%)	0%	0%
D) halo (detachment of podocytes from GBM)	0	0
MESANGIOPATHIC CHANGES	0	0
A) Increased mesangial cell cellularity	0	0
segmental mesangial hypercellularity (<50%)	0	0
global mesangial hypercellularity (≥50%)	0	0
* estimate AGGREGATE percentage mesangial hypercellularity (%)	0%	0%
B) Increased mesangial matrix	0	0
segmental increased mesangial matrix without hypercellularity (<50%	0	0
global increased mesangial matrix without hypercellularity (≥50%)	0	0
* estimate AGGREGATE percentage mesangial matrix (%)	0%	0%
GBM - SPIKES (%)	0	0
segmental spikes (<50%)	0	0
global spikes (≥50%)	0	0
* estimate AGGREGATE percentage spikes (%)	0%	0%
INFLAMMATION intracapillary leukocytes	0	0

C. Modified NDPSS2

CASE ID:		XX_XXXXX _XXX_XXX	XX_XXXXX _XXX_XXX
GLOMERULUS NUMBER:	Overall severity?	1	2
IMAGE LEVEL NUMBER:		L#?	L#?
STAIN TYPE:		Stain?	Stain?
Did this image have poor quality? If yes (1), please comment.		0	0
NO/MINIMAL CHANGES		0	0
ANY SCLEROSIS, WRINKLING, OR TIP		0	0
A) ANY SCLEROSIS OR TIP (%)		0%	0%
B) ANY WRINKLING (%)		0%	0%
any deflation (%)		0%	0%
any collapse (%)		0%	0%
GLOBAL OBLITERATION		0	0
A) GLOBAL SCLEROSIS (100%)		0	0
B) GLOBAL WRINKLING OF GBM (≥80%)		0	0
C) GLOBAL MESANGIAL SCLEROSIS (≥80%)		0	0
SEGMENTAL OBLITERATION/LESION		0	0
A) SEGMENTAL SCLEROSIS (<100%)		0	0
B) TIP LESION		0	0
C) WRINKLING OF GBM (<80%)		0	0
D) HYALINOSIS		0	0
E) SEGMENTAL LESIONS		0	0
glomerular foam cells		0	0
sinechia		0	0
PERIGLOMERULAR FIBROSIS		0	0
PODOCYTE INJURY		0	0
A) Any Hypertrophy (%)		0%	0%
B) Any Hyperplasia (%)		0%	0%
C) Hyalin droplets in epithelial cells (podocyte)		0	0
D) Halo (detachment of podocytes from GBM)		0	0
MESANGIOPATHIC CHANGES		0	0
A) Any Increased mesangial cell cellularity (%)		0%	0%
B) Any Increased mesangial matrix (%)		0%	0%
GBM - SPIKES		0	0
INFLAMMATION intracapillary leukocytes		0	0

Figure 1.

Nephrotic Syndrome Study Network Digital Pathology Scoring System (NDPSS) scoring matrices used in the original NDPSS (A), first modification (NDPSS1) (B), and second modification (NDPSS2) (C). Each descriptor was scored by using a drop-down menu that appeared when the appropriate cell was clicked. Abbreviations: GBM, glomerular basement membrane; glom, glomerulus; L#, image level number; WSI, whole slide imaging.

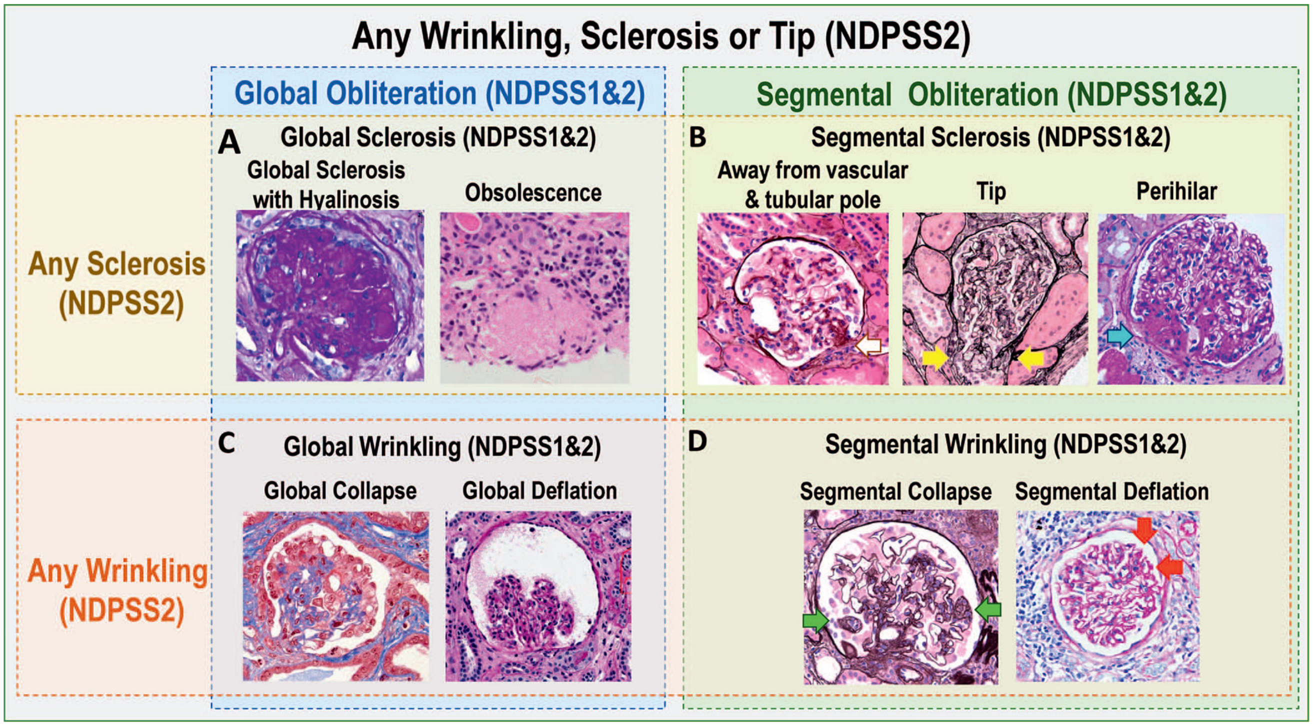

Any Wrinkling, Sclerosis or Tip (NDPSS2)
Global Obliteration (NDPSS1&2)
Segmental Obliteration (NDPSS1&2)
Any Sclerosis (NDPSS2)
A
Global Sclerosis (NDPSS1&2)
Global Sclerosis with Hyalinosis
Obsolescence
B
Segmental Sclerosis (NDPSS1&2)
Away from vascular & tubular pole
Tip
Perihilar
Any Wrinkling (NDPSS2)
C
Global Wrinkling (NDPSS1&2)
Global Collapse
Global Deflation
D
Segmental Wrinkling (NDPSS1&2)
Segmental Collapse
Segmental Deflation

Figure 2.

Example of classes and subclasses of descriptors (images) and how they are organized in the modified Nephrotic Syndrome Study Network Digital Pathology Scoring System (NDPSS), NDPSS1, and NDPSS2. The NDPSS2 class any sclerosis, wrinkling, or tip includes the subclasses any sclerosis (A and B) and any wrinkling (C and D). The NDPSS2 subclass any sclerosis contains additional subclasses global sclerosis (A) and segmental sclerosis (C); the NDPSS2 subclass any wrinkling contains additional subclasses global wrinkling (C) and segmental wrinkling (D). The NDPSS1 and 2 class global obliteration includes the subclasses global sclerosis (A) and global wrinkling (C); the NDPSS1 and 2 class segmental obliteration includes the subclasses segmental sclerosis (C) and segmental wrinkling (D). Examples of descriptors in the various classes and subclasses: A, The descriptors global sclerosis with hyalinosis (periodic acid–Schiff) and obsolescence (hematoxylin-eosin) are grouped in the NDPSS1 and 2 subclass global sclerosis, the NDPSS2 subclass any sclerosis, and the NDPSS1 and 2 class global obliteration. B, The descriptors segmental sclerosis away from vascular and tubular pole (silver stain), tip lesion (silver stain; yellow arrows), and segmental perihilar sclerosis (periodic acid–Schiff; blue arrow) are grouped in the NDPSS1 and 2 subclass segmental sclerosis, the NDPSS2 subclass any sclerosis, and the NDPSS1 and 2 class segmental obliteration. C, The descriptors global collapse (trichrome) and global deflation (silver stain) are grouped in the NDPSS1 and 2 subclass global wrinkling, the NDPSS2 subclass any wrinkling, and the NDPSS1 and 2 class global obliteration. D, The descriptors segmental collapse (silver stain; green arrows) and segmental deflation (periodic acid–Schiff; red arrows) are grouped in the NDPSS1 and 2 subclass segmental wrinkling, the NDPSS2 subclass any wrinkling, and the NDPSS1 and 2 class segmental obliteration. The descriptors global collapse and segmental collapse are also grouped in the NDPSS2 any collapse, and the descriptors global deflation and segmental deflation are also grouped in the NDPSS2 subclass any deflation (not shown in figure) (original magnifications ×60 [A, global sclerosis with hyalinosis] and ×40 [A, obsolescence, and B through D]).

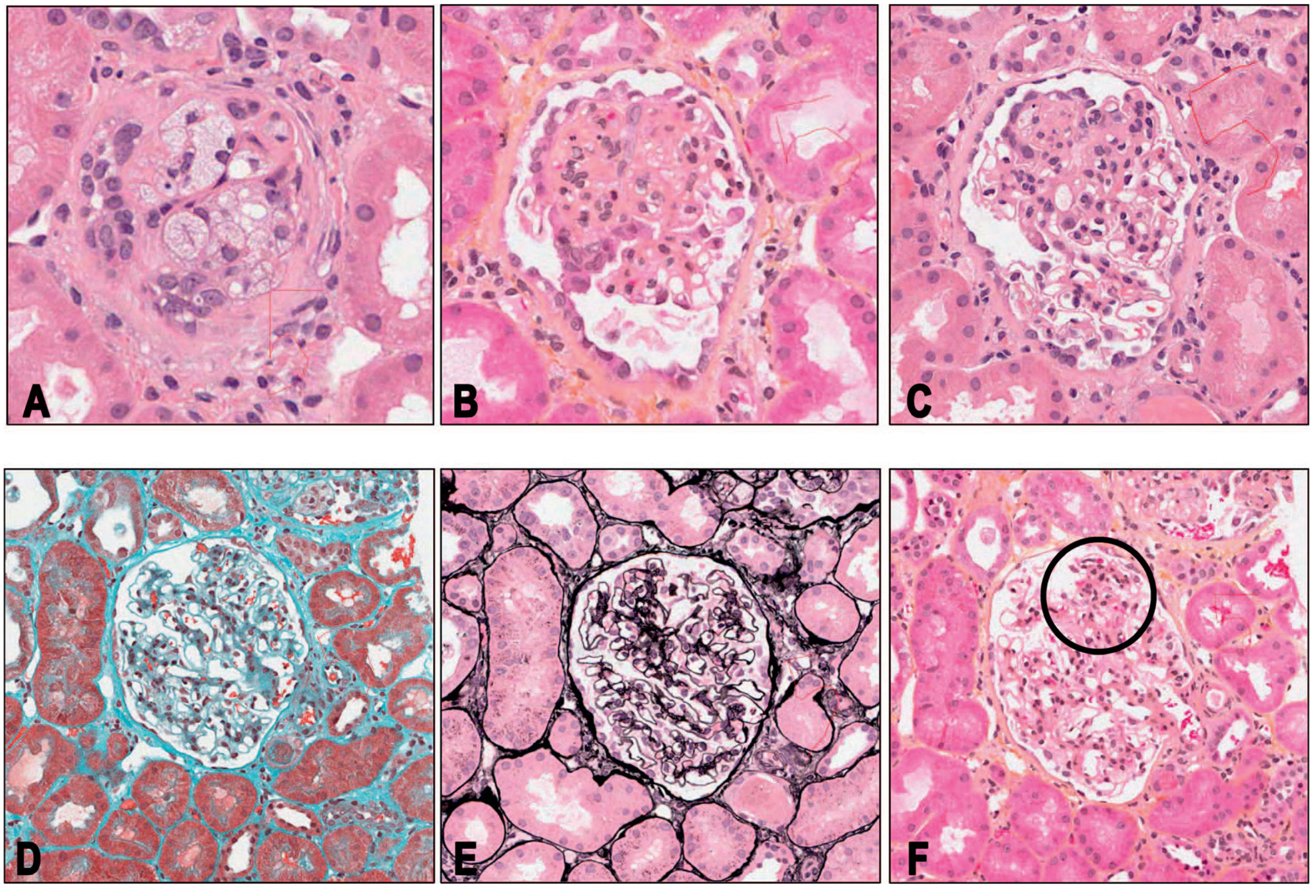


Figure 3.

Multilevel representation of a single glomerulus showing different descriptors in different levels. A, Level 2, intraglomerular foam cells. B, Level 5, an example of segmental obliteration involving at least 75% of the glomerular tuft, with foam cells and segmental podocyte hypertrophy and hyperplasia. C, Level 7; here the segmental obliteration involves less than 50% of the glomerular tuft. Other descriptors present in this section are foam cells and segmental podocyte hypertrophy. D, Level 10, no/minimal changes. E, Level 11, no/minimal changes. F, Level 12, segmental mesangial proliferation (circled) (hematoxylin-eosin, original magnification ×40 [A through C and F]; trichrome, original magnification ×40 [D]; silver, original magnification ×40 [E]).

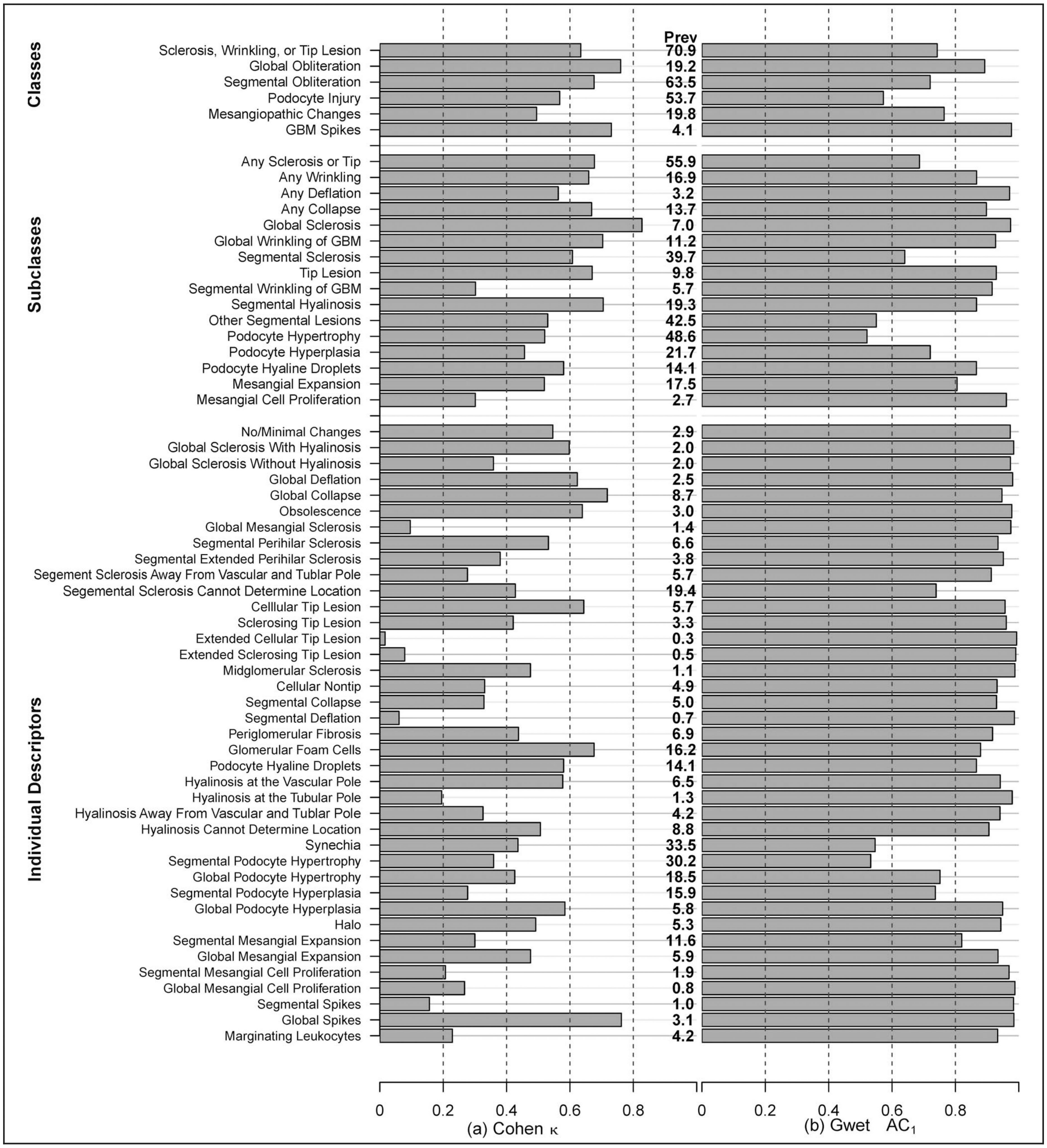


Figure 4.

Interrater agreement on original Nephrotic Syndrome Study Network Digital Pathology Scoring System descriptors by classes, subclasses, and individual descriptors. Agreement was assessed by the Cohen κ (a) and the Gwet agreement coefficient 1 (AC_1) (b). Prevalence (Prev) of each descriptor is listed to aid interpretation. Abbreviation: GBM, glomerular basement membrane.

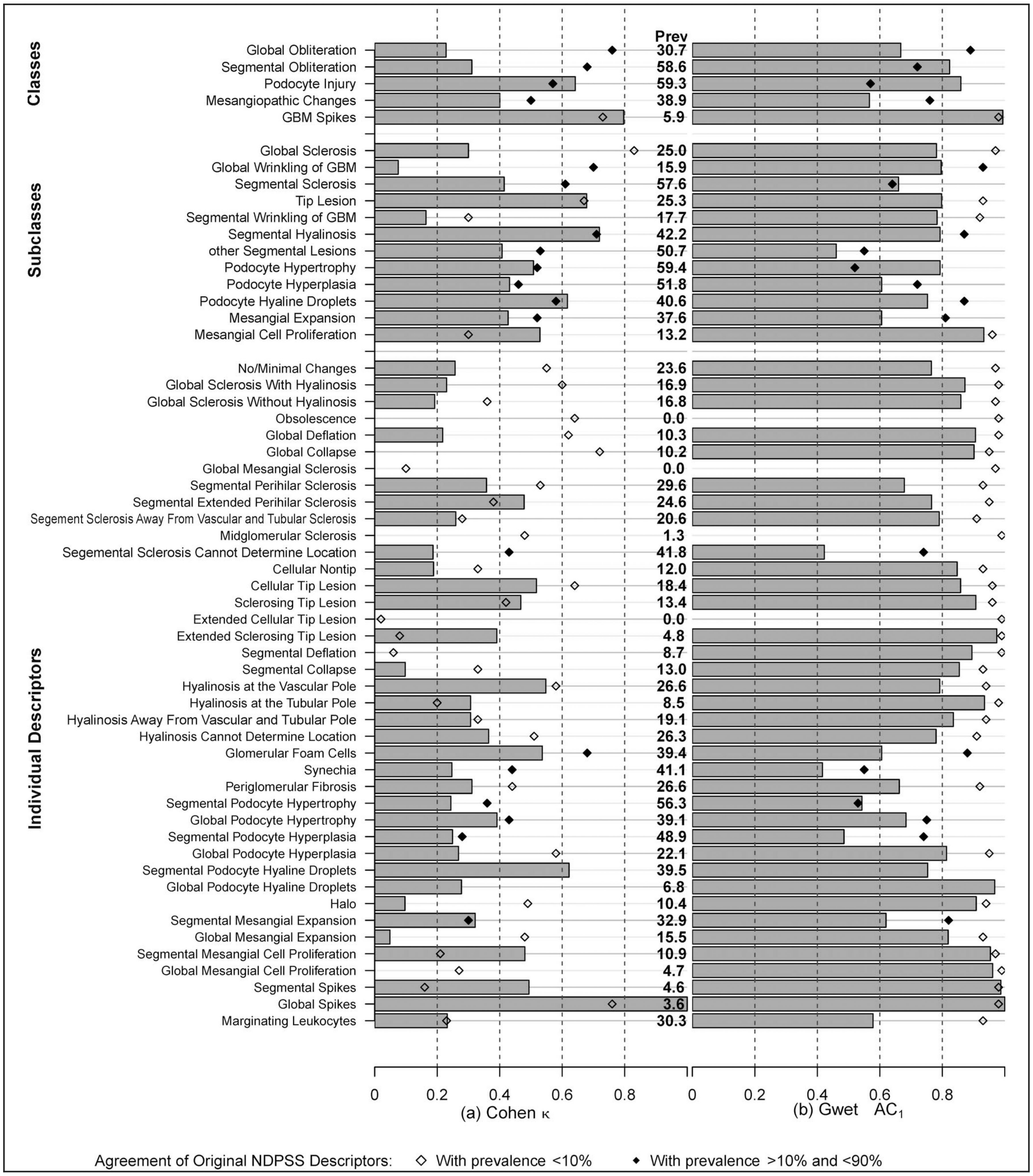


Figure 5.

Interrater agreement on descriptors from first modification of the Nephrotic Syndrome Study Network Digital Pathology Scoring System (NDPSS1) by classes, subclasses, and individual descriptors. Agreement was assessed by the Cohen κ (a) and the Gwet agreement coefficient 1 (AC_1) (b). Prevalence (Prev) of each descriptor from NDPSS1 is listed to aid interpretation. Agreement estimates from the original NDPSS are plotted for comparison for those with original prevalence less than 10% (open diamonds) and those with original prevalence between 10% and 90% (filled diamonds). Abbreviation: GBM, glomerular basement membrane.

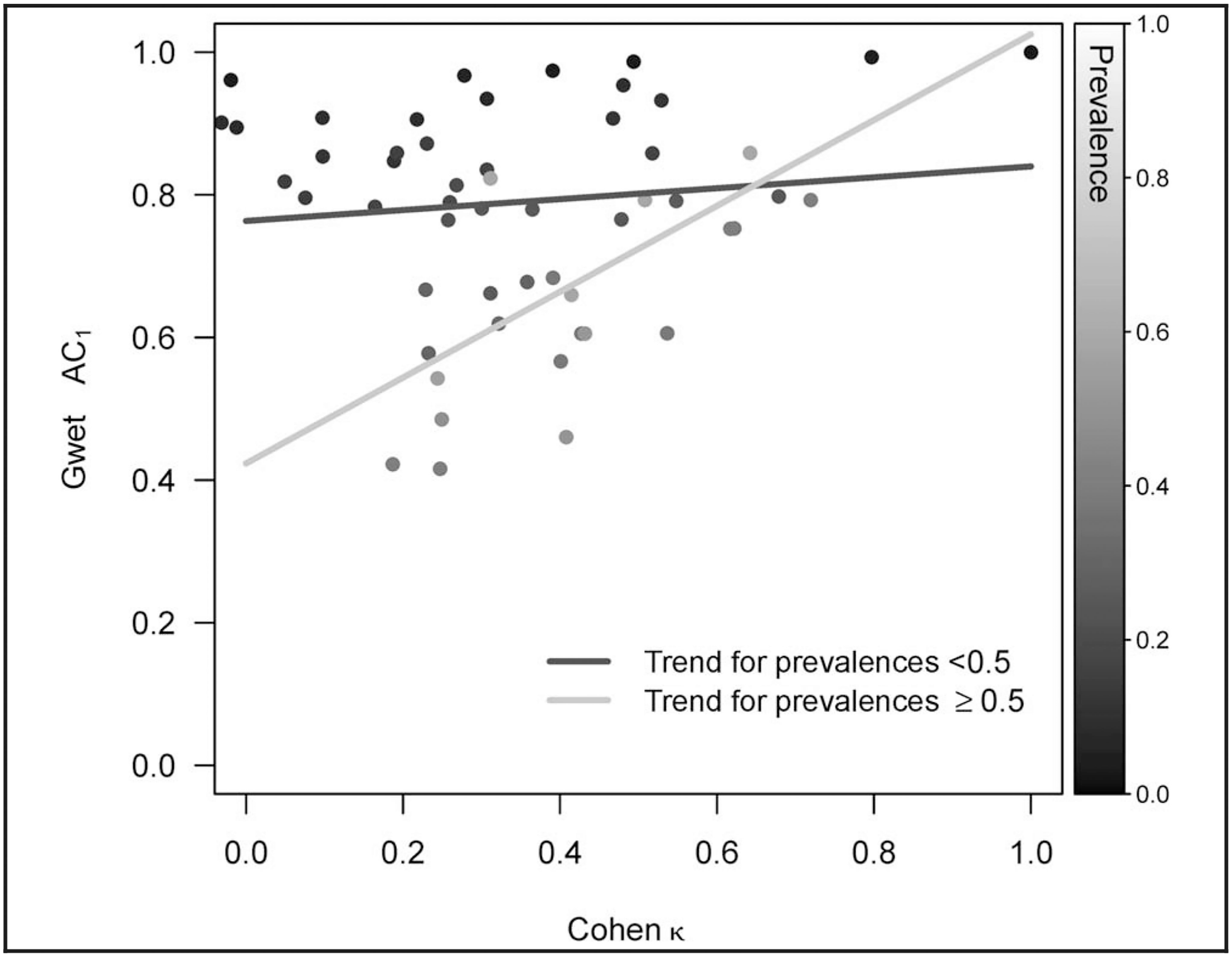


Figure 6.

Comparison between Gwet agreement coefficient 1 (AC_1) and Cohen κ estimates of interrater agreement of classes, subclasses, and individual descriptors from the first modification of the Nephrotic Syndrome Study Network Digital Pathology Scoring System. Prevalence of each descriptor is indicated by shades of gray and trend lines for high-prevalence (≥50%) and low-prevalence (<50%) descriptors are shown.

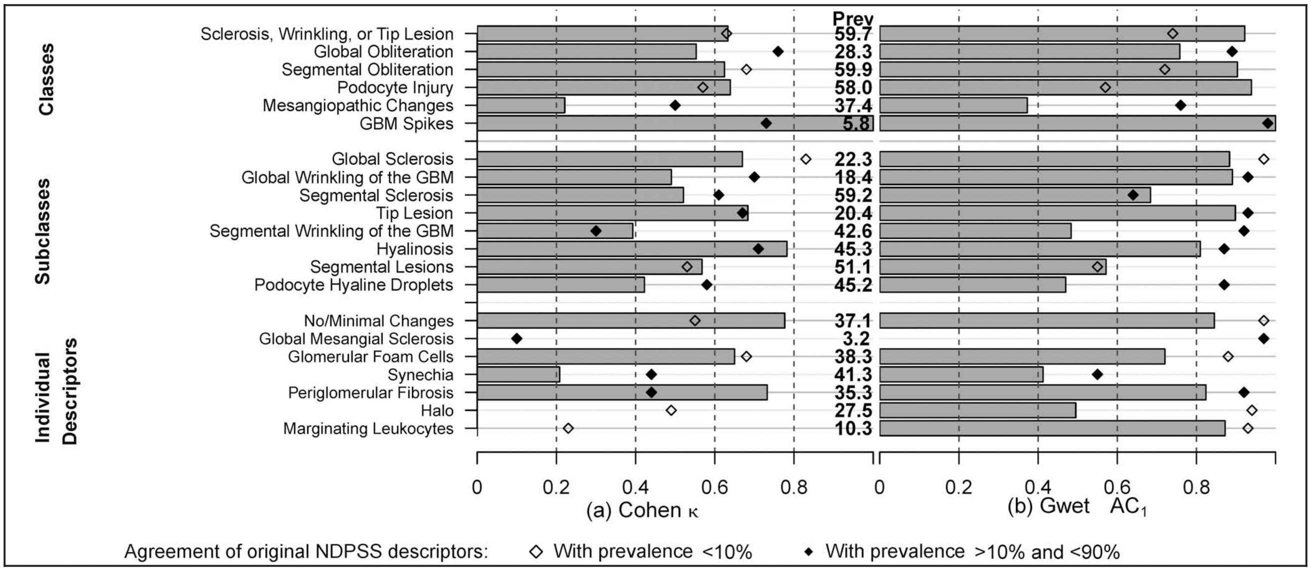


Figure 7.

Interrater agreement on descriptors from second modification of the Nephrotic Syndrome Study Network Digital Pathology Scoring System (NDPSS2) by classes, subclasses, and individual descriptors. Agreement was assessed by the Cohen κ (a) and the Gwet agreement coefficient 1 (AC_1) (b). Prevalence (Prev) of each descriptor in NDPSS2 is listed to aid interpretation. Agreement estimates from the original NDPSS are plotted for comparison for those with original prevalence less than 10% (open diamonds) and those with original prevalence between 10% and 90% (filled diamonds). Abbreviation: GBM, glomerular basement membrane.

Table 1

Groupings of Individual Descriptors Into Classes Included in Each Set of Scoring Strategies

	Classes					
	Any Sclerosis, Wrinkling, or Tip	**Global Obliteration**	**Segmental Obliteration**	**Podocyte Injury**	**Mesangiopathic Changes**	**GBM Spikes**
Individual descriptor						
No/minimal changes						
Global sclerosis with hyalinosis	X	X				
Global sclerosis without hyalinosis	X	X				
Global collapse	X	X				
Global deflation	X	X				
Obsolescence	X	X				
Global mesangial sclerosis		X				
Segmental perihilar sclerosis	X		X			
Segmental extended perihilar sclerosis	X		X			
Segmental sclerosis away from vascular and tubular pole	X		X			
Segmental sclerosis cannot determine location	X		X			
Cellular tip lesion	X		X			
Sclerosing tip lesion	X		X			
Extended cellular tip lesion	X		X			
Extended sclerosing tip lesion	X		X			
Midglomerular sclerosis			X			
Cellular nontip	X		X			
Segmental collapse	X		X			
Segmental deflation	X		X			
Periglomerular fibrosis						
Glomerular foam cells			X			
Segmental podocyte hyaline droplets[a]				X		
Global podocyte hyaline droplets[a]				X		
Hyalinosis at the vascular pole			X			
Hyalinosis at the tubular pole			X			

	Classes					
	Any Sclerosis, Wrinkling, or Tip	**Global Obliteration**	**Segmental Obliteration**	**Podocyte Injury**	**Mesangiopathic Changes**	**GBM Spikes**
Hyalinosis away from vascular and tubular pole			X			
Hyalinosis cannot determine location			X			
Synechia			X			
Segmental podocyte hypertrophy				X		
Global podocyte hypertrophy				X		
Segmental podocyte hyperplasia				X		
Global podocyte hyperplasia				X		
Halo				X		
Segmental mesangial expansion					X	
Global mesangial expansion					X	
Segmental mesangial cell proliferation					X	
Global mesangial cell proliferation					X	
Segmental spikes						X
Global spikes						X
Marginating leukocytes						
Set of Scoring Strategies						
Original NDPSS[b]	D	D	D	D	D	D
NDPSS1		P	P	P	P	P%
NDPSS2	D	D	D	D	D	D

Abbreviations: D, dichotomous; GBM, glomerular basement membrane; NDPSS, Nephrotic Syndrome Study Network Digital Pathology Scoring System; P, probability; %, percentage.

[a] Podocyte hyaline droplets was originally scored as a single descriptor and split into 2 individual descriptors (global versus segmental) for the NDPSS1 only.

[b] Postscoring grouping into classes.

Table 2

Groupings of Individual Descriptors Into Subclasses Included in Each Set of Scoring Strategies

	Subclasses					
	Any Sclerosis or Tip	**Any Wrinkling**	**Any Deflation**	**Any Collapse**	**Global Sclerosis**	**Global Wrinkling**
Individual descriptor						
No/minimal changes						
Global sclerosis with hyalinosis	X				X	
Global sclerosis without hyalinosis	X				X	
Global collapse		X		X		X
Global deflation		X	X			X
Obsolescence	X				X	
Global mesangial sclerosis						
Segmental perihilar sclerosis	X					
Segmental extended perihilar sclerosis	X					
Segmental sclerosis away from vascular and tubular pole	X					
Segmental sclerosis cannot determine location	X					
Cellular tip lesion	X					
Sclerosing tip lesion	X					
Extended cellular tip lesion	X					
Extended sclerosing tip lesion	X					
Midglomerular sclerosis	X					
Cellular nontip	X					
Segmental collapse		X		X		
Segmental deflation		X	X			
Periglomerular fibrosis						
Glomerular foam cells						
Segmental podocyte hyaline droplets[a]						
Global podocyte hyaline droplets[a]						
Hyalinosis at the vascular pole						
Hyalinosis at the tubular pole						

	Subclasses					
	Any Sclerosis or Tip	Any Wrinkling	Any Deflation	Any Collapse	Global Sclerosis	Global Wrinkling
Hyalinosis away from vascular and tubular pole						
Hyalinosis cannot determine location						
Synechia						
Segmental podocyte hypertrophy						
Global podocyte hypertrophy						
Segmental podocyte hyperplasia						
Global podocyte hyperplasia						
Halo						
Segmental mesangial expansion						
Global mesangial expansion						
Segmental mesangial cell proliferation						
Global mesangial cell proliferation						
Segmental spikes						
Global spikes						
Marginating leukocytes						
Set of Scoring Strategies						
Original NDPSS[b]	D	D	D	D	D	D
NDPSS1			%	%	P	P
NDPSS2	%	%	%	%	D	D

Subclasses									
Segmental Sclerosis	Segmental Wrinkling	Tip Lesions	Segmental Hyalinosis	Other Segmental Lesions	Mesangial Expansion	Mesangial Cell Proliferation	Podocyte Hypertrophy	Podocyte Hyperplasia	Podocyte Hyaline Droplets[a]
X									

Subclasses									
Segmental Sclerosis	Segmental Wrinkling	Tip Lesions	Segmental Hyalinosis	Other Segmental Lesions	Mesangial Expansion	Mesangial Cell Proliferation	Podocyte Hypertrophy	Podocyte Hyperplasia	Podocyte Hyaline Droplets[a]
X									
X									
X									
		X							
		X							
		X							
		X							
X									
X									
	X								
	X								
				X					
									X
									X
			X						
			X						
			X						
			X						
				X					
							X		
							X		
								X	
					X			X	
					X	X			
						X			
D	D	D	D	D	D	D	D	D	D
P	P	P	P	P	P%	P%	P%	P%	P%
D	D	D	D	D	%	%	%	%	D

Abbreviations: D, dichotomous; NDPSS, Nephrotic Syndrome Study Network Digital Pathology Scoring System; P, probability; %, percentage.

[a] Podocyte hyaline droplets was originally scored as a single descriptor and split into 2 individual descriptors (global versus segmental) for the NDPSS1 only.

[b] Postscoring grouping into subclasses.

Table 3

Scoring Strategies Tested in First (Nephrotic Syndrome Study Network Digital Pathology Scoring System [NDPSS] 1) and Second (NDPSS2) Modifications of the NDPSS

Purpose	**Scoring Strategy**
NDPSS1	
To evaluate the agreement in descriptor scoring using all biopsy levels rather than a single image	All tuft cross sections for a given annotated glomerulus were reviewed and collectively used to generate a single descriptor score (Figure 3), ie, the presence of an individual or group of descriptors was recorded if it appeared in one or more tuft cross sections. Although this strategy is part of the NDPP and NDPSS, our previously published study tested agreement using individual JPEG images only. One of the 39 individual descriptors from the original NDPSS was split into 2 (segmental versus global) for NDPSS1, so NDPSS1 included 40 individual descriptors.
To test if grouping descriptors with common characteristics prior to scoring improves agreement	40 individual glomerular descriptors relevant to MCD, FSGS, and MN were organized into 5 classes and 12 subclasses (Tables 1 and 2; Figure 2). In contrast to the previously published study where grouping was performed after the scoring process, pathologists directly scored classes, subclasses, and individual descriptors in a hierarchical fashion. Each class or subclass was endorsed if any one of the component descriptors was present.
To identify the descriptors for which pathologists had some uncertainty and to test whether scoring on an ordinal scale would improve agreement	Pathologists indicated their confidence in scoring the presence of any given class, subclass, or individual descriptor as a probability (0 = no, 0.25 = probably not, 0.50 = maybe, 0.75 = probably yes, or 1 = yes).
To test whether scoring on a continuous measure improves agreement compared with a dichotomous approach	The percentage of the glomerular tuft involved (0%, 5%, 10%, 20%, …, 90%, 100%) was indicated for 8 classes or subclasses of descriptors (Tables 1 and 2).
NDPSS 2	
To test whether reproducibility was modulated by having pathologists focus on a single glomerular level at a time, independently from descriptors present in other levels	Biopsy sections/levels containing each annotated glomerulus were individually scored using separate columns in the NDPSS2 scoring matrix. These section/level-specific scores were later combined to obtain a glomerulus-specific score, such that presence in any section implies presence in the glomerulus (Figure 3).
To test whether reproducibility could be increased by grouping descriptors in different ways than previously done	Descriptors were reorganized into 6 classes and 16 subclasses (Tables 1 and 2; Figure 2). Only 7 individual descriptors were included in NDPSS2 for scoring.
To test whether scoring on a continuous measure improves agreement compared with a dichotomous approach	In 8 of 16 subclasses, the score was recorded as a percentage of the glomerular tuft involved (Tables 1 and 2). For the remaining 8 subclasses, 6 classes, and 7 individual descriptors, dichotomous metrics (ie, present versus absent) were used for scoring.
To test reproducibility of each pathologist's subjective interpretation of the overall severity of damage in the biopsy	Pathologists were asked to indicate a gestalt overall damage score (from 1 = good prognosis to 5 = really bad prognosis). No cross-training was provided for this measure.
To evaluate whether removal of poor quality images or stratification by stain type affected reproducibility results	Pathologists indicated the stain type for each biopsy section and whether there were any images with poor quality.

Abbreviations: FSGS, focal segmental glomerulosclerosis; MCD, minimal change disease; MN, membranous nephropathy; NDPP, Nephrotic Syndrome Study Network Digital Pathology Protocol; NDPSS, Nephrotic Syndrome Study Network Digital Pathology Scoring System.