The Journal of Pathology: Clinical Research
J Pathol Clin Res September 2024; **10:** e12395
Published online in Wiley Online Library
(wileyonlinelibrary.com). DOI: 10.1002/2056-4538.12395

ORIGINAL ARTICLE

Validation of a whole slide image management system for metabolic-associated steatohepatitis for clinical trials

Hanna Pulaski[1], Shraddha S Mehta[1], Laryssa C Manigat[1], Stephanie Kaufman[1], Hypatia Hou[1], ILKe Nalbantoglu[2], Xuchen Zhang[2], Emily Curl[3], Ross Taliano[4], Tae Hun Kim[5], Michael Torbenson[6], Jonathan N Glickman[1†], Murray B Resnick[1‡], Neel Patel[1], Cristin E Taylor[1], Pierre Bedossa[7], Michael C Montalto[1], Andrew H Beck[1] and Katy E Wack[1]*

[1]*PathAI, Inc, Boston, MA, USA*
[2]*Department of Pathology, Yale University School of Medicine, New Haven, CT, USA*
[3]*Foundation Medicine, Inc, Boston, MA, USA*
[4]*Department of Pathology, Rhode Island Hospital, Providence, RI, USA*
[5]*Pathology Medical Group of Riverside, Providence, RI, USA*
[6]*Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA*
[7]*Liverpat and University of Paris, Paris, France*

**Correspondence to: Katy E Wack, PathAI, Inc, 1325 Boylston St, Suite 10000, Boston, MA 02215, USA. E-mail: katy.wack@pathai.com*

[†]*Present address:* Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
[‡]*Present address:* Department of Pathology and Laboratory Medicine, Rhode Island Hospital, Brown University, Providence, RI, USA

Abstract

The gold standard for enrollment and endpoint assessment in metabolic dysfunction-associated steatosis clinical trials is histologic assessment of a liver biopsy performed on glass slides. However, obtaining the evaluations from several expert pathologists on glass is challenging, as shipping the slides around the country or around the world is time-consuming and comes with the hazards of slide breakage. This study demonstrated that pathologic assessment of disease activity in steatohepatitis, performed using digital images on the AISight whole slide image management system, yields results that are comparable to those obtained using glass slides. The accuracy of scoring for steatohepatitis (nonalcoholic fatty liver disease activity score ≥4 with ≥1 for each feature and absence of atypical features suggestive of other liver disease) performed on the system was evaluated against scoring conducted on glass slides. Both methods were assessed for overall percent agreement with a consensus "ground truth" score (defined as the median score of a panel of three pathologists' glass slides). Each case was also read by three different pathologists, once on glass and once digitally with a minimum 2-week washout period between the modalities. It was demonstrated that the average agreement across three pathologists of digital scoring with ground truth was noninferior to the average agreement of glass scoring with ground truth [noninferiority margin: −0.05; difference: −0.001; 95% CI: (−0.027, 0.026); and $p < 0.0001$]. For each pathologist, there was a similar average agreement of digital and glass reads with glass ground truth (pathologist A, 0.843 and 0.849; pathologist B, 0.633 and 0.605; and pathologist C, 0.755 and 0.780). Here, we demonstrate that the accuracy of digital reads for steatohepatitis using digital images is equivalent to glass reads in the context of a clinical trial for scoring using the Clinical Research Network scoring system.

Keywords: digital pathology; nonalcoholic steatohepatitis; NASH; metabolic-associated steatohepatitis; MASH; clinical trials; validation

Received 1 September 2023; Revised 4 July 2024; Accepted 15 July 2024

Conflict of interest statement: Hanna Pulaski, Shraddha S Mehta, Hypatia Hou, Jonathan Glickman, Murray Resnick, Neel Patel, Michael C Montalto, Andrew H Beck, and Katy Wack are employees of PathAI, Inc. Laryssa C Manigat, Stephanie Kaufman, and Cristin E Taylor are former employees of PathAI, Inc. ILKe Nalbantoglu, Xuchen Zhang, Emily Curl, Ross Taliano, Tae Hun Kim, and Michael Torbenson are contractors of PathAI, Inc. Pierre Bedossa has no conflict.

Introduction

Metabolic dysfunction-associated steatotic liver disease (MASLD; formerly referred to as nonalcoholic fatty liver disease) is rising in prevalence globally, with an estimated 25% of the world's population affected [1]. Due to this increased burden of disease, liver decompensation due to the progression of metabolic-associated

steatohepatitis (MASH; formerly referred to as nonalcoholic steatohepatitis or NASH) is the leading cause of liver transplant in women [2] and expected to become the overall leading cause of liver transplant [3]. There are currently no approved therapies for MASH, and there is a large unmet need for clinical intervention in this patient population. Many clinical trials are ongoing to identify therapies for MASH, with composite score-based changes in histologic features, as the primary endpoint for accelerated or conditional approval. The MASH Clinical Research Network (CRN) scoring system [4] is utilized for enrollment and composite score primary endpoint criteria recommended by the European Medicine Agency and US Food and Drug Administration (FDA) [5–7] but is subject to high inter- and intrareader variability. Lack of standardization, consistency, and bias present during enrollment and follow-up timepoints may cause potentially effective therapies to fail in phase 2b or 3 trials. Due to the significant variation between expert pathologists, regulatory bodies are recommending multiple, or consensus, reads to reduce individual bias and increase quality and consistency [8–10]. However, obtaining the evaluations from several expert pathologists for the same participant on glass slides to meet enrollment windows and during follow-up time point reads is challenging, as shipping the slides around the country or in some cases around the world is time-consuming and comes with the hazards of slide breakage during the shipment.

Currently, the practice of pathology is increasingly adopting and incorporating digital pathology into clinical workflows. Numerous studies have shown high accuracy in providing primary diagnoses, on the categorical level, using whole slide images (WSIs) of glass slides [11–21]. However, although the current gold standard to establish the diagnosis of MASH is histopathologic analysis of a liver biopsy, no studies have been published to our knowledge to validate WSI for MASH diagnosis. Importantly, no studies evaluate the accuracy of histologic component scoring and specific score-based enrollment criteria for the clinical trial context of use, which moves beyond categorical diagnosis. Demonstration of equivalency to glass reads for MASH clinical trials has also been a request from the FDA [22]. The current gold standard to establish the diagnosis of MASH is histopathologic analysis of a liver biopsy. The diagnosis is established by the presence of a characteristic histologic pattern of ≥5% hepatic steatosis, lobular inflammation, and hepatocellular ballooning in the appropriate clinical setting and provided other potential cases of metabolic-associated liver disease such as significant alcohol intake have been excluded. In 1999, a semiquantitative grading and staging system to describe and unify the approach of pathologists to the histopathologic lesions of MASH was proposed by Brunt *et al* [23] and the MASH CRN. A semiquantitative activity grade (MASLD activity score or MAS) was assigned by a combination of parameters including steatosis, lobular inflammation, and hepatocyte ballooning. MAS ≥4 is used in phase 2b and 3 clinical trials as a definition of steatohepatitis for enrollment criteria [24–29]. Therefore, because the WSI image management system (IMS) is utilized in the context of clinical trials, we chose to define steatohepatitis as MAS ≥4 with a score of ≥1 for each feature and absence of atypical features suggestive of other liver disease, and we evaluated the accuracy of digital reads compared to glass slide reads by individual study readers compared to an independent consensus ground truth (GT). Importantly, as a secondary analysis, we determined linearly weighted kappa concordance for intrareader, intermodality assessment of steatosis, lobular inflammation, hepatocellular ballooning, fibrosis, and overall MAS. Exploratory analysis compared the score-based accuracy of individually assessing steatosis, hepatocellular ballooning, lobular inflammation, and fibrosis for expert pathologist readers evaluating on glass and digital compared to an independent consensus GT. Per this context of use, a study population was chosen which represents MASH trial screening and enrolled populations from multiple, completed trials as well as commercially available clinical non-MASH samples. This population was enriched for borderline MASLD/steatohepatitis cases, along with non-MASH cases, in order to thoroughly assess the ability to use digital pathology as a surrogate for glass reads in MASH trials. Importantly, College of American Pathologists (CAP) recommendations for validation design were considered and incorporated [30] but a larger sample size was chosen to adequately power the study for the score-based endpoint criteria and multiple study pathologists were utilized, to account for the known intra- and interrater variability. Additionally, although a noninferiority design was chosen to assess the accuracy of the specific score-based inclusion criteria in MASH trials, similar to many of the study designs utilized for WSI image validation in FDA submissions for primary diagnosis [31,32], secondary and exploratory endpoints evaluate other measures of accuracy on the individual histologic component score level, as is crucial for the MASH trial context of use, none of which has been described in the literature thus far.

Materials and methods

Ethics

De-identified liver cases were obtained from a third-party vendor and from completed clinical trials from PathAI partners where proper ethical approval(s) were obtained at the time of tissue collection/storage. Additional consent was not obtained as the samples were de-identified to this study, no protected health information (PHI) was received by PathAI, and the research involved no more than minimal risk to subjects. This study received a waiver of consent and expedited approval from WCG™ Institutional Review Board (IRB00000533).

WSI image management system

The AISight Clinical Trials platform (v3.3.1) is a Good Clinical Practice (GCP) compliant research use-only cloud-based WSI IMS that serves as an interface for viewing WSIs (Figure 1). System configurability allows for maximum flexibility in leveraging digital pathology to improve subject outcomes in clinical research. All pathologists performing digital reads in this study were trained on the use of the system and completed practice cases prior to the study start.

The WSI IMS is specifically designed for use in clinical trials, all samples (including slide labels) are de-identified by the trial Sponsor prior to sharing the glass slides or images. No PHI is shared through the WSI IMS. Each Sponsor has their own setup in the system and can only access their own samples and data.

Digital pathology IMS image handling

All WSIs are uploaded to the IMS and tiled image pyramids in DZI format are generated. The tiles are JPGs generated at 75% quality. The IMS can also do color calibration to the viewer tiles by applying the WSI's International Color Consortium profile if it exists. The color calibration is a customer configuration setting and can be turned on or off depending on the clinical trial needs.

Case selection and scanning

Existing de-identified glass slides from a third-party vendor and from completed clinical trials (screen failures and enrolled cases) were utilized in this study.

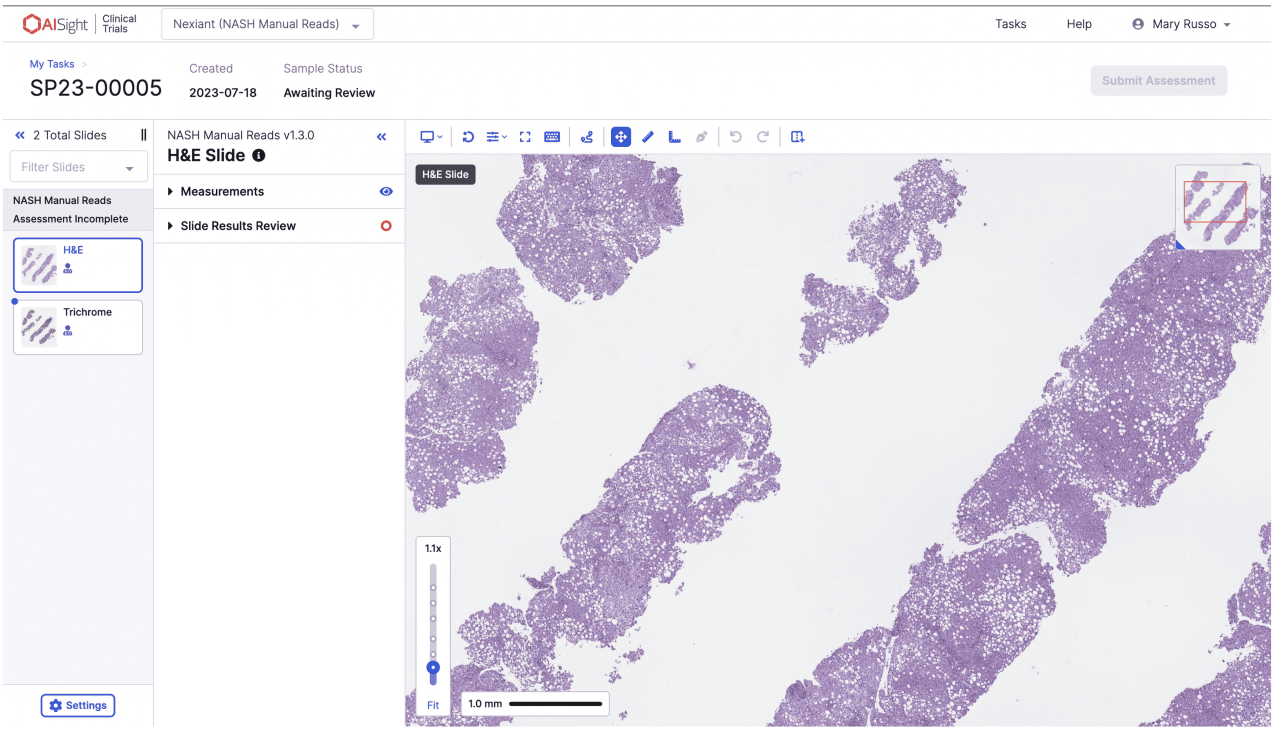

Figure 1. The WSI IMS user interface. In the WSI IMS, the pathologist has access to the cases (each case consisting of an H&E and Masson's trichrome) where they can review the slides, move around the slide, zoom in and out, and finally enter their scores.

Each case in this study had two slides – H&E and Masson's trichrome. Slides were first scanned at a single CAP-accredited, Clinical Laboratory Improvement Amendments-certified lab on a Leica Aperio AT2 scanner (scanner outputs are .svs files) at ×40 magnification, uploaded to WSI IMS via Amazon Web Services S3 bucket after image quality control and then distributed for glass reads.

The slide set consisted of 160 cases from liver needle biopsies. Two-thirds of the cases were chosen from patients with steatohepatitis (defined as MAS ≥4 with a score of ≥1 for each feature and absence of atypical features suggestive of other liver disease) based on the original trial central pathology scores, and the remaining one of three was from MASLD and other liver disorders encountered during clinical trial screening and follow-up timepoints (for inclusion of atypical features suggestive of other liver diseases). Based on the original trial individual central pathologist scores, 5–10% of the 160 cases were chosen to be diagnostically challenging or borderline steatohepatitis. This borderline category was defined as MAS ≥4 with a score of 0 for at least one of the histologic features (steatosis, lobular inflammation, and hepatocellular ballooning), MAS = 4 with a score of ≥1 for each of the features or MAS = 3. For glass reads, the 160 cases were split into three batches and the pathologists read one batch at a time (all WSIs read first, and glass slides read after a minimum of 2-week washout).

Pathologists' reads

Overall, six board-certified pathologists who have demonstrated proficiency in reading steatohepatitis cases, have liver subspecialty experience and sign-out MASH cases in their clinical practice participated in this study (three for GT and three for study glass and WSI reads). All pathologists were trained on the study protocol prior to the study start.

The GT reads were collected on glass slides using a light microscope by a group of three pathologists. Each of these pathologists read 160 cases on glass once.

A different set of three pathologists performed the study reads. They read all cases twice, first utilizing WSIs on the WSI IMS and after a 2-week washout, on glass slides with light microscopy (Figure 2).

Statistics and bioinformatics

The primary endpoint was to evaluate for noninferior overall percent agreement (OPA) of individual pathologist's steatohepatitis evaluation (defined as MAS ≥4 with a score of ≥1 for each feature and absence of atypical features suggestive of other liver disease) on the WSI with glass GT compared to the OPA of their glass reads with glass GT with a 0.05 noninferiority margin. Bootstrap 95% confidence intervals and p values were also computed. The 0.05 noninferiority margin utilized in this study was chosen based on those WSI platform validations for primary diagnosis, where a noninferiority margin of 0.04 is often used [12,13]. Although inter- and intrareader variability is higher in a complex condition like steatohepatitis, a challenging noninferiority margin of 0.05 was chosen as the primary endpoint.

The glass GT consensus score was determined as the mode if at least two of three pathologists were in agreement. If there was no agreement, the GT was considered to be the median of all three scores. Additionally, the majority of the three GT pathologists' responses were used to assess the presence of atypical features.

The GT median scores were computed using the following method:

- For scores
  - If the median for the scores was an integer, that was the final score.

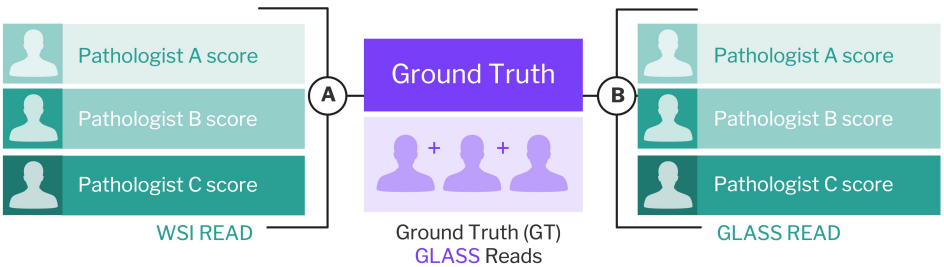

Figure 2. Study design. The study design is noninferior overall percent agreement (OPA) of individual pathologist's steatohepatitis evaluation (defined as MAS > 4 with a score of >1 for each feature and absence of atypical features suggestive of other liver disease) on the WSI with glass GT as compared to the OPA of their glass reads with glass GT with a 0.05 noninferiority margin.

- If the median for the scores was not an integer, analysis was performed with rounding up a score/stage, then again rounding down a score/stage, and the average of the two was used.

- For presence of atypical features (yes/no/sample not evaluable)
  - If at least two GT pathologists agreed, that was the final answer.
  - If at least two GT pathologists did not agree, analysis was performed with the answer yes, then again with the answer as no, and the average of the two was used.

In any case where two of the three glass GT pathologists indicated either H&E or Masson's trichrome slide was not evaluable for scoring, the whole case was removed from data analysis and if possible, was replaced with a new case, which fulfilled the target inclusion criteria. Eighteen cases were deemed inadequate by the GT panel and 17 of the cases were replaced. If a study pathologist indicated that any slide (either H&E or Masson's trichrome) was not adequate for scoring, that slide was removed from data analysis for that pathologist only.

The secondary endpoint consisted of study pathologist scores for the three steatohepatitis features, CRN fibrosis, and the overall MAS score between WSI and glass read. This endpoint was evaluated as follows: Linearly weighted kappa concordance statistics between glass and WSI read for each of the pathologists (intrapathologist and intermodality), each of the four histologic features, and overall MAS score. Overall, linearly weighted kappa was computed for each feature and overall MAS score by averaging the weighted kappa for the three pathologists. Bootstrap 95% confidence intervals are provided on the overall linearly weighted kappa. These concordance estimates are compared to the published range in Table 1. These analyses are based on observed data.

The exploratory endpoint determined overall weighted kappas for steatosis, hepatocellular ballooning, lobular inflammation fibrosis, and MAS for WSI with glass GT as compared to the weighted kappas for glass reads with glass GT.

Determination of sample size

The 2022 CAP updated guidance for validating WSI systems [30] for pathology applications recommends using a sample set of at least 60 cases, based on evidence from 33 publications reviewed. Based on these 33 studies, CAP also recommends that the ideal validation study endpoint is 95% intrarater diagnostic concordance between digital and glass slides. However, they note that noninferiority design is also acceptable. Therefore, with known substantial interrater variability in MASH scoring and diagnosis, a noninferiority design was determined to be more appropriate for the MASH trial population than a direct comparison of agreement between glass and digital reads. A sample size of 160 slides was selected based on a combination of *a priori* calculations and practical factors mentioned above to provide a degree of precision around the estimates and to account for not evaluable slides, and any incidental breakage of glass slides.

Table 1. Reference kappa scores for intrareader variability

Feature	Publication	Intrareader variability (weighted kappa scores)
Steatosis	Kleiner *et al* [4]	0.83
	Gawrieh *et al* [33]	0.72 (pre)* and 0.75 (post)*
	Davison *et al* [34]	0.666
Lobular inflammation	Kleiner *et al* [4]	0.60
	Gawrieh *et al* [33]	0.37 (pre)* and 0.48 (post)*
	Davison *et al* [34]	0.227
Hepatocellular ballooning	Kleiner *et al* [4]	0.66
	Gawrieh *et al* [33]	0.32 (pre)* and 0.56 (post)*
	Davison *et al* [34]	0.487
CRN fibrosis stage	Kleiner *et al* [4]	0.85
	Gawrieh *et al* [33]	0.64 (pre)* and 0.75 (post)*
	Davison *et al* [34]	0.679
MAS	Davison *et al* [34]	0.372

*Pathologists in this study read slides before and after an intervention. The intervention consisted of review of illustrative histologic images of MASLD with the study pathologists and use of scoring sheet with written diagnostic criteria for different MASLD phenotypes.

Results

One hundred fifty-nine cases were enrolled in the study by three GT pathologists by reading glass slides using a light microscope. The distribution of slides based on slide-level score from glass GT is listed in Table 2. Based on the study glass GT, the slide set included 38.9% of cases that met the definition for challenging, borderline cases (defined as MAS ≥4 with a score of 0 for at least one of the features, MAS = 4 with a score of ≥1 for each of the features or MAS = 3).

Overall, the three study pathologists indicated the presence of atypical features 57 times, and, for 40 of these, the categorization was identical on glass and on WSI. For the discrepant cases, the WSI agreed with the GT 10 times, and glass agreed with GT six times (list of atypical features in supplementary material, Table S1).

Table 2. Distribution of slides based on glass ground truth (GT)

Feature	Score	% (n) N = 159
Steatosis	0	8.2 (13)
	1	32.7 (52)
	2	30.2 (48)
	3	28.9 (46)
Inflammation	0	1.9 (3)
	1	62.3 (99)
	2	34.0 (54)
	3	1.9 (3)
Ballooning	0	22.6 (36)
	1	56.6 (90)
	2	20.8 (33)
MAS	0	0.6 (1)
	1	6.3 (10)
	2	10.1 (16)
	3	19.5 (31)
	4	2.5 (4)
	4, all features have a score of at least 1	17.0 (27)
	5	22.6 (36)
	6	12.6 (20)
	7	8.2 (13)
	8	0.6 (1)
Fibrosis	0	6.9 (11)
	0.5*	0.6 (1)
	1	27.7 (44)
	2	28.9 (46)
	3	28.3 (45)
	4	7.5 (12)

*Fibrosis stage is 0.5 because median stages for all three GT pathologists were used.

The OPA for steatohepatitis was 74.3% (95% CI, 70.0%, 78.8%) for WSI versus GT and 74.5% for glass versus GT (Table 3). These results are in line with the published MASH diagnosis values in Davison *et al* [34], where the % agreement ranged from 69.5% to 81.4%. However, the primary endpoint in this study compares the average single reader agreement with a consensus read (three pathologists), whereas Davison *et al* [34] illustrate average individual-to-individual (pairwise) agreements.

The acceptance criteria for noninferiority (with a margin of 0.05) agreement for steatohepatitis evaluations between reads on WSI and glass GT compared to reads on glass and glass GT was met with a difference of −0.001 (95% CI, −0.027, 0.026; $p < 0.0001$; Table 3). Additionally, in the worst-case scenario that is compatible with the observed data (WSI versus GT CI lower bound = 0.700 and glass versus GT CI lower bound = 0.703), the digital method is at most 3% worse than glass.

Weighted kappas between WSI read and glass read for each steatohepatitis feature (Table 4) overall were determined (intrareader and intermodality). For each histologic feature, the overall weighted kappas were higher than the published values (Table 1). Average weighted kappas between WSI read and glass GT compared to glass read and glass GT for each steatohepatitis feature and CRN fibrosis (Table 5) were determined. For steatosis, the weighted kappas for WSI versus GT were 0.58 (95% CI, 0.505, 0.64) and glass versus GT 0.593 (0.519, 0.655); for lobular inflammation, the weighted kappas for WSI versus GT were 0.367 (0.3, 0.432) and glass versus GT 0.38 (0.315, 0.445); for hepatocellular ballooning, the weighted kappas for WSI versus GT were 0.537 (0.457, 0.608) and glass versus GT 0.522 (0.435, 0.595); and for fibrosis, the weighted kappas for WSI versus GT were 0.64 (0.574, 0.695) and glass versus GT 0.525 (0.473, 0.571). All these results are similar or slightly higher than previously published results (0.609, 0.328, 0.517, and 0.484 for steatosis, lobular inflammation, hepatocellular ballooning, and fibrosis, respectively). Weighed kappas between WSI reads and glass reads per histologic feature by pathologist are shown in supplementary material, Table S2.

OPA for steatohepatitis evaluations between reads on WSI and glass GT compared to reads on glass and glass GT were similar for all three pathologists (Table 6). For pathologist A, the difference between WSI reads and glass GT versus glass reads and glass GT was −0.006 (95% CI, −0.031, 0.0196). For pathologist B, the difference between WSI reads and glass GT versus glass reads and glass GT was 0.0278 (95% CI, −0.034, 0.089) and the difference for pathologist C was −0.025 (95% CI, −0.069, 0.016).

Table 3. Primary endpoint: overall percent agreement (OPA) between reads on WSI and glass ground truth (GT) versus reads on glass and glass GT (outlined in Figure 2)

Modality	N	Agreement rate (95% CI)	Difference (95% CI)	p value
WSI versus GT	159	0.743 (0.7, 0.788)	−0.001 (−0.027, 0.026)	<0.0001
Glass versus GT	159	0.745 (0.703, 0.786)		

Table 4. Average weighted kappa between WSI reads and glass reads per histologic feature (intrareader and intermodality)

Feature	N	Weighted kappa (95% CI), glass to WSI
Steatosis	159	0.882 (0.844, 0.916)
Inflammation	159	0.761 (0.707, 0.809)
Ballooning	159	0.788 (0.732, 0.835)
Fibrosis	159	0.872 (0.837, 0.901)
MAS	159	0.795 (0.76, 0.825)

Table 5. Average weighted kappa between individual reads on WSI and glass ground truth (GT) versus individual pathologist reads on glass and glass GT (outlined in Figure 2) for each histologic feature

Feature	Modality	*N*	Average weighted kappa (95% CI)
Steatosis	WSI versus GT	159	0.58 (0.505, 0.64)
	Glass versus GT	159	0.593 (0.519, 0.655)
Inflammation	WSI versus GT	159	0.367 (0.3, 0.432)
	Glass versus GT	159	0.38 (0.315, 0.445)
Ballooning	WSI versus GT	159	0.537 (0.457, 0.608)
	Glass versus GT	159	0.522 (0.435, 0.595)
Fibrosis	WSI versus GT	159	0.64 (0.574, 0.695)
	Glass versus GT	159	0.604 (0.536, 0.662)
MAS	WSI versus GT	159	0.527 (0.476, 0.573)
	Glass versus GT	159	0.525 (0.473, 0.571)

The number of discrepant reads between glass and WSI for pathologists A, B, and C were 6, 24, and 14 cases, respectively. The percentage of agreement for these discrepant cases for WSI versus GT and glass versus GT, respectively, was, for pathologist A, 40% and 60%; for pathologist B, 59% and 41%; and for pathologist C, 33% and 67% (list of discrepant cases in supplementary material, Table S3).

Difficulty of case reads per feature on WSI and on glass was determined by dividing cases into easy (all three pathologists agreed on the score), medium (two of the three pathologists agreed on the score), and hard (all three pathologists disagreed on a score). For steatosis, the agreement rate on case difficulty between glass and WSI was 81.9% and there were no cases that were classified as easy on glass and difficult on WSI or vice versa (Figure 3A). For lobular inflammation, the agreement rate on case difficulty between glass and WSI was 67.8%, and there was one case classified as hard on WSI and easy on glass. There were no cases that were classified as easy on WSI and hard on glass (Figure 3B). For hepatocellular ballooning, the agreement rate on case difficulty between glass and WSI was 76.2% and there was one case classified as hard on glass and easy on WSI. There were no cases classified as easy on glass and hard on WSI (Figure 3C). For fibrosis, the agreement rate on case difficulty between glass and WSI was 72.3% and there was one case on each WSI and glass classified as easy with one modality and hard with the other modality (Figure 3D).

Discussion

Due to the challenges with a lack of standardization and consistency in scoring these four different histologic components, all of which comprise a composite score which plays a crucial role in determining whether a patient is enrolled in a MASH study and whether a drug candidate has been effective, there is an urgent, unmet need in providing tools that can help solve these issues for such a prevalent disease with very limited treatment options. Currently, a multiple pathologist-derived consensus score is the gold standard employed in MASH trials, with the goal of reducing error and providing a more precise, standardized score. However, this consensus approach is burdensome, especially with challenging enrollment windows. The use of a cloud-based WSI IMS with a highly configurable, GCP compliant data capture system can reduce this burden and allow for MASH experts around the world to efficiently participate in this consensus. Before these systems and additional tools, such as AI-assistive algorithms, can be used by pathologists in a trial, equivalence to glass reads must be established [22]. This WSI IMS validation study demonstrates that the accuracy of steatohepatitis digital reads on the WSI IMS is equivalent to reads performed with traditional light microscopy with glass slides, specifically for the MASH trial context of use. Additionally, the agreement between WSI and glass

Table 6. Overall percent agreement (OPA) between reads on WSI and glass ground truth (GT) versus reads on glass and glass GT by individual pathologist

Pathologist	Pathologist overall experience*	Pathologist WSI experience†	Modality	*N*	Agreement rate (95% CI)	Difference (95% CI)
A	5 years	4 years	WSI versus GT	159	0.843 (0.786, 0.899)	−0.006 (−0.031, 0.019)
			Glass versus GT	159	0.849 (0.792, 0.906)	
B	20 years	8 years	WSI versus GT	158	0.633 (0.56, 0.707)	0.0278 (−0.034, 0.089)
			Glass versus GT	157	0.605 (0.529, 0.679)	
C	2 years	6 years	WSI versus GT	159	0.755 (0.686, 0.824)	−0.025 (−0.069, 0.016)
			Glass versus GT	159	0.780 (0.711, 0.843)	

*Experience since fellowship.
†Includes WSI experience prior to fellowship completion.

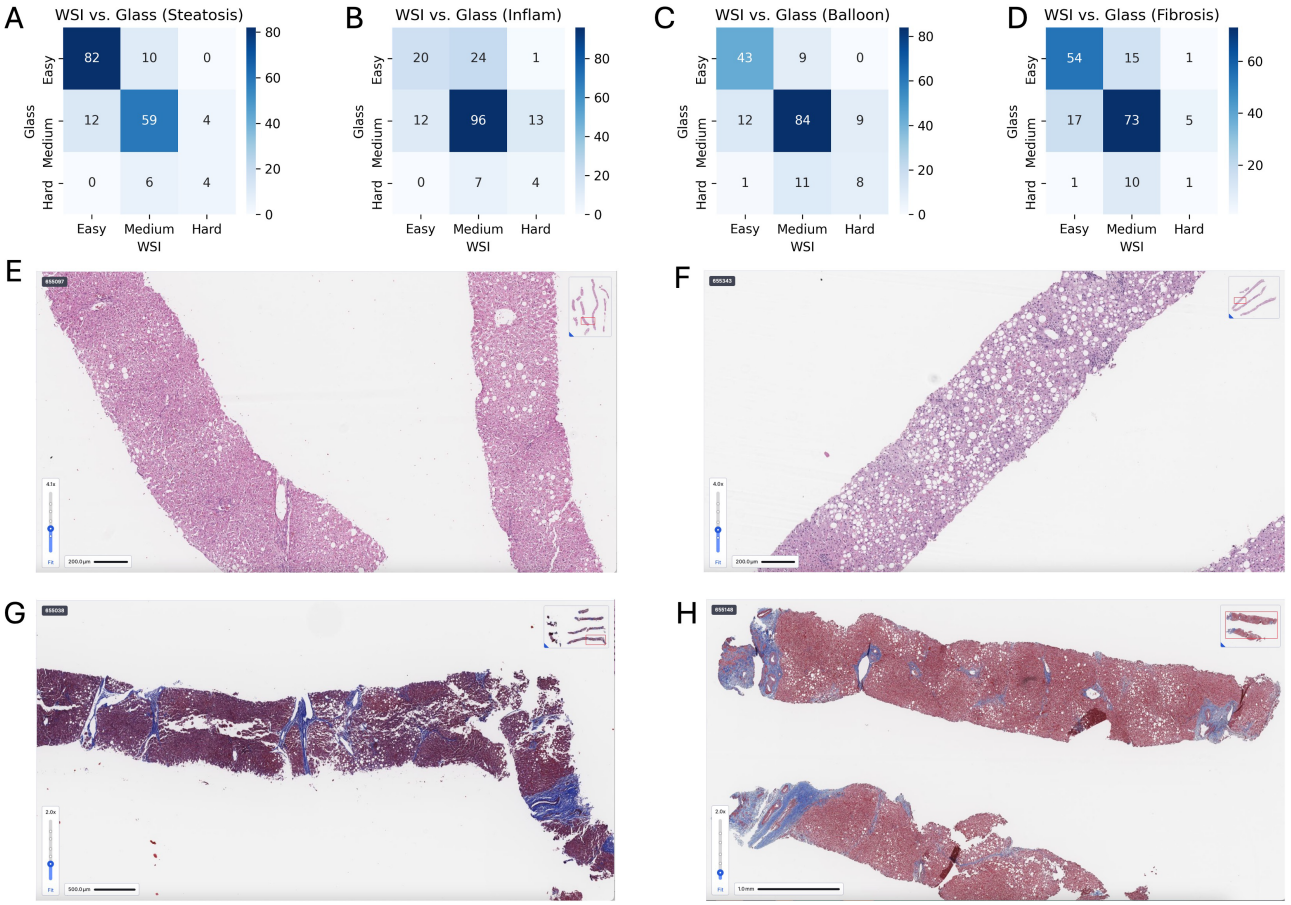

Figure 3. Comparison of easy, medium, and hard cases for glass and WSI reads. All cases in the study were divided into hard, medium, and easy based on pathologist agreement. Easy cases were defined as cases where all study pathologists agreed on the component score, medium cases as cases where two of three pathologists agreed on a score, and hard cases where all study pathologists disagreed on the component score. (A–D) Contingency tables of easy, medium, and hard cases for each of the categories (steatosis, inflammation, ballooning, and fibrosis). (E) An example of an easy H&E slide (all three study pathologists agreed on a score for each component). (F) An example of a medium/hard H&E slide (all 2/3 or 3/3 study pathologists disagreed on a score). (G) An example of an easy Masson's trichrome slide (all three study pathologists agreed on a score for each component). (H) An example of a medium/hard Masson's trichrome slide (all 2/3 or 3/3 study pathologists disagreed on a score).

GT reads versus glass and glass GT reads was shown to be similar for each individual participating pathologist. The use of a digital pathology platform facilitates multiple independent pathologists' reads in parallel and in consensus sessions, as is now commonly performed for MASH trials and recommended by the FDA [8–10]. This study design and these results are in line with studies performed for primary diagnoses by Leica [13] and Philips [12], despite this study dataset being enriched with borderline cases (Table 2) and considering the additional score-based requirements, demonstrating a significant noninferior OPA of steatohepatitis assessment between average individual WSI and GT reads versus average individual glass and GT reads [NI margin: 0.05; difference: −0.001; 95% CI: (−0.027, 0.026); and $p < 0.0001$; Table 3]. Additionally, the OPA between WSI and GT reads versus glass and GT reads were shown to be similar for each individual participating pathologist. Average intrareader, intermodality weighted kappas for each histologic score feature in this study were higher than weighted kappas in published literature (Table 1). Varying level of intrareader agreement was observed per pathologist per histologic feature, which is expected, as a wide range of intrareader weighted kappas have been demonstrated in the literature [4,33,34]. Importantly, results from all three pathologists were within the published ranges for intrareader

weighted kappas, with two of the three pathologists exceeding the published average weighted kappas for all four histologic features and all three pathologists exceeding the weighted kappa for overall MAS score. However, it is important to note key differences between the gold standard reference studies (e.g. Kleiner *et al* [4]) and this study dataset, which comprises a larger sample size and an interventional trial population versus the smaller, non-interventional clinical dataset examined in the Kleiner paper; additionally, the Kleiner study results do not specify individual intrapathologist values or confidence intervals.

To ensure that the platform was validated across steatohepatitis spectrum, the target study population was intended to be enriched with 5–10% challenging or borderline steatohepatitis cases (defined as MAS ≥4 with a score of 0 for at least one of the features, MAS = 4 with a score of ≥1 for each of the features or MAS = 3). Previously collected single central pathologist scores were used during study enrollment; however, based on the final study consensus GT, around 40% of the cases met the definition of being borderline or challenging. The observed difference in target versus actual percent challenging enrichment is consistent with published literature describing inter-pathologist agreement rates for MAS of approximately only 30% [34]. This level of enrichment contrasts with primary diagnosis studies where only around 5% of the cases included were considered to be borderline and/or challenging, and only major discordances in categorical diagnoses were counted toward disagreement rates, whereas any histologic feature score discrepancy was considered here. However, even with ∼40% challenging cases in this study, the 0.05 noninferiority margin was met. Overall, the diverse analyses performed here on the categorical diagnostic, individual component score, and composite score levels, provide strong evidence that individual pathologists can achieve equivalent levels of performance using this WSI IMS as they can using glass slides for MASH trial evaluations.

The cases read on WSI and glass were also categorized as easy (all three study pathologists agreed), medium (two of three pathologists agreed), or hard (all disagreed) (Figure 3). For the most components, agreement and disagreement were balanced around the agreement levels. In a couple of scenarios, glass considered more cases as hard than WSI and, in other scenarios, the reverse was true. However, these situations are to be expected, especially for inflammation and ballooning, where both inter- and intrareader agreements are quite low. For example, intrareader/intermodality agreements shown in Table 4 show that, on average, readers are the least consistent with inflammation scoring, and this is the same observation noted in the literature. These inconsistencies would lead to different levels of agreement, independent of those due to the modality (glass or WSI).

One possible limitation of the study includes the read order for the WSI and glass slides. Here, for study reads, digital slides were read before glass slides due to workflow and timeline constraints. We recognize that this is not ideal; however, we believe that the potential for bias that this introduces is minimal, given there is a sufficient wash-out period [30] between reads, the large sample size, and given the known intrareader variability in scoring, especially in a population so enriched for borderline cases.

The results from this validation study support the conclusion that the WSI IMS platform is equivalent to the glass read in reference to a robust glass GT when used by pathologists to evaluate steatohepatitis trial populations for diagnosis and trial-based scoring criteria and histologic feature scoring during enrollment and for follow-up timepoints. This IMS, along with the Leica AT2 WSI scanner, can therefore be utilized for individual and consensus steatohepatitis reads in clinical trials. The IMS and workflow presented here were designed specifically for use in clinical trials, utilizing a robust design with MASH trial-specific endpoints in a challenging study population, representative of both screened and enrolled patient samples. Incorporating an IMS with GCP-compliant and configurable data report forms into clinical trial workflows makes trial management more efficient, improves upon data integrity, allows for multiple reads in parallel, and provides opportunities to utilize the most experienced pathologists on reader panels as geographic location is no longer a limiting factor for selecting pathologists or shipping glass slides. Utilization of an IMS platform will allow pathologists from all over the world to work on the same cases simultaneously and provide their results within hours of slide upload, shortening trial enrollment timelines while allowing for accurate, gold-standard assessments. Additionally, other tools such as annotation and measurement tools can facilitate and add to the consensus experience. The details of how a WSI IMS will be utilized in trials as a part of the end-to-end, tissue-to-reporting trial read-out should be documented in a trial protocol and/or related documentation as outlined in the 2021 SPIRIT-Path extension guidelines [35], and this validation work can now be referenced. Finally, with the crucial foundation of this validation evidence, which established glass-to-WSI equivalence specific to the MASH trial use case, assistive AI-based tools can now be offered to pathologists as a part of the trial IMS to help solve the challenge of

lack of standardization in histologic scoring. The validation and establishment of glass-to-digital equivalence for this specific context of us were a crucial first step in a field, where there is an urgent need for histology read standardization for enrollment and primary endpoints. This approach can serve as a validation framework for other challenging diagnostic areas, particularly in clinical trials where pathology plays an important role and where multiple histologic features are evaluated and/or assessed by score (e.g. inflammatory bowel disease) [36,37].

Acknowledgements

This work was fully supported by PathAI, Inc.

Author contributions statement

HP contributed to the conception and design of the study, interpreted results and wrote the manuscript. SSM contributed to the conception and design of the study, performed data analysis, interpreted results and made critical revisions to the article. LCM, SK, HH, IN, XZ, EC, THK, RT and MT contributed to acquisition and cleaning of data. JNG and MBR contributed to the conception and design of the study and interpretation of data. NP performed data analysis and interpreted results. CET contributed to the conception and design of the study, interpreted results and made critical revisions to the article. PB, MCM and AHB contributed to the conception and design of the study and made critical revisions to the article. KEW contributed to the conception and design of the study, interpreted results and made critical revisions to the article.

Data availability statement

The data that support the findings of this study are not openly available due to reasons of sensitivity and are available from the corresponding author upon reasonable request. Data are located in an electronic quality management system at PathAI.

References

1. Younossi ZM, Koenig AB, Abdelatif D, *et al*. Global epidemiology of nonalcoholic fatty liver disease—meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology* 2016; **64**: 73–84.
2. Noureddin M, Vipani A, Bresee C, *et al*. NASH leading cause of liver transplant in women: updated analysis of indications for liver transplant and ethnic and gender variances. *Am J Gastroenterol* 2018; **113**: 1649–1659.
3. Friedman SL, Neuschwander-Tetri BA, Rinella M, *et al*. Mechanisms of NAFLD development and therapeutic strategies. *Nat Med* 2018; **24**: 908–922.
4. Kleiner DE, Brunt EM, Van Natta M, *et al*. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology* 2005; **41**: 1313–1325.
5. European Medicines Agency. Reflection Paper on Regulatory Requirements for the Development of Medicinal Products for Non-Alcoholic Steatohepatitis (NASH) (EMA/CHMP/111529/2024). 2023. [Accessed 12 July 2024]. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-regulatory-requirements-development-medicinal-products-non-alcoholic-steatohepatitis-nash_en.pdf
6. Food and Drug Administration. Nonalcoholic Steatohepatitis with Compensated Cirrhosis: Developing Drugs for Treatment Guidance for Industry – Draft. 2019. [Accessed 12 July 2024]. Available from: https://www.fda.gov/media/127738/download
7. Food and Drug Administration. Noncirrhotic Nonalcoholic Steatohepatitis with Liver Fibrosis: Developing Drugs for Treatment Guidance for Industry. 2018. [Accessed 12 July 2024]. Available from: https://www.fda.gov/media/119044/download
8. Sanyal AJ, Loomba R, Anstee QM, *et al*. Utility of pathologist panels for achieving consensus in NASH histologic scoring in clinical trials: data from a phase 3 study. *Hepatol Commun* 2024; **8**: e0325.
9. Matsubayashi T. Drug Development for Nonalcoholic Steatohepatitis (NASH) with Fibrosis: A Regulatory Perspective [Presentation]. Regulatory Perspectives for Development of Drugs for Treatment of NASH [Webinar]. Food and Drug Administration. 2021. [Accessed 12 July 2024]. Available from: https://www.fda.gov/drugs/news-events-human-drugs/regulatory-perspectives-development-drugs-treatment-nash-01292021-01292021
10. Anania FA, Dimick-Santos L, Mehta R, *et al*. Nonalcoholic steatohepatitis: current thinking from the Division of Hepatology and Nutrition at the Food and Drug Administration. *Hepatology* 2021; **73**: 2023–2027.
11. Buck TP, Dilorio R, Havrilla L, *et al*. Validation of a whole slide imaging system for primary diagnosis in surgical pathology: a community hospital experience. *J Pathol Inform* 2014; **5**: 43.
12. Mukhopadhyay S, Feldman MD, Abels E, *et al*. Whole slide imaging versus microscopy for primary diagnosis in surgical pathology. *Am J Surg Pathol* 2018; **42**: 39–52.
13. Borowsky AD, Glassy EF, Wallace WD, *et al*. Digital whole slide imaging compared with light microscopy for primary diagnosis in surgical pathology. *Arch Pathol Lab Med* 2020; **144**: 1245–1253.
14. Mills AM, Gradecki SE, Horton BJ, *et al*. Diagnostic efficiency in digital pathology. *Am J Surg Pathol* 2018; **42**: 53–59.
15. Amin S, Mori T, Itoh T. A validation study of whole slide imaging for primary diagnosis of lymphoma. *Pathol Int* 2019; **69**: 341–349.
16. Tabata K, Mori I, Sasaki T, *et al*. Whole-slide imaging at primary pathological diagnosis: validation of whole-slide imaging-based primary pathological diagnosis at twelve Japanese academic institutes. *Pathol Int* 2017; **67**: 547–554.

17. Lee JJ, Jedrych J, Pantanowitz L, *et al*. Validation of digital pathology for primary histopathological diagnosis of routine, inflammatory dermatopathology cases. *Am J Dermatopathol* 2018; **40**: 17–23.
18. Krishnamurthy S, Mathews K, McClure S, *et al*. Multi-institutional comparison of whole slide digital imaging and optical microscopy for interpretation of hematoxylin-eosin–stained breast tissue sections. *Arch Pathol Lab Med* 2013; **137**: 1733–1739.
19. Al-Janabi S, Huisman A, Nikkels PGJ, *et al*. Whole slide images for primary diagnostics of paediatric pathology specimens: a feasibility study. *J Clin Pathol* 2013; **66**: 218–223.
20. Hanna MG, Reuter VE, Hameed MR, *et al*. Whole slide imaging equivalency and efficiency study: experience at a large academic center. *Mod Pathol* 2019; **32**: 916–928.
21. Hanna MG, Reuter VE, Ardon O, *et al*. Validation of a digital pathology system including remote review during the COVID-19 pandemic. *Mod Pathol* 2020; **33**: 2115–2127.
22. Food and Drug Administration. Use of Biomarkers for Diagnosing and Assessing Treatment Response in Noncirrhotic NASH Trials. Webinar. 2023. [Accessed 12 July 2024]. Available from: https://www.fda.gov/drugs/news-events-human-drugs/use-biomarkers-diagnosing-and-assessing-treatment-response-noncirrhotic-nash-trials-09182023
23. Brunt EM, Janney CG, Di Bisceglie AM, *et al*. Nonalcoholic steatohepatitis: a proposal for grading and staging the histological lesions. *Am J Gastroenterol* 1999; **94**: 2467–2474.
24. Novo Nordisk A/S. Research Study on Whether Semaglutide Works in People with Non-Alcoholic Steatohepatitis (NASH) (ESSENCE). [Accessed 11 July 2024]. Available from: https://clinicaltrials.gov/study/NCT04822181
25. Akero Therapeutics, Inc. A Study of Efruxifermin in Non-Cirrhotic Subjects with Histologically Confirmed Nonalcoholic Steatohepatitis (NASH) (Harmony). [Accessed 11 July 2024]. Available from: https://clinicaltrials.gov/study/NCT04767529
26. Navidea Biopharmaceuticals. An Evaluation of the Safety of Intravenous Tc 99m Tilmanocept and a Comparison of Imaging with Sulfur Colloid in Subjects with and without NASH. [Accessed 11 July 2024]. Available from: https://clinicaltrials.gov/study/NCT03332940
27. Gannex Pharma Co. Ltd. Study to Evaluate the Safety, Tolerability, and Efficacy of ASC41 in Adults with NASH. [Accessed 11 July 2024]. Available from: https://clinicaltrials.gov/study/NCT05118360
28. Mayo Clinic. Statins for the Treatment of NASH (STAT NASH). [Accessed 11 July 2024]. Available from: https://clinicaltrials.gov/study/NCT04679376
29. 89bio, Inc. Study Evaluating the Safety, Efficacy and Tolerability of BIO89-100 in Subjects with Biopsy-Confirmed Nonalcoholic Steatohepatitis (NASH) (ENLIVEN). [Accessed 11 July 2024]. Available from: https://clinicaltrials.gov/study/NCT04929483
30. Evans AJ, Brown RW, Bui MM, *et al*. Validating whole slide imaging systems for diagnostic purposes in pathology. *Arch Pathol Lab Med* 2022; **146**: 440–450.
31. Proscia. Proscia Concentriq Dx 510(k). [Accessed 11 July 2024]. Available from: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K230839
32. PathAI. PathAI Novo 510(k). [Accessed 11 July 2024]. Available from: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K212361
33. Gawrieh S, Knoedler DM, Saeian K, *et al*. Effects of interventions on intra- and interobserver agreement on interpretation of nonalcoholic fatty liver disease histology. *Ann Diagn Pathol* 2011; **15**: 19–24.
34. Davison BA, Harrison SA, Cotter G, *et al*. Suboptimal reliability of liver biopsy evaluation has implications for randomized clinical trials. *J Hepatol* 2020; **73**: 1322–1332.
35. Kendall TJ, Robinson M, Brierley DJ, *et al*. Guidelines for cellular and molecular pathology content in clinical trial protocols: the SPIRIT-path extension. *Lancet Oncol* 2021; **22**: e435–e445.
36. Pell R, Oien K, Robinson M, *et al*. The use of digital pathology and image analysis in clinical trials. *J Pathol Clin Res* 2019; **5**: 81–90.
37. Barisoni L, Hodgin JB. Digital pathology in nephrology clinical trials, research, and pathology practice. *Curr Opin Nephrol Hypertens* 2017; **26**: 450–459.

SUPPLEMENTARY MATERIAL ONLINE

Table S1. List of atypical features as noted by the study pathologists

Table S2. Weighed kappas between WSI reads and glass reads per histologic feature by pathologist

Table S3. List of discrepant cases per pathologist