OPEN

Reproducibility of Rejection Grading in Uterus Transplantation: A Multicenter Study

Verena Broecker, MD,[1,2] Mats Brännström, PhD,[2] Hans Bösmüller, MD,[3] Eva Sticová, MD,[4] Jana Malušková, MD,[4] Andres Chiesa-Vottero, MD,[5] and Johan Mölne, MD, PhD[1,6]

Background: Diagnosis of rejection after uterus transplantation is based on histopathological examination of ectocervical biopsies. Inflammation at the stromal–epithelial interface is the backbone of the histopathological classification proposed by our group in 2017. However, the reproducibility of this grading scheme has not been tested, and it is unclear whether it covers the full morphological spectrum of rejection. **Methods:** We present a multicenter study in which 5 pathologists from 4 uterus transplantation centers performed 2 rounds of grading on 145 and 48 cervical biopsies, respectively. Three of the centers provided biopsies. Additionally, the presence of perivascular stromal inflammation was recorded. During discussions after the first round, further histological lesions (venous endothelial inflammation and apoptosis) were identified for closer evaluation and added to the panel of lesions to score in the second round. All participants completed a questionnaire to explore current practices in handling and reporting uterus transplant biopsies. **Results:** Cervical biopsies were commonly performed in all centers to monitor rejection. Intraobserver reproducibility of rejection grading (performed by 1 rater) was excellent, whereas interobserver reproducibility was moderate and did not improve in the second round. Reproducibility of perivascular stromal inflammation was moderate but unsatisfactory for venous endothelial inflammation and apoptosis. All lesions were more frequent in, but not restricted to, biopsies with rejection patterns. **Conclusions:** Grading of rejection in cervical biopsies is reproducible and applicable to biopsies from different centers. Diagnosis of rejection may be improved by adding further histological lesions to the grading system; however, lesions require rigorous consensus definition.(Transplantation Direct 2023;9: e1535; doi: 10.1097/TXD.0000000000001535.)

Uterus transplantation has become an alternative to surrogacy and adoption for women who wish to have children but experience absolute uterine factor infertility. Underlying causes include uterine agenesis as in the Mayer-Rokitansky-Küster-Hauser syndrome, with an estimated prevalence of 0.02% among females,[1,2] and hysterectomy for various causes, affecting between 1.7% and 14% of women under the age of 40.[2] Although still considered to be an experimental procedure, >80 transplantations have been performed during the past few years, and >40 children have been born after both live and deceased donor transplantations.[3] It is common practice for all types of solid organ transplants to take tissue biopsies from the allograft to detect rejection and guide immunosuppressive treatment. Although some of the morphological hallmarks of rejection are shared between all types of allografts, every organ has its characteristic spectrum of histological findings, reflected in organ-specific classification schemes for grading of rejection.[4-7] In 1999, the Banff working classification on renal allograft pathology was published, with the objective to "guide therapy and to establish

Received 30 June 2023. Revision received 27 July 2023.

Accepted 9 August 2023.

[1] Department of Clinical Pathology, Sahlgrenska University Hospital, Region Västra Götaland, Gothenburg, Sweden.

[2] Department of Obstetrics and Gynecology, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden.

[3] Institute of Pathology and Neuropathology, University Hospital Tübingen, Tübingen, Germany.

[4] Clinical and Transplant Pathology Department, Institute for Clinical and Experimental Medicine (IKEM), Prague, Czech Republic.

[5] Department of Anatomic Pathology, Cleveland Clinic, Cleveland, OH.

[6] Department of Laboratory Medicine, Institute of Biomedicine, Sahlgrenska Academy, University of Gothenburg, Sweden.

Correspondence: Verena Broecker, MD, Department of Clinical Pathology, Gula Straket 8, Sahlgrenska University Hospital, 43145 Gothenburg, Sweden. (verena.brocker@vgregion.se).

V.B. designed the study, performed the research and data analysis, and wrote the article. M.B. participated in the research design and writing of the article. H.B., E.S., J.M., and A.C.-V. participated in the performance of the research and writing of the article. J.M. participated in the research design, performing of the research, and writing of the article.

The data that underlie the results of the study presented are available from the corresponding author upon reasonable request.

The authors declare no conflicts of interest.

The study was funded by the Jane and Dan Olsson Foundation for Science, Swedish Research Council, and Knut and Alice Wallenberg Foundation.

Supplemental digital content (SDC) is available for this article. Direct URL citations appear in the printed text, and links to the digital files are provided in the HTML text of this article on the journal's Web site (www.transplantationdirect.com).

ISSN: 2373-8731

DOI: 10.1097/TXD.0000000000001535

an objective end point for clinical trials."[8] Since its first publication, the classification has been continuously under development to account for evolving knowledge and changing clinical needs.[9,10] Reproducibility of grading of rejection in kidney allografts has previously been reported to be unsatisfactory[11,12]; however, high interobserver agreement on scoring of histological lesions in allografts can be reached among experienced pathologists in a well-established field.[13]

In 2017, our group proposed a provisional histological grading scheme for the diagnosis of rejection on cervical biopsies after uterus transplantation.[14] This grading scheme focused on morphological patterns, rather than scoring of individual lesions, based on the spectrum of histological findings, which were present in 163 cervical biopsies from the first clinical trial of uterus transplantation.[15] Notably, all rejections were clinically silent. The degree and distribution of inflammation at the stromal–epithelial interface is the backbone of the proposed classification, and individual lesions, such as epithelial apoptosis, are mentioned but not considered in more detail. Perivascular stromal inflammation (PVSI) was not seen at the time but was described in a subsequent publication on transplant hysterectomies.[16] A slightly different grading scheme of rejection has recently been proposed by Agarwal et al,[17] in which lesions other than interface inflammation are incorporated. However, the prevalence of other lesions in transplant cervical biopsies and their association with an overall pattern of rejection has not been systematically investigated.

As uterus transplantation is being more widely implemented at transplant centers worldwide, it is important to standardize reporting of transplant cervical biopsies in a classification system that covers the full spectrum of rejection-associated morphological findings and is sufficiently reproducible. This is required to compare studies from different centers addressing treatment regimens based on pathology. As recently reported, rejection affects 33% of recipients 0 to 5 mo posttransplantation and 21% of recipients 6 to 10 mo posttransplantation.[18] In the present study, our objective was to test the reproducibility of our previously proposed grading scheme among pathologists from different centers and to identify additional histological lesions that are currently not considered in the classification but are potentially important for a diagnosis of rejection.

MATERIALS AND METHODS

Samples

The present study included 193 transplant cervical biopsies performed for monitoring of rejection within clinical trials of human uterus transplantation from 3 different transplant centers active in uterus transplantation. The clinical trials were approved by the respective institutional review boards (the Joint Ethics Committee of the Institute for Clinical and Experimental Medicine and Thomayer Hospital Prague [number 2044/15 (NM-15-01], Ethics Committee of the University of Tübingen [project identification code 211/2016A], and Regional Ethics Committee of Gothenburg, Sweden [#88-12]).[19-21] To cover the full range of the histological spectrum, from normal biopsies to severe rejection (Table 1), biopsies were selected by 1 pathologist (V.B.) not involved in the grading of rejection (128 biopsies from Gothenburg, 20 from Prague, and 45 from Tübingen). The composition of the cohort accounts for a sufficient overall sample size with enough cases in each category. However, the original diagnosis was not regarded as the gold standard. For each biopsy, 1 level from a routine hematoxylin and eosin–stained glass slide was scanned at ×40 magnification (Hamamatsu NanoZoomer S210, Kista, Sweden), omitting the slide ID. Slides were anonymized to all participants, and recipient-related clinical data were not collected. Pathologists from 4 centers (Gothenburg, Prague, Tübingen, and Cleveland) participated in grading of the biopsies.

TABLE 1.

Composition of the cohort

Diagnostic category	First-round cases (n)	Second-round cases (n)
No rejection	80	19
Borderline changes	37	20
Mild rejection	15	8
Moderate rejection	5	1
Severe rejection	8	0
Total	145	48

Three centers contributed cases for the first round. Centers contributed more cases than were finally selected for the analysis by the study lead (V.B.), who was ignorant of the original diagnosis and clinical data from the outside centers and who did not participate in the grading of rejection. Cases were selected to cover the full spectrum of histological findings. The numbers in the diagnostic categories are based on the study lead's judgment but did not serve as the gold standard. Please note that no previously unknown case of severe rejection could be retrieved for the second round.

Study Layout

To evaluate the current clinical practice for handling and reporting uterus transplant biopsies, each participating pathologist completed an online questionnaire before grading the biopsies. The study consisted of 2 rounds of grading (Figure 1). For the first round, 145 scanned slides were made available to participants on an online platform (Smart in Media, Köln, Germany) together with supporting documentation, including a brief introduction of the study outline, previously published articles that describe the rejection classification,[14,16] a schematic handout of the rejection classification (extracted from[14]), and sample photographs of PVSI. The latter has been described as "a cuff of mononuclear inflammatory cells, surrounding mainly small venules or capillaries, occasionally involving small arteries and arterioles."[16] For each electronic slide, participants had to choose 1 category according to the previously proposed 5-tier grading system for rejection[14] (no rejection/other, borderline changes, mild/grade 1 rejection, moderate/grade 2 rejection, and severe/grade 3 rejection) and recorded the presence or absence of PVSI. A free-text comment was optional. Five pathologists (H.B., E.S., A.C.V., J.M., and Ja.Ma.) independently graded rejection, and 5 pathologists (H.B., E.S., A.C.V., J.M., and V.B.) independently recorded the presence or absence of PVSI (which was not part of the original classification). After the first round of grading, participants met virtually to discuss difficult cases and to identify additional histological lesions for further assessment that might potentially be relevant but are currently not systematically considered for a diagnosis of rejection in the proposed grading system. The time between the start of the first and the start of the second grading round was 8 mo. For the second round of grading, the same 5 pathologists as before (H.B., E.S., A.C.V., J.M., and Ja.Ma.) graded rejection according to the proposed grading scheme on 48 new biopsies. Additionally, 5 pathologists (H.B., E.S., A.C.V., J.M., and V.B.) recorded the presence or absence of PVSI, venous

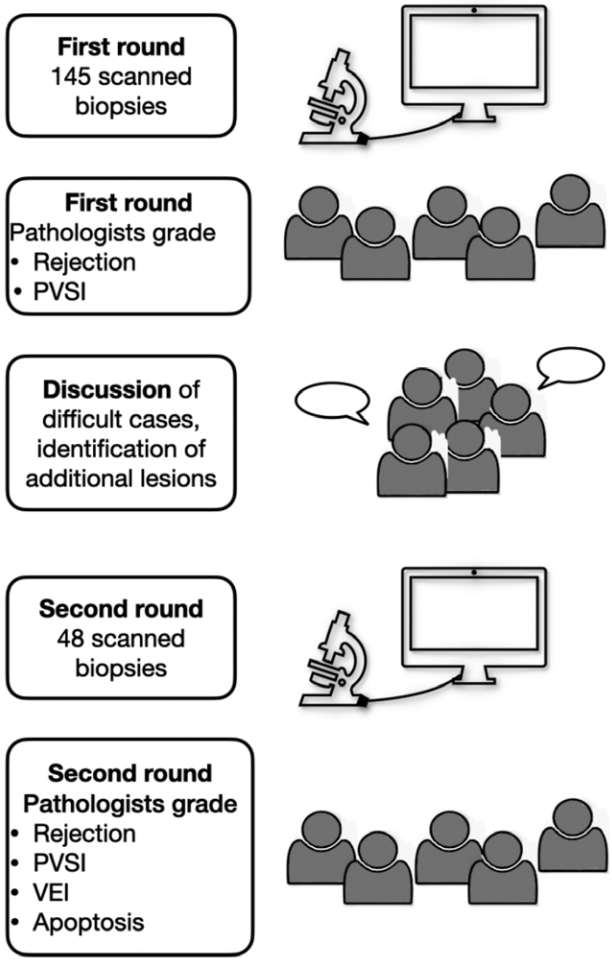

FIGURE 1. Study layout. The study was divided into 2 rounds of grading, based on 145 and 48 scanned slides, respectively. Five pathologists independently graded rejection according to the proposed grading scheme[13] (no rejection, borderline changes, mild, moderate, and severe rejection). In the first round, the presence/absence of PVSI was also recorded. After the first round of grading, participants met virtually to discuss controversial cases and to identify histological lesions, which are currently not considered for a diagnosis of rejection, and, however, might be part of the spectrum of rejection. Two additional lesions (venous endothelial inflammation and apoptosis) were recorded as absent/present in the second round, together with the grading of rejection and the recording of PVSI. PVSI, perivascular stromal inflammation; VEI, venous endothelial inflammation.

endothelial inflammation, and apoptosis in the basal epithelial layer (Figure 2). Venous endothelial inflammation was defined as the presence of inflammatory cells underneath the endothelium, lifting the endothelial cells.

Statistical Analysis

As the study design (multiple raters and a subset of variables on the ordinal scale) did not fit the required assumptions for calculating Cohen's κ or Fleiss' κ agreement scores, we sought an agreement measure applicable to the current study. Both Gwet's AC1 and Krippendorff's alpha can be applied to multiple raters performing ratings on ordinal scales. Krippendorff's alpha has rarely been used in pathology.[22] However, as pathologists commonly use κ coefficients, we applied several agreement scores to facilitate comparison with previous reproducibility studies within pathology without compromising the required statistical assumptions.

To determine the overall level of inter- and intrarater agreement, Krippendorff's alpha coefficient was calculated, which is applicable to any number of raters (in contrast to Cohen's κ) and ratings on an ordinal scale (in contrast to Fleiss' κ).[23,24] Krippendorff's alpha and κ scores were compared with Fleiss'/Cohen's κ coefficients for those analyses meeting the required assumptions. Gwet's AC1 agreement score has recently been used in pathology[25,26] and has also been calculated for comparison.[27] Intrarater agreement score was assessed for 1 pathologist (J.M.) from the coordinating center (assessed on the subset of 80 cases from that center for the first grading round), by comparing the participant's study diagnosis with the original clinical diagnosis, previously rendered by the same pathologists (wash-out time >4 y in 86% of biopsies, >1 y in 99%, and 1 biopsy with 4-mo wash-out time). This pathologist was not involved in the selection of biopsies for the study. As per convention, an agreement coefficient <0.2 was regarded as poor, >0.2 to 0.4 as fair, >0.4 to 0.6 as moderate, >0.6 to 0.8 as good, and >0.8 as excellent. Linear weights were used for ordinal variables. The irrCAC package in R/Bioconductor (version 3.3.2.) was used for the calculation of agreement scores. To analyze whether PVSI, venous endothelial inflammation, or apoptosis were associated with rejection, the 2-sided Fisher exact test was used. For this analysis, a reference diagnosis for each lesion had to be established, which was done by calculating the mean of pairwise interrater κ agreement scores between participants. The judgment of the participant with the highest mean κ score for any variable served as the reference for that specific variable, respectively. Data collection and statistical analysis were performed in IBM SPSS Statistics 28. Graphs were created using Prism 9, Microsoft PowerPoint version 2202, and Affinity Designer 1.10.5. A P value of <0.05 was considered to be statistically significant. In the case of multiple testing, P values were adjusted using the Bonferroni method.

RESULTS

The Previously Proposed Grading Scheme of Rejection Is Commonly Applied to Uterus Transplant Cervical Biopsies

To explore the current clinical practice regarding reporting uterus transplant biopsies, the participating pathologists were asked to fill out a questionnaire before grading the biopsies. The results of the survey are shown in Table 2. All participants were specialist pathologists and were regularly involved in reporting uterus transplant biopsies. All participants reported transplant pathology other than uterus transplant pathology in their daily practice, whereby kidney, liver, and intestines were most frequent. Most participants (5/6) apply the grading classification previously proposed by our group[14] when reporting transplant cervical biopsies, whereas 1 participant normally makes a descriptive diagnosis. Taking punch biopsies from the cervix to monitor rejection was common practice at all the participating centers.

Agreement Scores Calculated Using κ and Krippendorff's Alpha Give Comparable Results

Agreement scores using different measures as described in detail in the Materials and Methods section are displayed in Table 3, showing that Krippendorff's alpha produced very similar results compared with Cohen's and Fleiss' κ scores in the correct settings, whereas Gwet's AC1 appears to result

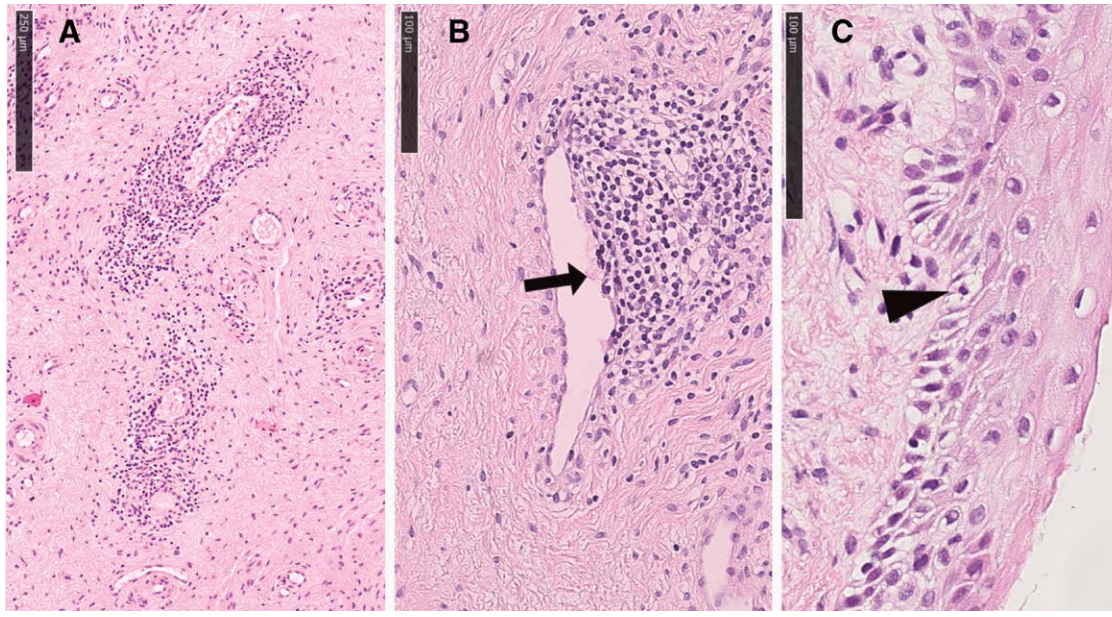

FIGURE 2. Histological lesions. A, Perivascular stromal inflammation: inflammatory infiltrates, dominated by lymphocytes, surrounding venules in the cervical stroma. Perivascular stromal inflammation was regarded as being present in this sample by 5 of 5 pathologists. B, Venous endothelial inflammation: inflammatory cells underneath the endothelium, lifting the endothelial cells (arrow). Venous endothelial inflammation was regarded as being present in this sample by 3 of 5 pathologists. C, Apoptosis: an apoptotic cell is seen in the basal layer of the surface squamous epithelium (arrowhead). Apoptosis was regarded as being present in this sample by 4 of 5 pathologists. (All hematoxylin and eosin–stained samples were originally scanned at ×40 magnification).

TABLE 2.

Answers to questionnaire

Question	Answer
Are you a specialist in pathology or a pathologist in training?	6/6[a] specialized pathologists
Are you working subspecialized?	5/6 subspecialized pathologists
	1/6 general pathologist
Are you involved in reporting biopsies from uterus transplants?	6/6 regularly involved in reporting of uterus transplant biopsies
Do you report transplant pathology other than uterus transplants?	6/6 report transplant pathology other than the uterus
If you do report other transplant pathology, which organ(s)?	Kidney: 5/6
	Liver: 4/6
	Intestines: 5/6
	Heart: 1/6
	Skin: 1/6
Do you report gynecological pathology other than uterus transplants in your daily practice?	5/6 report gynecological pathology other than uterus transplants
How many uterus transplant biopsies have you reported approximately?	Range from 50 to 200
Do you apply any grading classification when reporting transplant cervical biopsies?	5/6 apply the grading system proposed by Mölne et al[14]
	1/6 make a descriptive diagnosis
Where are biopsies from uterus transplants normally taken at your center?	3/4[b] cervix (ectocervix)
	1/4 cervix (transformation zone)
What type of biopsy do you normally get from uterus transplants?	4/4 punch biopsies

[a]Please note that 6 pathologists in total were involved in the study but only 5 performed the grading in each round, as described in the Materials and Methods section.
[b]Questions that apply to the respective center's practice refer to n = 4 in total.

in comparatively high agreement scores, likely because some agreement statistics are more susceptible than others to the prevalence of lesions in the cohort and to each rater's likelihood to use certain categories as discussed by Wongpakaran et al.[28]

Intrarater Agreement on the Grading of Rejection Is Excellent, Interrater Agreement Moderate

Grading of rejection in the first round was performed on 145 transplant cervical biopsies, based on the previously proposed 5-tier grading scheme.[14] Interrater agreement was moderate (Krippendorff's alpha 0.574; Table 3 and Figure 3) and did not improve when the 3 rejection grades were lumped to "rejection" in the analysis (Krippendorff's alpha 0.520, data not shown). In 38 of 145 cases (26.2%), 5 pathologists agreed on the same grade, in 73 of 145 pathologists (50.3%) were divided between 2 different grades (4:1 in 39 cases, 3:2 in 34 cases), and in 30 of 145 (20.7%) and 4 of 145 (2.8%) cases, 3 and 4 different grades were considered, respectively. No case received 5 different opinions. In cases with 2 different gradings proposed, the differential was mostly between neighboring categories (64/73 cases; 88%), across the full spectrum of gradings. In cases with 3 different gradings proposed, opinions were mostly divided between "no rejection"/"borderline

TABLE 3.
Comparison between agreement measures

Variable	Variable level	No. of raters (participant's initials)	Type of agreement measure			
			Cohen's κ (unweighted), mean of pairwise agreements	Fleiss' κ (>2 raters, unweighted)	Weighted κ; mean of pairwise agreements	Krippendorff's alpha (>2 raters, weighted or unweighted as applicable)	Gwet's AC1 (>2 raters, weighted or unweighted as applicable)
Rejection grading[a] first round	Ordinal	5 (J.M., H.B., A.C.V., E.S., Ja.Ma.)	Unweighted test not applicable to ordinal variable	Unweighted test not applicable to ordinal variable	0.576	0.574 (SD 0.038; 95% CI, 0.49-0.65)	0.706 (SD 0.02; 95% CI, 0.66-0.75)
Rejection grading[a] second round	Ordinal	5 (J.M., H.B., A.C.V., E.S., Ja.Ma.)	Unweighted test not applicable to ordinal variable	Unweighted test not applicable to ordinal variable	0.506	0.498 (SD 0.07; 95% CI, 0.35-0.65)	0.743 (SD 0.04; 95% CI, 0.66-0.82)
PVSI first round	Nominal	5 (J.M., H.B., A.C.V., E.S., V.B.)	0.536	0.538 (SD 0.03; 95% CI, 0.49-0.59)	Weighted test not applicable to nominal variable	0539 (SD 0.058; 95% CI, 0.42-0.65)	0.797 (SD 0.03; 95% CI, 0.73-0.86)
PVSI second round	Nominal	5 (J.M., H.B., A.C.V., E.S., V.B.)	0.502	0.513 (SD 0.046; 95% CI, 0.42-0.6)	Weighted test not applicable to nominal variable	0.517 (SD 0.078; 95% CI, 0.35-0.67)	0.735 (SD 0.071; 95% CI, 0.59-0.88)
VEI second round	Nominal	5 (J.M., H.B., A.C.V., E.S., V.B.)	0.231	0.219 (SD 0.046; 95% CI, 0.129-0.31)	Weighted test not applicable to nominal variable	0.221 (SD 0.05; 95% CI, 0.01-0.338)	0.82 (SD 0.05; 95% CI, 0.78-0.923)
Apoptosis second round	Nominal	5 (J.M., H.B., A.C.V., E.S., V.B.)	0.235	0.197 (SD 0.046; 95% CI, 0.107-0.287)	Weighted test not applicable to nominal variable	0.206 (SD 0.06; 95% CI, 0.08-0.32)	0.400 (SD 0.074; 95% CI, 0.25-0.55)

Cohen's κ, Fleiss' κ, and weighted κ values for 2 raters were calculated using SPSS version 28, and Krippendorff's alpha and Gwet's AC1 were calculated using irrCAC package in R/Bioconductor (version 3.3.2).
[a]Based on 5-tier grading (no rejection, borderline changes, mild, moderate, and severe rejection).
CI; confidence interval; PVSI, perivascular stromal inflammation; SD; standard deviation; VEI, venous endothelial inflammation.

changes"/"mild rejection" (20/30 cases; 67%). Less variation in opinions was observed at the extreme ends of the spectrum; of all cases that were regarded as "no rejection" by any of the participants (n = 98), 47% received the same diagnosis by at least 4 participants; of all cases that were regarded as "severe rejection" by any of the participants (n = 13), 69% received the same diagnosis by at least 4 participants. These percentages were much lower for cases regarded as "borderline changes," "mild rejection," or "moderate rejection" by any of the participants (13.5%, 11.3%, and 14%, respectively).

Intrarater agreement score performed on the 145 first-round samples was 0.865 (SD 0.039; 95% confidence interval [CI], 0.789-0.941; same values for weighted κ and Krippendorff's alpha).

Grading of rejection in the second round was performed on a new set of 48 transplant cervical biopsies using the same grading scheme. Again, rejection grading reached a moderate interrater agreement (Krippendorff's alpha 0.498; Table 3 and Figure 3) and did not improve when 3 rejection grades were lumped to "rejection" in the analysis (Krippendorff's alpha 0.498, data not shown).

Interrater Agreement on PVSI is Moderate but Unsatisfactory on Venous Endothelial Inflammation and Apoptosis

As we have previously shown that PVSI can be seen in hysterectomy specimens that show other features of rejection,[16] we included PVSI as a lesion to be scored in the study. Assessment of PVSI (absent/present) reached a moderate interrater agreement in both the first and the second grading rounds (Krippendorff's alpha 0.539 and 0.517, respectively; Table 3 and Figure 3). Both venous endothelial inflammation (present/absent) and apoptosis (present/absent) were assessed in the second round and showed low agreement scores (Krippendorff's alpha 0.221 and 0.206, respectively; Table 3 and Figure 3).

PVSI, Apoptosis, and Venous Endothelial Inflammation Are More Frequent in Biopsies With a Rejection Pattern

To explore whether PVSI, venous endothelial inflammation, and apoptosis might be part of the morphological spectrum of rejection and should therefore be systematically considered for diagnosis, we assessed whether there was an association between the grading and these lesions. For this purpose, a reference diagnosis had to be established for each variable as described in the Materials and Methods section. Although venous endothelial inflammation was more frequent in mild rejection, only PVSI (in the first round) and apoptosis were significantly associated with rejection (Figure 4) but did not increase with the grade of rejection. However, all lesions were also seen in rare cases without rejection or borderline pattern (thus cases with no or only mild patchy inflammation at the stromal–epithelial interface). To see whether PVSI increased from "no rejection" to "rejection" biopsies for each center, we analyzed the presence of PVSI separately for each center (**Figure S1, SDC**, http://links.lww.com/TXD/A567), showing that the pattern was similar among centers.

DISCUSSION

Uterus transplantation is the most recent achievement among solid organ transplantations, with the first live birth occurring

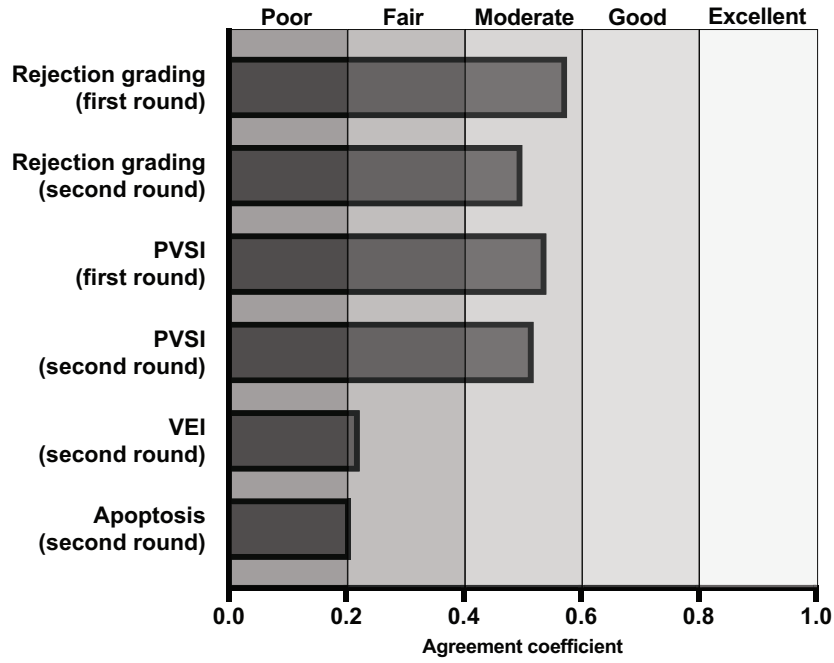

FIGURE 3. Interrater agreement coefficients. As described in the Materials and Methods section, the interrater agreement coefficients displayed here were calculated using Krippendorff's alpha coefficient. Both grading of rejection (based on the 5-tier grading scheme) and recording of PVSI were performed in both the first and second rounds. VEI and apoptosis were recorded in the second round, after the group discussion. PVSI, perivascular stromal inflammation; VEI, venous endothelial inflammation.

in 2014.[29] Protocol biopsies are regularly taken from the uterine cervix to monitor rejection after transplantation and guide immunosuppressive treatment. In 2017, we proposed a pattern-based histopathological grading scheme for rejection in human uterus transplants,[14] but it has not yet been tested if this grading scheme is reproducible among pathologists. To test this was the primary aim of the present study. Furthermore, it is uncertain whether the proposed grading scheme considers the full morphological spectrum of rejection or whether some morphological lesions are missing. Therefore, we also aimed to identify additional histological lesions, which may be part of the spectrum of rejection, and to test their reproducibility.

We found that

- Cervical biopsies to monitor rejection are common practice in centers performing uterus transplantation, and our previously proposed grading scheme has gained wide acceptance.
- Grading of rejection according to that proposed 5-tier grading scheme achieved moderate interrater and excellent intrarater agreement. Grading was more reliable at the extreme ends of the morphological spectrum.
- Additional histological lesions, which might be part of the morphological spectrum of rejection in uterus cervical biopsies, included PVSI, apoptosis in the basal epithelial layer, and venous endothelial inflammation in the cervical stroma. Interrater reproducibility was moderate for PVSI but unsatisfactory for both venous endothelial inflammation and apoptosis.
- All lesions mentioned previously were more frequent in biopsies with a rejection pattern but could also be seen in biopsies without epithelial–stromal interface inflammation.

The first consensus paper on the classification of rejection in kidney transplant biopsies was published in 1999, driven by the apparent need for a standardized reporting system of kidney allograft pathology to standardize treatment and support clinical trials.[8] Although this classification has its limitations in terms of inter- and intraobserver reproducibility,[9] it has gained acceptance internationally to the extent that it is used in clinical trials and is continuously under development.[7] There are several possible explanations for this seeming discrepancy: statistical agreement measures may not sufficiently reflect the real world, in which pathologists at the same center discuss cases among each other and with clinicians against a clinical background. Scoring individual lesions is more difficult than deciding on an overall morphological pattern. Participants' vigilance might be lower when asked to score multiple parameters in an artificial setting compared with real-world diagnostics. The development and continuous improvement of the rejection classification in kidneys can nonetheless serve as a raw model for rejection classification in uterus transplants.

In contrast to rejection classifications in most other solid organ transplants, there is no independent biomarker after uterus transplantation, such as serum creatinine or liver enzyme test, which histological changes in cervical biopsies can be validated against at the time of biopsy. Histological rejections in uterus transplants are typically clinically silent,[15] and there is no clear association with pregnancy and childbirth.[16] The lack of a granular short-term outcome parameter is one of the main obstacles in developing a clinically meaningful histological grading scheme for uterus transplants, no matter how well reproducible it may be. Although the current study was not designed to test the prognostic implication of the proposed classification, it will help to establish a robust basis for further studies that address immunosuppression protocols and outcomes in uterus transplantation. The analysis of the transcriptome in transplant (and nontransplant) cervical biopsies or measurement of plasma cell-free DNA[30] may

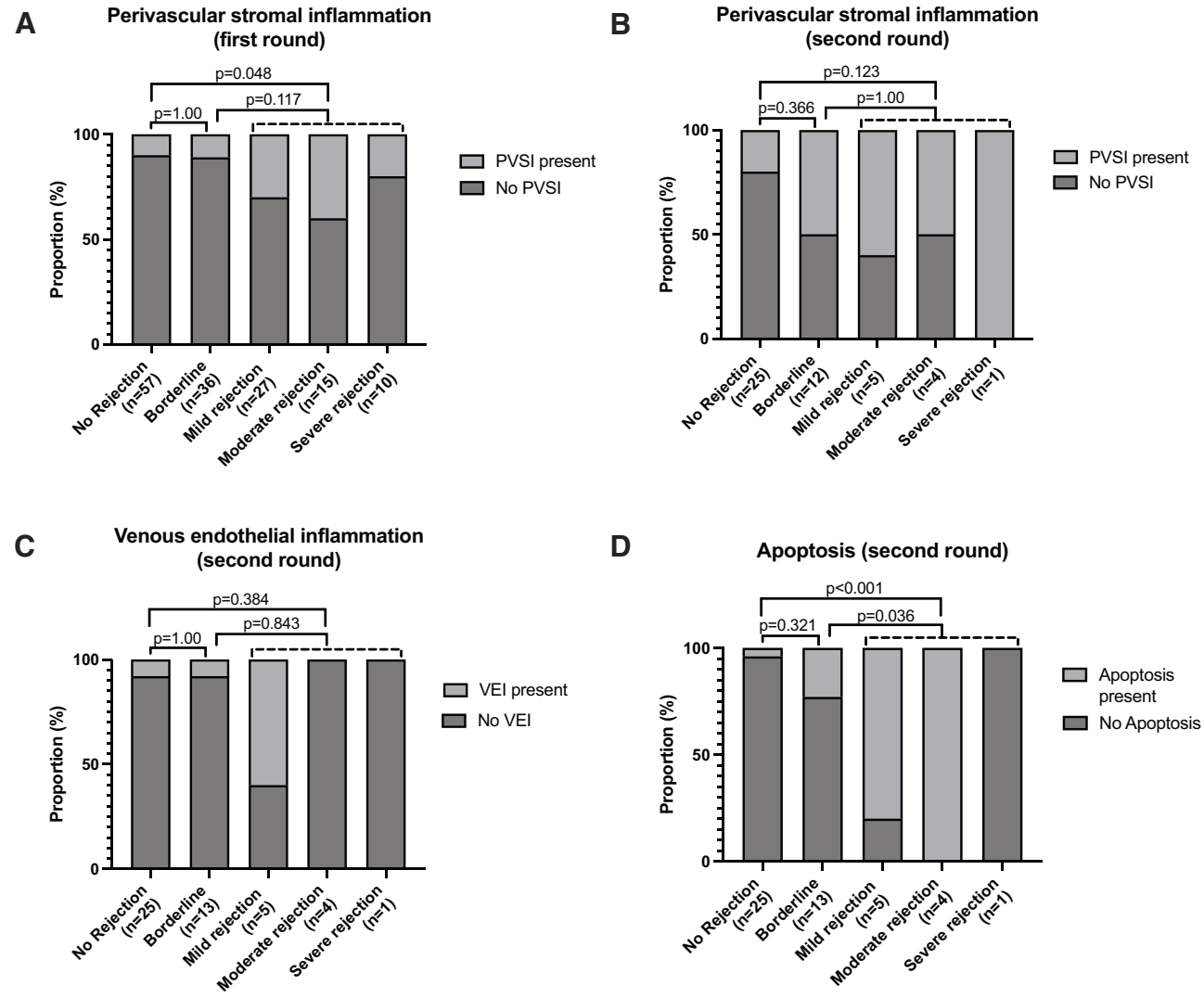

FIGURE 4. Association of histological lesions with rejection. For this analysis, a gold-standard diagnosis was established for each parameter (rejection grade, presence/absence of VEI, PVSI, and apoptosis, respectively) in the following manner: For each pathologist, the mean of all pairwise interrater agreement coefficients was calculated. The pathologist with the highest mean interrater agreement score for a specific parameter served as the gold standard for that particular parameter. Please note that the numbers in each diagnostic group differ from the numbers in Table 1 (which represent the study lead's judgment but not the gold standard). Groups were compared as follows: no rejection versus borderline; no rejection versus rejection (all); borderline versus rejection (all). All displayed *P* values were adjusted for multiple testing. A, Association of PVSI with rejection in the first round. B, Association of PVSI with rejection in the second round. Please note that 1 case was missed during grading by one of the participants; therefore, a total of n = 47. C, Association of VEI with rejection in the second round. D, Association of apoptosis with rejection in the second round. PVSI, perivascular stromal inflammation; VEI, venous endothelial inflammation.

further help to address the need for alternative tools to diagnose rejection after uterus transplantation.

Interobserver reproducibility of rejection was moderate in our study, which is similar to what other studies have found for grading rejection after kidney, heart, and lung transplantation (reviewed in reference [9]).[22,31] Scoring of individual histological lesions was affected by lower interobserver agreement than an overall diagnosis of rejection in the reviewed studies,[9] which was also true for our study. Concerning kidney rejection grading, Furness et al[12] explored whether continuous feedback to pathologists on their deviation from the group average could improve interobserver agreement and found that it had little effect on the reproducibility of histological lesions. In our study, the discussion of difficult cases did not affect agreement in the second round of grading. Rosales et al[32] assessed interobserver agreement for histological lesions related to epithelial damage and vascular inflammation in porcine skin-related vascular composite allografts, which bore some similarities to uterine allografts. They reported comparably high weighted κ scores of 0.673 and 0.663 for scoring perivascular inflammation and the presence of keratinocyte apoptosis and necrosis, respectively.[32] However, reproducibility was assessed between 2 experienced pathologists working at the same center. A recent study addressing the reproducibility of histological lesions in late pediatric transplant liver biopsies showed excellent observer agreement.[13] The study was performed by experts in a well-established field, which is different from the present study, exploring an emerging field in pathology in which each participant has limited experience so far.

In the present study, we assessed PVSI, apoptosis, and venous endothelial inflammation as additional morphological features, which have not been systematically considered for a diagnosis of rejection in the previously proposed grading

scheme. Notably, these lesions were not seen in the series of 163 uterus transplant cervical biopsies, which served as the basis for the proposed grading scheme.[14] However, PVSI was later described in transplant hysterectomies after childbirth from the first completed clinical trial of uterus transplantations.[16] This underscores the need for constant evaluation and amendment of existing classification schemes with more data becoming available to incorporate the full morphological spectrum of rejection and nonrejection changes in uterus transplants.

A strength of the current study is the inclusion of biopsies from different centers and involving pathologists working at different centers, who provided their input from various angles of transplant pathology. Apoptosis, for example, is an important criterion for both rejection and graft-versus-host disease in intestinal biopsies although this feature does not play a role in rejection diagnosis in kidney allografts. Although apoptosis as a feature of rejection was mentioned in the proposed grading scheme, it had not been studied in detail. Kreft et al[33] assessed the interobserver reproducibility of diagnosis and grading of graft-versus-host disease, for which apoptosis is a key lesion, and found that a rigorous definition of consensus criteria improved κ values for grading from 0.322 (fair) to 0.455 (moderate). Venous endothelial inflammation is important in liver allografts but is ignored in kidney allografts. All histological lesions analyzed in our study (PVSI, venous endothelial inflammation, and apoptosis) were more frequent in biopsies diagnosed with rejection compared with nonrejection; however, they were also seen in biopsies without any interface inflammation. Additionally, the lesions did not become more frequent with higher grades of rejection. Therefore, the significance of these lesions for the diagnosis of rejection remains uncertain, and more studies are required to address whether, for example, isolated PVSI without inflammation at the stromal–epithelial interface or apoptosis in the absence of inflammation should be regarded as a form of rejection. Agarwal et al[17] recently proposed a scoring system of rejection in uterus transplants based on interface inflammation, epithelial changes, and stromal and vascular inflammation. However, the frequency of the different lesions at different grades of rejection, the weighting of individual lesions for an overall diagnosis of rejection in each diagnostic category, and the reproducibility of lesions were not described.

Although our study included biopsies from different centers, it was not primarily designed to assess whether the histological spectrum differs between centers, potentially related to differences in immunosuppression. Notably, the proportion of biopsies showing PVSI was higher in all diagnostic categories in biopsies from the second round, which originated from more recent transplants under a slightly reduced immunosuppressive protocol.

In conclusion, we found that both a histological pattern-based diagnosis of rejection and PVSI as an individual histological lesion were reproducible on cervical biopsies from uterus transplants. In contrast, the reproducibility of additional lesions, such as venous endothelial inflammation and apoptosis, was unsatisfactory. Grading of rejection and lesions may be improved after rigorous consensus definitions, training, and with help of artificial intelligence, when larger data sets become available. PVSI and apoptosis were significantly associated with a diagnosis of rejection and may therefore be important lesions to consider for a diagnosis of rejection. We suggest documenting PVSI, apoptosis, and venous endothelial inflammation in pathology reports from transplant cervical biopsies as individual lesion scores as a supplement to an overall diagnosis of rejection to facilitate further studies, addressing the prevalence and clinical significance of these findings.

ACKNOWLEDGMENTS

The authors thank Toshima Parris for her valuable comments.

REFERENCES

1. Herlin M, Bjorn AM, Rasmussen M, et al. Prevalence and patient characteristics of Mayer-Rokitansky-Küster-Hauser syndrome: a nationwide registry-based study. *Hum Reprod.* 2016;31:2384–2390.
2. Sallee C, Margueritte F, Marquet P, et al. Uterine factor infertility, a systematic review. *J Clin Med*. 2022;11:4907.
3. Brannstrom M, Belfort MA, Ayoubi JM. Uterus transplantation worldwide: clinical activities and outcomes. *Curr Opin Organ Transplant.* 2021;26:616–626.
4. Stewart S, Winters GL, Fishbein MC, et al. Revision of the 1990 working formulation for the standardization of nomenclature in the diagnosis of heart rejection. *J Heart Lung Transplant.* 2005;24:1710–1720.
5. Cendales LC, Kanitakis J, Schneeberger S, et al. The Banff 2007 working classification of skin-containing composite tissue allograft pathology. *Am J Transplant.* 2008;8:1396–1400.
6. Demetris AJ, Bellamy C, Hubscher SG, et al. 2016 comprehensive update of the Banff working group on liver allograft pathology: introduction of antibody-mediated rejection. *Am J Transplant.* 2016;16:2816–2835.
7. Loupy A, Haas M, Roufosse C, et al. The Banff 2019 kidney meeting report (I): updates on and clarification of criteria for T cell- and antibody-mediated rejection. *Am J Transplant.* 2020;20:2318–2331.
8. Racusen LC, Solez K, Colvin RB, et al. The Banff 97 working classification of renal allograft pathology. *Kidney Int.* 1999;55:713–723.
9. Mengel M, Sis B, Halloran PF. SWOT analysis of Banff: strengths, weaknesses, opportunities and threats of the international Banff consensus process and classification system for renal allograft pathology. *Am J Transplant.* 2007;7:2221–2226.
10. Mengel M, Loupy A, Haas M, et al. Banff 2019 meeting report: molecular diagnostics in solid organ transplantation-consensus for the Banff Human Organ Transplant (B-HOT) gene panel and open source multicenter validation. *Am J Transplant.* 2020;20:2305–2317.
11. Furness PN, Taub N; Convergence of European Renal Transplant Pathology Assessment Procedures (CERTPAP) Project. International variation in the interpretation of renal transplant biopsies: report of the CERTPAP Project. *Kidney Int.* 2001;60:1998–2012.
12. Furness PN, Taub N, Assmann KJ, et al. International variation in histologic grading is large, and persistent feedback does not improve reproducibility. *Am J Surg Pathol.* 2003;27:805–810.
13. Hubscher SG, Feng S, Gouw ASH, et al. Standardizing the histological assessment of late posttransplantation biopsies from pediatric liver allograft recipients. *Liver Transpl.* 2022;28:1475–1489.
14. Mölne J, Broecker V, Ekberg J, et al. Monitoring of human uterus transplantation with cervical biopsies: a provisional scoring system for rejection. *Am J Transplant.* 2017;17:1628–1636.
15. Johannesson L, Kvarnstrom N, Mölne J, et al. Uterus transplantation trial: 1-year outcome. *Fertil Steril*. 2015;103:199–204.
16. Broecker V, Brännstrom M, Ekberg J, et al. Uterus transplantation: histological findings in explants at elective hysterectomy. *Am J Transplant*. 2021;21:798–808.
17. Agarwal A, Johannesson L, Findeis SK, et al. Clinicopathological analysis of uterine allografts including proposed scoring of ischemia reperfusion injury and T-cell-mediated rejection—Dallas uterus transplant study: a pilot study. *Transplantation.* 2022;106:167–177.
18. Brannstrom M, Tullius SG, Brucker S, et al. Registry of the International Society of Uterus Transplantation: first report. *Transplantation*. 2023;107:10–17.
19. Brucker SY, Strowitzki T, Taran FA, et al. Living-donor uterus transplantation: pre-, intra-, and postoperative parameters relevant to

surgical success, pregnancy, and obstetrics with live births. *J Clin Med*. 2020;9:2485.
20. Brannstrom M, Dahm-Kahler P, Kvarnstrom N, et al. Reproductive, obstetric, and long-term health outcome after uterus transplantation: results of the first clinical trial. *Fertil Steril.* 2022;118:576–585.
21. Fronek J, Kristek J, Chlupac J, et al. Human uterus transplantation from living and deceased donors: the interim results of the first 10 cases of the Czech trial. *J Clin Med*. 2021;10:586.
22. Angelini A, Andersen CB, Bartoloni G, et al. A web-based pilot study of inter-pathologist reproducibility using the ISHLT 2004 working formulation for biopsy diagnosis of cardiac allograft rejection: the European experience. *J Heart Lung Transplant.* 2011;30:1214–1220.
23. Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. *Commun Methods Meas.* 2007;1:77–89.
24. Krippendorff K. *Content Analysis: An Introduction to its Methodology*. 3rd ed. SAGE; 2013.
25. Chen Y, Zee J, Smith A, et al. Assessment of a computerized quantitative quality control tool for whole slide images of kidney biopsies. *J Pathol.* 2021;253:268–278.
26. Royal V, Zee J, Liu Q, et al. Ultrastructural characterization of proteinuric patients predicts clinical outcomes. *J Am Soc Nephrol.* 2020;31:841–854.
27. Gwet KL. *Handbook of Inter-Rater Reliability*. 4th ed. Advanced Analytics, LLC; 2014:410.
28. Wongpakaran N, Wongpakaran T, Wedding D, et al. A comparison of Cohen's kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Med Res Methodol.* 2013;13:61.
29. Brannstrom M, Johannesson L, Bokstrom H, et al. Livebirth after uterus transplantation. *Lancet.* 2015;385:607–616.
30. Hamidian Jahromi A, Horen SR, Johannesson L. Plasma cell-free DNA as a novel method for early detection of acute rejection in uterine transplant. *Exp Clin Transplant.* 2022;20:440–441.
31. Bhorade SM, Husain AN, Liao C, et al. Interobserver variability in grading transbronchial lung biopsy specimens after lung transplantation. *Chest.* 2013;143:1717–1724.
32. Rosales IA, Foreman RK, DeFazio M, et al. Systematic pathological component scores for skin-containing vascularized composite allografts. *Vasc Compos Allotransplant*. 2018;3:62–74.
33. Kreft A, Mottok A, Mesteri I, et al; Gastrointestinal Pathology Group of the German-Austrian-Swiss GvHD Consortium. Consensus diagnostic histopathological criteria for acute gastrointestinal graft versus host disease improve interobserver reproducibility. *Virchows Arch.* 2015;467:255–263.