Seminars in Arthritis and Rheumatism 80 (2026) 153041

Contents lists available at ScienceDirect

Seminars in Arthritis and Rheumatism

journal homepage: www.elsevier.com/locate/semarthrit

# Interobserver variability in synovial tissue biopsy markers in Rheumatoid Arthritis: Results from OMERACT special interest group on synovial tissue biopsy

Mario Ferraioli [a,b,*], Annabelle Small [c], Stefano Alivernini [d], João Eurico Fonseca [e], Søren Andreas Just [f], Walter Maksymowych [g], Fraser Morton [a], Ioana Nicorescu [h], Carl Orr [i], Simone Perniola [j], Arthur Pratt [h], Vasco C. Romão [e], Marleen G.H. van de Sande [k], Vibeke Strand [l], Sander W. Tas [k], Douglas Veale [m], Mihir D Wechalekar [c,1], Aurélie Najm [a,1], on behalf of OMERACT Synovial Tissue Biopsy Special Interest Group

[a] School of Infection and Immunity, College of Medical Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom
[b] Rheumatology, Allergology and Clinical Immunology, Department of Systems Medicine, University of Rome Tor Vergata, Rome, Italy
[c] Rheumatology Synovial Tissue Translational Research Group, College of Medicine & Public Health, Flinders University, and Department of Rheumatology, Flinders Medical Centre, Adelaide, Australia
[d] Division of Rheumatology and Clinical Immunology - Fondazione Policlinico Universitario A. Gemelli IRCCS - Università Cattolocia del Sacro Cuore
[e] Rheumatology Department, Faculdade de Medicina, Universidade de Lisboa and ULS Santa Maria, Lisbon, Portugal
[f] Section of Rheumatology, Department of Medicine, Svendborg Hospital - Odense University Hospital, Svendborg, Denmark
[g] Department of Medicine, University of Alberta, Edmonton, Canada
[h] Translational and Clinical Research Institute and NIHR Newcastle Biomedical Research Centre, Newcastle University and Newcastle Hospitals NHS Foundation Trust, Newcastle upon Tyne, United Kingdom
[i] Centre for Arthritis and Rheumatic Disease, University College Dublin, Dublin, Ireland
[j] Rheumatology Unit, Department of Precision and Regenerative Medicine and Ionian Area (DiMePRe-J), University of Bari, Italy
[k] Department of Rheumatology & Clinical Immunology and Department of Experimental Immunology, Amsterdam UMC, Amsterdam Institute for Immunology and Infectious Diseases, University of Amsterdam, Amsterdam, Netherlands
[l] Division of Immunology/Rheumatology, Stanford University School of Medicine, Palo Alto, CA, USA
[m] Dublin Academic Medical Centre, Centre for Arthritis and Rheumatic Disease, St Vincent's University Hospital, Dublin, Ireland

ARTICLE INFO

*Keywords:*
Synovial tissue
OMERACT
Histology
Standardization
Agreement
Experience

ABSTRACT

*Objectives:* to evaluate interobserver agreement in synovial tissue (ST) biopsy histology and immunohistochemistry (IHC) analysis across expert centers, and the influence of observers' experience on interobserver agreement.

*Methods:* ST slides, collected from 38 biologic-naïve rheumatoid arthritis patients, stained for H&E, FVIII, CD68, CD3 and CD19/20, were uploaded to an online platform. Using semiquantitative analysis, independent observers scored ST slides for immune and vascularity markers (0 to 4), and Krenn's synovitis score (0 to 9). Presence of lymphoid aggregates and synovial pathotype were also recorded. After the initial scoring, a modified eDelphi process was conducted to reach consensus on a shared scoring approach among observers, followed by a second scoring.

Interobserver agreement was assessed with Kappa coefficients.

*Results:* Initial scoring showed variability across observers and centres, with higher agreement for immune markers (e.g. kappa for CD19/20=0.737; CD3=0.713) compared to vascular ones (FVIII=0.392, H&E = 0.316). Stratifying by observers' experience revealed lower agreement among highly experienced observers compared with low and intermediate-experienced observers on specific markers.

After discussion, a modified eDelphi process was conducted, and consensus on adopting a maximum-score approach (the highest score observable on the slide as the final scoring) for semiquantitative parameters was

* Corresponding author.
*E-mail address:* ferraioli.mar@gmail.com (M. Ferraioli).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.semarthrit.2026.153041

Available online 30 June 2026

reached. A rescoring exercise, applying this principle, demonstrated improved agreement among observers, across most parameters (CD68= 0.684, Krenn's score 0.545), although variability persists for certain parameters.
*Conclusion:* This study demonstrates that consensus-based methods can improve interobserver agreement in ST histology and IHC scoring across centres. Adoption of a maximum-score approach among observers represents a further step toward standardization of ST biopsy analysis, and may support its broader application
*Statement of clinical significance:* Histological and immunohistochemical analysis of synovial tissue (ST) biopsy is a promising tool in rheumatoid arthritis (RA), both in translational and clinical research. Notably, the differences in ST biopsy evaluation and interpretation persist across different centres of excellence hampers its widespread clinical application and use in multicentre trials. Our study demonstrates that consensus-based approaches, particularly the adoption of a maximum-score principle, can improve interobserver agreement of ST scoring across centres. Our findings represent a critical step toward the standardization of ST biopsy analysis.

## Introduction

Rheumatoid arthritis (RA) is a chronic autoimmune disease primarily affecting synovial joints, often leading to severe disability and premature mortality [1]. Over the past 20 years, drugs with different mechanisms of action have expanded our therapeutic arsenal. However, a considerable number of patients still do not achieve remission, and the assessment of disease activity remains highly dependent on the evaluation methods used [2,3]. Notably, considering the lack of validated biomarkers, treatment strategies are still based on a "trial and error" approach.

To overcome this unmet need in RA, synovial tissue (ST) biopsy analysis is an attractive tool for biomarkers discovery, to improve treatment responses of targeted therapeutics, and to assess prognosis [4]. Although various clinical trials have recently applied cellular and molecular analyses of ST biopsy, using cutting edge technology to stratify patients to different treatments [5,6], a more widely applicable approach for routine practice is the histological analysis of ST [like haematoxylin/eosin (H&E) and immunohistochemistry (IHC)]. Histology and IHC have already been utilised to stratify disease [7,8], and predict therapy outcomes [8,9], with some success. Furthermore, histology and IHC are still relevant in ST biopsy research contexts. Indeed, not only do they provide simple and useful insights on quality control, but they are also useful to select region of interest before applying higher-end techniques, such as imaging mass cytometry and spatial transcriptomics. However, standardisation of the methodology of interpretation of ST histology and IHC is critical to ensure reproducibility [8,10]. Indeed, substantial differences in ST biopsy evaluation and interpretation persist across different centres of excellence.

Several steps have been undertaken to address these issues. Rooney et al. [11] evaluated interobserver agreement for 2 IHC markers (CD3 and CD68) amongst 6 international centres, and obtained satisfactory results. Efforts by the OMERACT ST biopsy Special Interest Group (SIG), have progressively advanced the field of ST biopsy standardisation: firstly, minimally invasive ultrasound-guided synovial biopsy techniques were validated [12], following which, in collaboration with the EULAR Synovitis Study Group (ESSG), consensus recommendations on ST biopsy handling and analysis were published [13] as well as minimal reporting requirements in ST research [10]. The next step, defined during OMERACT 2018 by the OMERACT ST biopsy SIG, was to provide further insights into the analysis of ST histological markers in order to evaluate:

1) the interobserver variability in ST staining and analysis across expert centres and,
2) the influence of observers' experience on interobserver agreement.

## Materials and methods

According to OMERACT methodology, the work was conducted in 4 phases as follows:

### *Digital slide collection*

Each participating excellence centre was asked to contribute digital images of ST biopsy, collected with image-guided techniques [10], from patients affected by RA (fulfilling the ACR/EULAR 2010 criteria [1]), prior to the start of the first biologic disease modifying anti-rheumatic drug (bDMARDs). For each patient, 5 different images corresponding each to a different IHC staining were uploaded on an online platform (*carearthritis.com*). Staining was performed based on a priorly shared standardised protocol and included (Fig. 1): H&E and Factor VIII (FVIII) for vascularity, CD68, CD3 and CD19/20 for immune infiltrates (macrophages, T- and B-cells, respectively). Following quality control and exclusion of 10 images (5.2%), 180 ST biopsy slides were included.

### *Blinded online scoring exercise*

Independent observers (all rheumatologists/immunologists) scored each slide online, using a 5-point semiquantitative analysis scoring system (0–4), for immune [12–15] and vascularity [18,19] markers, a pre-defined atlas was also provided [20]. Centres with multiple observers scored independently, blinded to each other's results. Slides were also graded for synovitis severity according to Krenn et al. [21], assigning a single score (0–3) to its determinants (hyperplasia/enlargement of synovial lining cell layer; activation of resident cells/synovial stroma; inflammatory infiltrate), to a total score of 0–9. Observers also indicated presence/absence of lymphoid aggregates, as well as synovial phenotype (pathotype) according to Dennis et al. (lymphoid; myeloid; low inflammatory and fibroid) [7]. Results were presented at OMERACT 2023.

### *Delphi exercise*

Following OMERACT 2023, participants convened online to discuss scoring modalities to address existing discrepancies. Slides with the highest scoring heterogeneity were selected according to Shannon's equitability index (Shannon's J) and discussed. Shannon's J quantifies response heterogeneity, as normalised measures derived from Shannon's diversity index (Shannon's H) [22], ranging from 0 (maximum skewedness, no variability) to 1 (maximum evenness, high variability).

Following discussion, a modified Delphi process was initiated to establish consensus on key scoring principles. Items for voting were derived from critical discussion points, and formulated into two statements:

(a) The global score for semiquantitative parameters should reflect the highest score observed in the slide;
(b) A cut off is required to define lymphoid aggregates.

Voting was conducted via a one-stage eDelphi process, in which participants responded anonymously with "yes" or "no" to each item. Items were endorsed if agreement was $\geq$ 70%.

*Blinded online rescoring exercise*

The rescoring exercise used the same slides from the initial scoring round. Based on the eDelphi results, participants applied the maximum score approach. Each slide was rescored for immune [14–17] and vascularity [18,19] markers, as well as for the activation of resident cells/synovial stroma item of the Krenn score [21], using the methodology previously described. These parameters were selected during the online meeting, as the most contentious and requiring re-scoring.

*Statistical analysis*

Descriptive statistics were applied to describe patients' and observers' characteristics.

To assess interobserver variability across the scores performed, Cohen's Kappa was calculated for categorical variables (such as lymphoid aggregates and pathotype), while Cohen's weighted Kappa was used for continuous variables. Notably, Krenn's synovitis severity was analysed both as the total score and categorised as grades of synovitis: none (0–1), low (2–4), or high (5–9) [21].

Kappa values were calculated for both scoring exercises, and compared among matched observers that took part in both rounds. Kappa coefficients were interpreted as follows: poor ($\leq$0.20), fair (>0.20 and $\leq$0.40), moderate (>0.40 and $\leq$0.60), good (>0.60 and $\leq$0.80) and very good (>0.80 and $\leq$1.00) [23]. Light's Kappa was calculated for each parameter, and for both scoring exercises, to assess interobserver agreement. Light's Kappa, together with its 95% confidence intervals, was calculated for each parameter, and for both scoring exercises, to assess interobserver agreement.

Lastly, observers were stratified into three groups based on their experience in scoring ST biopsy: low ($\leq$10 years), intermediate (11–20 years) and high (>20 years); details available in Supplementary Table 1. Eventual differences in interobserver agreement across the experience groups were assessed based on results from the first scoring exercise, including scores for vascularity and immune markers, and Krenn's subcategories. For each score, a permuted Kruskal-Wallis test was performed (10,000 permutations/test), with Benjamini-Hochberg correction. Empirical p-values were calculated as the proportion of permuted statistics greater than or equal to the observed statistic. Permuted Dunn's tests (10,000 permutations) were applied for post-hoc comparisons.

Statistical analysis was performed using R 4.4.2 [24], using the "irr" [25] and "FSA'' [26] packages, while ggiraphExtra [27] and ggcorrplot2 [28] packages were used to produce graphs. $p < 0.05$ was considered as statistically significant.

*Ethics*

All participating centres obtained local ethical approval related to this study, in accordance with institutional and national regulations.

## Results

*Blinded online scoring exercise*

11 observers from 7 different centres from 3 continents participated in the first scoring exercise. ST biopsy slides were collected from 38 patients with RA (demographic and clinical data depicted in Supplementary Table 2).

Overall, correlation analysis showed moderate-good agreement amongst observers for different histology and IHC markers, that also reflected in the agreement analysis (Table 1). In particular, agreement for immune markers such as CD19/20 and CD3 achieved "good" or "very good". For these markers, Light's Kappa values resulted above 0.7. On the other hand, both vascularity staining (by H&E and Factor VIII obtained much lower agreement values (Light's Kappa values of 0.316 and 0.392, respectively).

Degrees of heterogeneity and variations in histology/IHC scores were noted across observers and centres (Fig. 2), reflecting differences in scoring approaches. Indeed, while some parameters (like CD3 and CD19/20) were scored consistently by the observers, greater variability was noted for other parameters (like H&E and FVIII). Remarkably, it appeared as some observers presented a tendency toward higher [like observer 8 (Fig. 2)] or lower [like observer 4 and 5 (Fig. 2)] scores in the semiquantitative analysis, contributing to the observed spread in results.

*Differences in interobserver agreement according to observers' experience*

Interobserver agreement varied among groups based on experience (Table 2). Surprisingly, for CD19 ($p = 0.005$) and CD68 ($p = 0.041$), highly experienced observers had lower agreement when compared with those less experienced. Similar results were obtained when comparing the high with the intermediate experience group for CD3 ($p = 0.001$), CD68 ($p = 0.015$), H&E vascularity ($p = 0.007$), hyperplasia of lining layer ($p = 0.007$) and inflammatory infiltrate ($p = 0.047$). Likewise, the latter also resulted statistically different between low and intermediate experience groups ($p = 0.047$). No other significant differences were observed for the other comparisons made, as well as for FVIII.

*Delphi exercise*

The differences observed during the first scoring exercise highlighted the need for a more homogeneous and shared approach to semiquantitative analysis. Hence, the SIG held an eDelphi survey aiming at standardizing the scoring framework. Voting results showed 77.78% agreement for adopting a "maximum score approach" for semiquantitative analysis, that is: each parameter's final score corresponds to the highest value observable. Similarly, there was 100% agreement to define a cutoff for lymphoid aggregates. Consequently, the SIG endorsed the "maximum score approach" for semiquantitative analysis of histology/IHC scores, while further work on lymphoid aggregates' definition

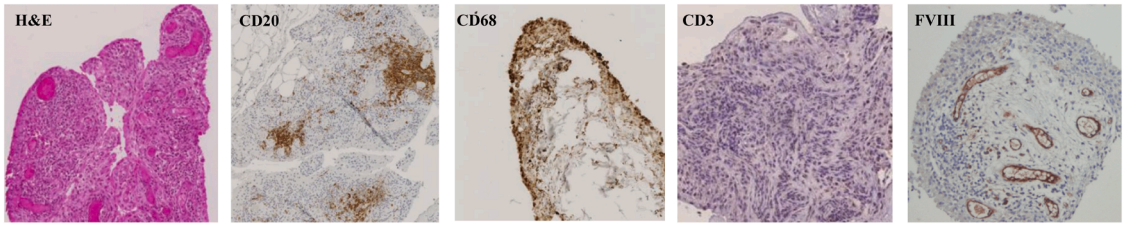


**Fig. 1.** Representative images of synovial tissues showing the stainings that were scored by observers: Haematoxylin/Eosin (H&E), CD20, CD68, CD3 and Factor VIII (FVIII).

**Table 1**
Analysis of interobserver agreement from the first scoring exercise, involving 11 observers, and calculated using unweighted (for Lymphoid Aggregates and Pathotype) and weighted (for all others) Cohen's Kappa for correlation coefficients, and Light's Kappa.

| | Cohen's Kappa correlation coefficients | | | | | Light's Kappa (95%CI) |
|---|---|---|---|---|---|---|
| | Poor | Fair | Moderate | Good | Very Good | |
| **CD19/20** | 0 (0%) | 0 (0%) | 7 (12.7%) | 27 (49.1%) | 21 (38.2%) | 0.737 (0.628–0.795) |
| **CD3** | 0 (0%) | 1 (1.9%) | 9 (16.7%) | 34 (63%) | 10 (18.5%) | 0.713 (0.609–0.777) |
| **CD68** | 0 (0%) | 4 (7.3%) | 12 (21.8%) | 34 (61.8%) | 5 (9.1%) | 0.653 (0.562–0.72) |
| **FVIII** | 5 (9.1%) | 25 (45.5%) | 20 (36.4%) | 5 (9.1%) | 0 (0%) | 0.392 (0.247–0.517) |
| **H&E vascularity** | 12 (21.8%) | 29 (52.7%) | 13 (23.6%) | 1 (1.8%) | 0 (0%) | 0.316 (0.185–0.412) |
| **Lymphoid Aggregates** | 0 (0%) | 1 (1.8%) | 15 (27.3%) | 30 (54.5%) | 9 (16.4%) | 0.675 (0.537–0.796) |
| **Pathotype** | 10 (18.2%) | 23 (41.8%) | 20 (36.4%) | 2 (3.6%) | 0 (0%) | 0.344 (0.261–0.422) |
| **Total Krenn's score (0–9 scale)** | 1 (1.8%) | 10 (18.2%) | 18 (32.7%) | 26 (47.3%) | 0 (0%) | 0.535 (0.441–0.597) |
| Activation of resident cells/stroma | 14 (25.5%) | 21 (38.2%) | 18 (32.7%) | 2 (3.6%) | 0 (0%) | 0.341 (0.231–0.43) |
| Hyperplasia of Lining Layer | 9 (16.4%) | 32 (58.2%) | 13 (23.6%) | 1 (1.8%) | 0 (0%) | 0.326 (0.205–0.421) |
| Inflammatory Infiltrate | 3 (5.5%) | 10 (18.2%) | 20 (36.4%) | 20 (36.4%) | 2 (3.6%) | 0.525 (0.416–0.601) |
| **Total Krenn's score (categorised)** | 4 (7.3%) | 13 (23.6%) | 22 (40.0%) | 16 (29.1%) | 0 (0%) | 0.474 (0.361–0.568) |

H&E: Hematoxylin/Eosin.

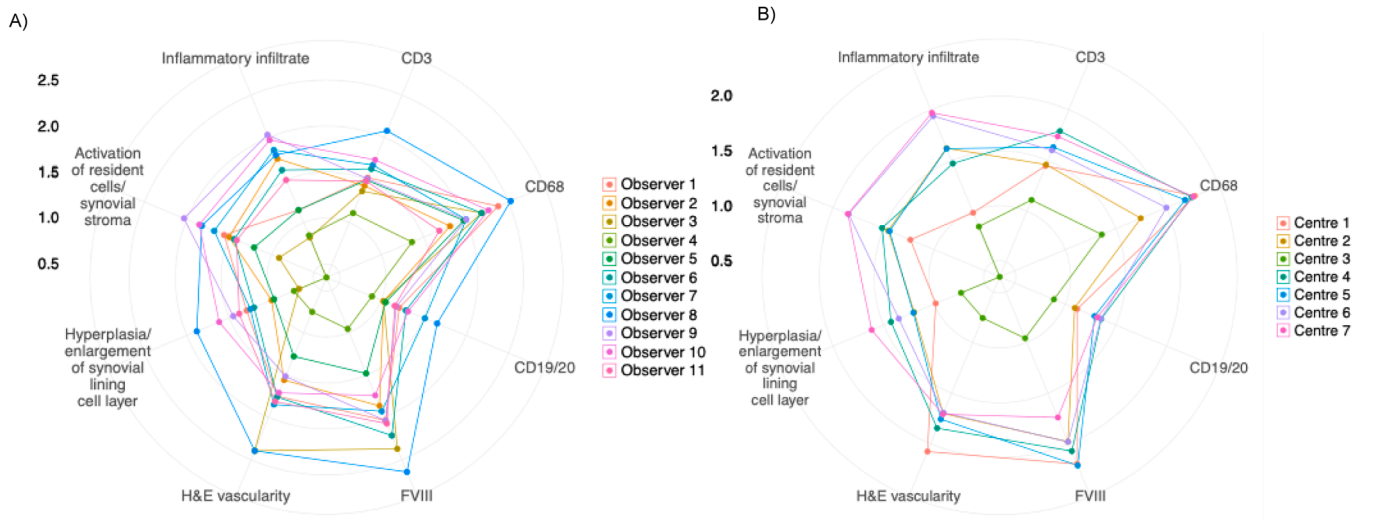


**Fig. 2.** Radar plot of the mean value for each score across all slides in the initial scoring exercise, depicted for all observers (panel A) and according to expert centres (panel B).

was planned.

*Second online scoring exercise*

Eight observers, from 6 centres, who participated in the first scoring exercise, took part in the rescoring exercise. Agreement improved between the first and second scores (correlation coefficients in Table 3). Notably, this improvement included activation of resident cells/stroma (64.3%), an outcome that also affected the total Krenn's score (42.9%). There were similar improvements with regards to the vascularity markers FVIII and H&E (60.7%, and 53.6%, respectively), as well as CD68 (60.7%).

The rescoring exercise led to improvements in the interobserver agreements for almost all parameters rescored. Looking at Light's Kappa values for matched observers who took part in both scores, these were particularly improved for activation of resident cells/stroma, H&E and CD68 (Supplementary Table 3).

## Discussion

In this manuscript we investigated interobserver variability in histological and IHC analysis of ST biopsy, recognising that these techniques are fundamental for ST analysis in the era of spatial transcriptomics. The presented study was conducted on ST biopsy from a large cohort of bDMARD-naïve patients with RA.

To the best of our knowledge, this represents the first study to systematically evaluate the inter-observer agreement for histological and IHC ST markers by a high number of international experts and centers.

Results show a substantial level of agreement between observers, particularly for CD19/20, CD3 and CD68 markers. Despite a strong agreement, we observed discrepancies on specific IHC markers which led to further discussions. Indeed, through a structured eDelphi exercise, consensus was reached on scoring methods, which subsequently led to improvement in the interobserver agreement.

Previous efforts to address interobserver variability in ST biopsy analysis have been limited. Rooney et al. achieved a strong interobserver agreement for synovial sublining CD3+ and CD68+ infiltration using semiquantitative analysis [11], though their study encompassed a smaller patient population. Mussawy et al. found a comparable Cohen's kappa of 0.712 between 2 observers for H&E staining, in a 20-patient arthritis cohort that included 8 RA patients [29]. Our results appear more robust than those available in the literature, as they are based not only on a larger cohort, but especially since they include multiple histology and IHC markers and various other scores, fully implementing the EULAR/OMERACT consensus on the set of items needed for ST biopsy analysis [13].

Through discussion and an eDelphi process, we were able to reach agreement on scoring methods and add supplementary items to the

**Table 2**
Comparison of interobserver agreement among observers who participated in the first online score, divided according to their years of experience in scoring synovial tissue biopsy. That is 1) low: 10 years or less, 2) intermediate: between 11 years and up to 20 years, and 3) high: >20 years.

| | | Light's Kappa (95%CI) | | | | Post-hoc Dunn's test | | |
|---|---|---|---|---|---|---|---|---|
| | | Experience | | | Kruskal-Wallace permutated adjusted p-value: | Observed Z | Permutated p value (unadjusted) | Permutated p value (adjusted) |
| | | Low | Intermediate | High | | | | |
| **CD19/20** | | 0.838 (0.717–0.895) | 0.783 (0.67–0.851) | 0.597 (0.448–0.7) | **0.023** | | | |
| experience | low vs intermediate | | | | | 0.903 | 0.399 | 0.399 |
| | low vs high | | | | | 2.740 | 0.001 | **0.005** |
| | high vs intermediate | | | | | 2.002 | 0.040 | 0.0613 |
| **CD3** | | 0.737 (0.603–0.819) | 0.797 (0.648–0.884) | 0.539 (0.401–0.642) | **0.023** | | | |
| experience | low vs intermediate | | | | | −1.742 | 0.084 | 0.126 |
| | low vs high | | | | | 1.528 | 0.130 | 0.130 |
| | high vs intermediate | | | | | 2.951 | <0.0001 | **0.001** |
| **CD68** | | 0.732 (0.608–0.818) | 0.778 (0.685–0.837) | 0.454 (0.302–0.596) | **0.023** | | | |
| experience | low vs intermediate | | | | | −0.516 | 0.650 | 0.650 |
| | low vs high | | | | | 2.160 | 0.027 | **0.041** |
| | high vs intermediate | | | | | 2.582 | 0.005 | **0.015** |
| **FVIII** | | 0.425 (0.238–0.584) | 0.441 (0.195–0.603) | 0.259 (0.078–0.417) | 0.345 | | | |
| experience | low vs intermediate | | | | | −0.387 | 0.727 | 0.727 |
| | low vs high | | | | | 1.159 | 0.266 | 0.4 |
| | high vs intermediate | | | | | 1.475 | 0.147 | 0.4 |
| **H&E vascularity** | | 0.343 (0.185–0.448) | 0.41 (0.221–0.55) | 0.109 (0.011–0.211) | **0.023** | | | |
| experience | low vs intermediate | | | | | −0.774 | 0.452 | 0.452 |
| | low vs high | | | | | 2.054 | 0.037 | 0.056 |
| | high vs intermediate | | | | | 2.687 | 0.002 | **0.007** |
| **Activation of resident cells/stroma** | | 0.227 (0.049–0.38) | 0.503 (0.32–0.646) | 0.326 (0.234–0.406) | **0.023** | | | |
| experience | low vs intermediate | | | | | −2.646 | 0.004 | **0.014** |
| | low vs high | | | | | −0.685 | 0.513 | 0.513 |
| | high vs intermediate | | | | | 1.475 | 0.155 | 0.233 |
| **Hyperplasia of lining layer** | | 0.379 (0.22–0.505) | 0.44 (0.211–0.625) | 0.143 (0.001–0.284) | **0.023** | | | |
| experience | low vs intermediate | | | | | −0.774 | 0.462 | 0.462 |
| | low vs high | | | | | 2.055 | 0.037 | 0.055 |
| | high vs intermediate | | | | | 2.687 | 0.002 | **0.007** |
| **Inflammatory Infiltrate** | | 0.445 (0.296–0.558) | 0.658 (0.535–0.744) | 0.406 (0.265–0.528) | **0.035** | | | |
| experience | low vs intermediate | | | | | −2.130 | 0.031 | **0.047** |
| | low vs high | | | | | 0.316 | 0.774 | 0.774 |
| | high vs intermediate | | | | | 2.055 | 0.031 | **0.047** |

research agenda. Indeed, the Delphi-based approach provided valuable insights into experts' scoring methods. In particular, the method used for semiquantitative scoring varied widely across observers due to use of either maximum scoring or average scoring across one slide. Following the eDelphi survey, the "maximum score approach" was adopted by the SIG, establishing a common scoring paradigm among observers. Experts agreed that the final grading of each semiquantitative score should reflect the highest available on a ST slide. The "maximum score approach" can offer a more representative reflection of synovial disease severity, and by leveraging onto the most pronounced histological features, it might enhance patient-level representativity by better capturing aggressive synovial changes. At the same time, "maximum score approach", resembling scoring paradigms used in oncology, may speculatively inform therapy choices, although this potential requires further investigation.

Results from the second score, performed applying the "maximum score approach", show an improvement in correlation analysis, as well as in Light's kappa values demonstrating a consistent, although in some cases modest, increase in interobserver agreement values. Importantly, the observed effect of the consensus-based approach, represents evidence that standardisation of scoring methodology can lead to measurable and directionally consistent improvements across different

**Table 3**
Correlation and agreement analysis for the second scoring exercise, with percentage of improved correlations confronted with the first score on matched observers. H&E: Hematoxylin/Eosin.

| | Cohen's Kappa interpretation | | | | | Impro-vement | Light's Kappa (95%CI) |
|---|---|---|---|---|---|---|---|
| | Poor | Fair | Moderate | Good | Very Good | | |
| **CD19/20** | 0 (0%) | 0 (0%) | 3 (10.7%) | 13 (46.4%) | 12 (42.9%) | **46.4%** | **0.759 (0.621–0.84)** |
| **CD3** | 0 (0%) | 0 (0%) | 4 (14.3%) | 18 (64.3%) | 6 (21.4%) | **53.6%** | **0.712 (0.611–0.788)** |
| **CD68** | 0 (0%) | 0 (0%) | 8 (28.6%) | 15 (53.6%) | 5 (17.9%) | **60.7%** | **0.684 (0.57–0.759)** |
| **FVIII** | 7 (25.0%) | 11 (39.3%) | 8 (28.6%) | 2 (7.1%) | 0 (0%) | **60.7%** | **0.342 (0.182–0.478)** |
| **H&E vascularity** | 5 (17.9%) | 12 (42.9%) | 9 (32.1%) | 2 (7.1%) | 0 (0%) | **53.6%** | **0.368 (0.233–0.474)** |
| **Krenn's Synovitis score (0–9 scale)** | 1 (3.6%) | 3 (10.7%) | 12 (42.9%) | 12 (42.9%) | 0 (0%) | **42.9%** | **0.545 (0.438–0.627)** |
| Activation of resident cells/stroma | 8 (28.6%) | 11 (39.3%) | 9 (32.1%) | 0 (0%) | 0 (0%) | **64.3%** | **0.322 (0.205–0.413)** |
| **Krenn's Synovitis score (categorised)** | 1 (3.6%) | 8 (28.6%) | 12 (42.9%) | 7 (25.0%) | 0 (0.0%) | **53.6%** | **0.472 (0.336–0.575)** |

expert centres.

Finally, the SIG emphasised how the "maximum score approach" should be conducted with care to any overestimation, that might represent a potential limitation. To ensure objectivity and fairness, the SIG recommends to evaluate multiple slides per patient, with different IHC staining conducted on serial sections and multi-site sampling per patient, as recommended in previous OMERACT led standardisation efforts [13].

Furthermore, tissue quality appeared in this work as a major driver of agreement, emphasizing the need for appropriate quality controls, as previously reported [10,13,30]. Notably, in present study, ST were processed across different centres and their image files were shared in different formats. In line with quality control principles, a standardised protocol was circulated prior to collection to all centres. However, these elements might have influenced agreement, and should be considered as an intrinsic limitation. Additionally, clinical data were available for 70% of included patients; yet, considering the focus of present study, and its scope as a methodological scoring exercise, impact of missing data on main outcomes is likely minimal.

Lastly, this study considered observers' experience and assessed its influence on interobserver agreement. Interestingly, while longer experience would conceivably be associated with higher agreement, we found results to the contrary, for multiple parameters. Similar findings have been reported in other fields. Comparable rates of interobserver agreement have been described among experienced and inexperienced pathologists [31], and even between senior and junior radiologists [32]. In another case junior pathologists even out-performed senior ones in agreement levels [33].

A plausible explanation for this discrepancy, is that observers' experience shape divergent approaches to ST biopsy analysis. Less experienced observers may adhere more strictly to scoring guidelines or atlases. Conversely, highly experienced observers may have developed - during the years - a more refined but possibly divergent perspective on ST scoring, based on personal interpretive habits developed over time. Hence, the variability observed in the highly experienced group can be based on this divergence, especially considering the semiquantitative nature of performed scores, where subjective judgement plays a role. Another possible explanation might involve centre-specific characteristics like scoring habits or volume of ST biopsies locally handled. However, all experts come from large third-level centres and – apart from 1 centre – there are 2 observers from each of them – that also have different experience levels among one another. Thus, even though these differences may have played a role, their effect appears minimised.

It would have been interesting to perform another analysis on the second scoring, however the dropout rate in the most experienced observers' group did not allow this.

The consensus at OMERACT 2025 following the presentation of these results was the need for integration of artificial intelligence (AI) technology into ST biopsy analysis as a key research priority. Indeed, digital image analysis (DIA) techniques have been previously applied to ST biopsy analysis [11,34], finding good correlation with semiquantitative analysis, but have not been validated furtherly. Similarly, a machine learning approach recently achieved good correlation with quantitative assessment of synovial cells on H&E staining [35]

AI technology can potentially further improve scoring of ST biopsy, potentially reducing interobserver variability by automating the quantification of ST histology and IHC, aiding selection of representative tissue regions and generating reproducible metrics. These could conceivably enable patient stratification with regards to the degree and type of inflammation and improve prediction of therapeutic outcomes.

The next steps for the ST biopsy SIG include validation of AI-assisted approaches for ST biopsy analyses.

## Conclusion

Considering the high potential of ST biopsy in clinical trials and clinical practice, continued efforts are needed to standardise ST biopsy analysis, in an effort to improve both utility and reproducibility. Our study highlights the need for common scoring methods across observers and centres, and demonstrates a framework to improve interobserver agreement through scoring, discussion about discrepancies and agreement through a Delphi survey, followed by a rescoring exercise. This work from the OMERACT ST biopsy SIG represents a further step towards standardization of ST biopsy analysis, including the future aim to integrate AI technology into the workflow.

## CRediT authorship contribution statement

**Mario Ferraioli**: conceptualization, methodology, formal analysis, investigation, data curation, writing – original draft, writing - review & editing, visualization. **Annabelle Small**: investigation, writing - review & editing. **Stefano Alivernini**: investigation, resources, writing - review & editing. **João Eurico Fonseca**: conceptualization, methodology, validation, investigation, resources, writing - review & editing. **Søren Andreas Just**: methodology, validation, investigation, resources, writing - review & editing. **Walter Maksymowych**: methodology, software, resources, writing - review & editing. **Fraser Morton**: conceptualization, methodology, formal analysis, data curation, writing - review & editing, visualization. **Ioana Nicorescu**: conceptualization, methodology, writing - review & editing. **Carl Orr**: validation, investigation, resources, writing - review & editing. **Simone Perniola**: investigation, writing - review & editing. **Arthur Pratt**: conceptualization, investigation, writing - review & editing. **Vasco C. Romão**: investigation, resources, writing - review & editing. **Marleen G.H. van de Sande**: conceptualization, investigation, resources, writing - review & editing. **Vibeke Strand:** conceptualization, writing - review & editing, project administration. **Sander W. Tas**: methodology, investigation, resources, writing - review & editing. **Douglas Veale**: conceptualization, methodology, investigation, resources. **Mihir D. Wechalekar**: conceptualization, methodology, validation, investigation, formal analysis, resources, writing - review & editing, supervision, project administration. **Aurélie Najm**: conceptualization, methodology, validation, formal analysis, investigation, resources, writing - review & editing, supervision, project administration, funding acquisition.

## Declaration of generative AI in scientific writing

During the preparation of this work the author(s) used AI-assisted technology in order to improve the readability under strict human oversight and control. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

## Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Marleen G.H. van de Sande reports a relationship with Beneke El that includes: speaking and lecture fees. Marleen G.H. van de Sande reports a relationship with Eli Lilly that includes: speaking and lecture fees. Marleen G.H. van de Sande reports a relationship with Novartis that includes: consulting or advisory, funding grants, and speaking and lecture fees. Marleen G.H. van de Sande reports a relationship with UCB that includes: consulting or advisory, funding grants, and speaking and lecture fees. Marleen G.H. van de Sande reports a relationship with MSD that includes: funding grants. Marleen G.H. van de Sande reports a relationship with Abbvie that includes: consulting or advisory. Marleen G.H. van de Sande reports a relationship with Janssen that includes: consulting or advisory, funding grants, and speaking and lecture fees. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

The authors wish to thank patients research partners - acknowledging their distinctive contribution within the OMERACT community - whose ongoing engagement continues to shape research priorities and ensure that outcomes remain meaningful to patients. Furthermore, the authors acknowledge the fundamental work of Mr. Joel Paschke and the IT team at CARE Arthritis for developing the online scoring platform.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.semarthrit.2026.153041.

## Data availability

Dataset used to generate data included in this manuscript will be made available at reasonable request.

## References

[1] Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO, et al. 2010 rheumatoid arthritis classification criteria: an American college of rheumatology/European league against rheumatism collaborative initiative. Ann Rheum Dis 2010;69(9):1580–8. https://doi.org/10.1136/ard.2010.138461.

[2] Koh JH, Lee Y, Kim HA, Kim J, Shin K. Comparison of remission criteria in patients with rheumatoid arthritis treated with biologic or targeted synthetic disease-modifying anti-rheumatic drugs: results from a nationwide registry. Ther Adv Musculoskelet 2022;14:1759720×221096363. https://doi.org/10.1177/1759720×221096363.

[3] Guo N, Li X, Movahedi M, Cesta A, Bombardier C. Biologics initiation in moderate versus severe rheumatoid arthritis: real-world experience from a Canadian registry. Clin Exp Rheumatol 2024 May;42(5):1067–74. https://doi.org/10.55563/clinexprheumatol/mpevid.

[4] Wechalekar MD, Najm A, Veale DJ, Strand V. The 2018 OMERACT synovial tissue biopsy special interest group report on standardization of synovial biopsy analysis. J Rheumatol 2019;46(10):1365–8. https://doi.org/10.3899/jrheum.181062.

[5] Rivellese F, Surace AEA, Goldmann K, Sciacca E, Çubuk C, Giorli G, et al. Rituximab versus tocilizumab in rheumatoid arthritis: synovial biopsy-based biomarker analysis of the phase 4 R4RA randomized trial. Nat Med 2022;28(6):1256–68. https://doi.org/10.1038/s41591-022-01789-0.

[6] Rivellese F, Nerviani A, Giorli G, Warren L, Jaworska E, Bombardieri M, et al. Stratification of biological therapies by pathobiology in biologic-naive patients with rheumatoid arthritis (STRAP and STRAP-EU): two parallel, open-label, biopsy-driven, randomised trials. Lancet Rheumatol 2023;5(11):e648–59. https://doi.org/10.1016/S2665-9913(23)00241-2.

[7] Dennis G, Holweg CT, Kummerfeld SK, Choy DF, Setiadi AF, Hackney JA, et al. Synovial phenotypes in rheumatoid arthritis correlate with response to biologic therapeutics. Arthritis Res Ther 2014;16(2):R90. https://doi.org/10.1186/ar4555.

[8] Alivernini S, Tolusso B, Gessi M, Gigante MR, Mannocci A, Petricca L, et al. Inclusion of synovial tissue–derived characteristics in a nomogram for the prediction of treatment response in treatment-naive rheumatoid arthritis patients. Arthritis Rheumatol 2021;16(9):1601–13. https://doi.org/10.1186/ar4555.

[9] Nerviani A, Di Cicco M, Mahto A, Lliso-Ribera G, Rivellese F, Thorborn G, et al. A Pauci-immune synovial pathotype predicts inadequate response to tnfα-blockade in rheumatoid arthritis patients. Front Immunol 2020;16:845. https://doi.org/10.1186/ar4555.

[10] Najm A, Costantino F, Alivernini S, Alunno A, Bianchi E, Bignall J, et al. EULAR points to consider for minimal reporting requirements in synovial tissue research in rheumatology. Ann Rheum Dis 2022;81(12):1640–6. https://doi.org/10.1136/annrheumdis-2021-221875.

[11] Rooney T, Bresnihan B, Andersson U, Gogarty M, Kraan M, Schumacher HR, et al. Microscopic measurement of inflammation in synovial tissue: inter-observer agreement for manual quantitative, semiquantitative and computerised digital image analysis. Ann Rheum Dis 2007;66(12):1656–60. https://doi.org/10.1136/ard.2006.061143.

[12] Humby F, Kelly S, Bugatti S, Manzo A, Filer A, Mahto A, et al. Evaluation of minimally invasive, ultrasound-guided synovial biopsy techniques by the OMERACT filter — determining validation requirements. J Rheumatol 2016;43(1):208–13. https://doi.org/10.3899/jrheum.141199.

[13] Najm A, Le Goff B, Orr C, Thurlings R, Cañete JD, Humby F, et al. Correction to: standardisation of synovial biopsy analyses in rheumatic diseases: a consensus of the EULAR synovitis and OMERACT synovial tissue biopsy groups. Arthritis Res Ther 2018;20(1):282. https://doi.org/10.1186/s13075-018-1795-5.

[14] Tak PP, Smeets TJM, Daha MR, Kluin PM, Meijers KAE, Brand R, et al. Analysis of the synovial cell infiltrate in early rheumatoid synovial tissue in relation to local disease activity. Arthritis Rheum 1997;40(2):217–25. https://doi.org/10.1002/art.1780400206.

[15] Tak PP, Taylor PC, Breedveld FC, Smeets TJM, Daha MR, Kluin PM, et al. Decrease in cellularity and expression of adhesion molecules by anti–tumor necrosis factor α monoclonal antibody treatment in patients with rheumatoid arthritis. Arthritis Rheum 1996;39(7):1077–81. https://doi.org/10.1002/art.1780390702.

[16] Tak PP, Thurkow EW, Daha MR, Kluin PM, Smeets TJM, Meinders AE, et al. Expression of adhesion molecules in early rheumatoid synovial tissue. Clin Immunol Immunopathol 1995;77(3):236–42. https://doi.org/10.1006/clin.1995.1149.

[17] Tak PP, Van Der Lubbe PA, Cauli A, Daha MR, Smeets TJM, Kluin PM, et al. Reduction of synovial inflammation after anti-cd4 monoclonal antibody treatment in early rheumatoid arthritis. Arthritis Rheum 1995;38(10):1457–65. https://doi.org/10.1002/art.1780381012.

[18] Kennedy A, Ng CT, Biniecka M, Saber T, Taylor C, O'Sullivan J, et al. Angiogenesis and blood vessel stability in inflammatory arthritis. Arthritis Rheum 2010;62(3):711–21. https://doi.org/10.1002/art.27287.

[19] Najm A, Orr C, Gallagher L, Biniecka M, Gaigneux E, Le Goff B, et al. Knee joint synovitis: study of correlations and diagnostic performances of ultrasonography compared with histopathology. RMD Open 2018;4(1):e000616. https://doi.org/10.1136/rmdopen-2017-000616.

[20] Najm A, Le Goff B, Venet G, Garraud T, Amiaud J, Biha N, et al. IMSYC immunologic synovitis score: a new score for synovial membrane characterization in inflammatory and non-inflammatory arthritis. Jt Bone Spine 2019;86(1):77–81. https://doi.org/10.1016/j.jbspin.2018.04.004.

[21] Krenn V, Morawietz L, Häupl T, Neidel J, Petersen I, König A. Grading of chronic synovitis — a histopathological grading system for molecular and diagnostic pathology. Pathol Res Pr 2002;198(5):317–25. https://doi.org/10.1078/0344-0338-5710261.

[22] Kohn AJ, Pielou EC. Ecological diversity. 7th ed. New York: Wiley & Sons; 1977. https://doi.org/10.4319/lo.1977.22.1.0174b.

[23] Altman DG. Practical statistics for medical research. 1st ed. London; New York: Chapman and Hall; 1991611.–p. https://doi.org/10.1201/9780429258589.

[24] R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. 2024. Available at: https://www.R-project.org/.

[25] Gamer M, Lemon J, Fellows I, Singh P. irr: various coefficients of interrater reliability and agreement. R package version 0.84.1, https://CRAN.R-project.org/package=irr; 2019.

[26] Ogle DH, Doll JC, Wheeler AP, Dinno A. FSA: simple fisheries stock assessment methods. R package version 0.9.5, https://CRAN.R-project.org/package=FSA; 2023.

[27] Moon K. ggiraphExtra: make interactive ggplot2. extension to ggplot2 and ggiraph. R package version 0.3.0, https://CRAN.R-project.org/package=ggiraphExtra; 2020.
[28] Cai J, Matheson G, Samson LD. ggcorrplot2: visualize a correlation matrix using ggplot2. R package version 0.1.2, https://github.com/caijun/ggcorrplot2; 2022.
[29] Mussawy H, Zustin J, Luebke AM, Strahl A, Krenn V, Rüther W, et al. The histopathological synovitis score is influenced by biopsy location in patients with knee osteoarthritis. Arch Orthop Trauma Surg 2021;142(11):2991–7. https://doi.org/10.1007/s00402-021-03889-x.
[30] Najm A, Costantino F, Weill C, Filer A, D'Agostino MA. Impact of synovial biopsy procedures and disease-specific aspects on synovial tissue outcome: a systematic literature review informing the EULAR points to consider for the minimal reporting requirements in synovial tissue research in rheumatology. RMD Open 2022;8(1):e002116. https://doi.org/10.1136/rmdopen-2021-002116.
[31] Rutgers JJ, Bánki T, Van Der Kamp A, Waterlander TJ, Scheijde-Vermeulen MA, Van Den Heuvel-Eibrink MM, et al. Interobserver variability between experienced and inexperienced observers in the histopathological analysis of Wilms tumors: a pilot study for future algorithmic approach. Diagn Pathol 2021;16(1):77. https://doi.org/10.1186/s13000-021-01136-w.
[32] Bai X, Sun SM, Xu W, Kang HH, Li L, Jin YQ, et al. MRI-based Bosniak classification of cystic renal masses, version 2019: interobserver agreement, impact of readers' Experience, and diagnostic performance. Radiology 2020;297(3):597–605. https://doi.org/10.1148/radiol.2020200478.
[33] Alpert L, Setia N, Ko HM, Lagana SM, Pittman ME, Johncilla M, et al. Interobserver agreement and the impact of mentorship on the diagnosis of inflammatory bowel disease–associated dysplasia among subspecialist gastrointestinal pathologists. Virchows Arch 2021;478(6):1061–9. https://doi.org/10.1007/s00428-020-02998-z.
[34] Kraan MC, Haringman JJ, Ahern MJ, Breedveld FC, Smith MD, Tak PP. Quantification of the cell infiltrate in synovial tissue by digital image analysis. Rheumatol (Oxf) 2000;39(1):43–9. https://doi.org/10.1093/rheumatology/39.1.43.
[35] Bell RD, Brendel M, Konnaris MA, Xiang J, Otero M, Fontana MA, et al. Automated multi-scale computational pathotyping (AMSCP) of inflamed synovial tissue. Nat Commun 2024;15:7503. https://doi.org/10.1038/s41467-024-51012-6.