

Supplementary Material 2

Use of artificial-intelligence and automation tools

Review — Definitions and Assessment of Residual Distal Radioulnar Joint Instability After Operative Fixation of Distal Radius Fractures: A Methodological Systematic Review

Prepared 5 August 2026; revised 6 August 2026. Reported in accordance with the *Journal of Musculoskeletal Trauma* Artificial Intelligence Guideline and with PRISMA 2020 item 8 (automation tools).

Tools

	Tool	Manufacturer	Version	Used at
A	Claude, accessed through the Claude desktop application for Windows	Anthropic, San Francisco, CA, USA	Application 1.24012.11 (build 09114b); Claude Opus model series, the release verified on 5 August 2026 being Claude Opus 5	Points 1–3, 5–8
B	ocrmypdf	open source	17.8.1	Point 4
C	Tesseract OCR (English)	open source	5	Point 4
D	R with the <i>readxl</i> package	R Foundation	R 4.3.3, readxl 1.4.3	Point 6

The work described below was carried out between 24 July and 5 August 2026. The application version recorded above is the version installed at the close of the review; the model release in use at each individual working session was not recorded prospectively, and is stated here only for 5 August 2026, the date on which it was verified. This limitation is disclosed rather than resolved retrospectively.

No eligibility decision, no coding decision and no classification decision in this review was made by an automated tool. Every such decision was made by a named human reviewer. No AI-generated image or video appears in the manuscript or in any supplementary file.

Where each use sits in the sequence

The protocol fixes the sequence: independent assessment → **lock** → **calculation of agreement** → disclosure of disagreement → resolution → adjudication → confirmation. The interval between the lock and the calculation is the only one in which an automated use could have changed a reported reliability figure.

	Use	When	Position in the sequence
1	Title-and-abstract pre-screening	Before registration	Before the lock
2	Preliminary full-text reading	Before registration	Before the lock
3	Translation of non-English reports	Full-text assessment	Before the lock
4	Optical character recognition	Full-text assessment	Before the lock
5	Drafting the re-adjudication after rule clarifications 2 and 3	4 Aug 2026	After the lock, before disclosure
6	Checking extraction-file independence; formatting the agreement output	4–5 Aug 2026	Is the calculation, on locked data
7	Drafting the written justification for resolved disagreements	4 Aug 2026	After disclosure
8	Extracting verbatim sentences for the study-design classification	5 Aug 2026	After the lock

No use of an automated tool falls between the lock and the calculation of agreement. Agreement statistics were calculated on locked data before any disagreement was disclosed and were not recalculated after resolution, so no use listed below could have altered a reported reliability figure after the fact. The timestamps for every step, with file hashes, are deposited in the procedural record.

Point 1 · Title-and-abstract pre-screening

When — during protocol development, before registration.

What — 629 of the 695 records deduplicated at that stage were passed to the model, which was given the title and abstract of each record and asked only to identify records that were obviously ineligible. The remaining 66 were screened by Reviewer 1 without assistance. Those 695 records are those of the development search of 28 July 2026; the search of record of 3 August 2026 yielded 697 unique records, and the counts reported in the review are those of the search of record.

Human control — between 1 and 3 August 2026, after registration, Reviewer 1 re-read the title and abstract of all 629 records and judged each independently **while masked to the label produced by the model**. **None of the 629 decisions differed** from the pre-screen. Every title-and-abstract decision used in this review is a decision of Reviewer 1.

Instruction given to the model

For each record you are given a title and an abstract. Identify only those records that are obviously ineligible for a review of residual distal radioulnar joint instability after operative fixation of a distal radius fracture. Do not make an eligibility decision. If there is any doubt, do not mark the record.

Point 2 · Preliminary full-text reading before registration

When — during protocol development, before registration.

What — assistant-supported preliminary reading of 51 reports.

Human control — **none of those decisions was carried forward**. After registration Reviewer 1 assessed every report entering the full-text stage again, on a form in which the decision fields were blank, as the registered protocol requires. No decision differed from the pre-registration record. That Reviewer 1 had seen the earlier values is reported as a limitation of the review.

Point 3 · Translation of non-English reports

When — during full-text assessment.

What — machine-assisted translation of material published in Czech and in Chinese, to allow it to be read. Four of the 30 included reports are not in English: three in Czech [33-35] and one in Chinese [30]. All four were read this way.

Human control — the material on which each coding decision rests is deposited verbatim **in the source language**, alongside the code assigned to it, in Supplementary Table 2 and in the deposited dataset; a reader with the language can therefore check every coding against its source. **Confirmation by a fluent speaker was not sought for any of the four**. For the three Czech reports the definition level was coded *Partial* or *None* — the coding records an **absence** — and an absence is harder to establish through translation than a presence. The Chinese report carries a full English abstract, but every value coded from it rests on the Chinese body text rather than on that abstract. Both points are reported as limitations of the review.

What changed after a rule query — during protocol development, one of these reports [35] was judged to be a duplicate publication of [34] and was placed with the excluded records. After the interpretation issued for rule query 3 (amendment C6) — that the duplicate-report code applies only where a report is **not independently eligible** — it was reassessed, found independently eligible, retained, and linked to [34] as an overlapping pair. It is the report set aside in the first sensitivity analysis. The basis for the original duplicate judgment was the recruitment dates, the institution, the author list and a self-citation, none of which depends on interpretation of the Czech text.

Point 4 · Optical character recognition

When — during full-text assessment.

What — two reports were supplied as image-only scans. For one, a 300 dpi optically recognized version was produced; for the other, an invisible text layer was added with **ocrmypdf 17.8.1 and Tesseract 5 (English)** while the original image layer was left unchanged.

Human control — no content was altered and the originals are retained unmodified. Reading was performed by the reviewers from the original page images.

Point 5 · Drafting the re-adjudication after rule clarifications 2 and 3

When — 4 August 2026, after the two interpretations were issued by the third reviewer and **before any disagreement was disclosed**.

What — drafting of the re-adjudication of the twelve affected reports and compilation of the consensus document.

Human control — **the judgment and the supporting quotation for every one of the twelve reports were verified against the full text by both reviewers, who confirmed each entry**; the confirmation is recorded in the consensus workbook. The model produced no initial independent assessment. The reported agreement statistics were computed and locked **before** either interpretation was issued, so they are unaffected.

Point 6 · Checking extraction-file independence, and formatting the agreement output

When — 4–5 August 2026, on locked data, before disclosure.

What — (i) a check of whether the two extraction files were independently produced; (ii) computation of the agreement statistics with the deposited R script, and formatting of its output.

Human control — the agreement statistics were produced by a deposited R script (R 4.3.3, *readxl* 1.4.3), not by the language model. **The script refuses to run unless both lock confirmations are present**, so no reported figure can have been produced after disclosure. The script and its full output are deposited with the review.

What the check found — the second reviewer's first submission of the first-stage extraction file reproduced the first reviewer's values verbatim. The file was set aside and re-created from the full texts alone. This is disclosed in the review and in the amendment log.

Point 7 · Drafting the written justification for resolved disagreements

When — 4 August 2026, during the face-to-face consensus meeting, **after** the agreement statistics had been calculated and locked.

What — drafting of the written rationale for each of the 131 resolved extraction disagreements, citing the source text and the relevant codebook or protocol clause, and checking of typography and consistency of notation. The consensus workbook contains 132 numbered rows: these 131 disagreements and one definitional decision, included so that both reviewers would confirm it.

Human control — **the final value for every disagreement was chosen by the two reviewers at the meeting**. The model did not select values and was not asked to interpret the source PDFs. Agreement had been calculated and locked before the meeting began and, as the protocol requires, was not recalculated afterwards.

Point 8 · Extracting verbatim sentences for the study-design classification

When — 5 August 2026.

What — extraction from the full texts of the sentences bearing on the direction of recruitment and data collection, for the twelve reports whose study-design item did not indicate direction. These sentences were presented to the reviewers as evidence.

Human control — **the classification of every report was made by the reviewers**, who read the presented sentences against the full text. Two of the twelve automatically ranked candidate sentences were wrong — one described a different study and one was a table fragment — and all twelve were re-selected by hand before being presented. The final assignments were checked against all five clauses of the issued interpretation, with no violation and no retrospective re-classification.

Points at which no automated tool was used

No automated tool took part in the initial independent eligibility assessment by either reviewer; the assignment of any exclusion code; the assignment of any item-level code, including the definition level, the lesion–instability relationship, the level of the positive criterion and the follow-up definition level; the resolution of any disagreement; any adjudication by the third reviewer; or the interpretation of any rule query.