npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-026-02431-w

# Large language models in systematic review and meta-analysis of surgical treatments for vaginal vault prolapse



Yunjeong Park[1,2], Hyun-Soo Zhang[3] & Sang Wook Bai[1,2]

Systematic reviews provide the highest level of evidence but remain resource-intensive. We evaluated the performance of a large language model (LLM; ChatGPT, OpenAI) in a PRISMA-guided review of randomized controlled trials on vaginal vault prolapse surgery. Prompts were carefully designed to minimize errors, and outputs were verified. Each task was completed within minutes. For title/abstract screening, recall was 69.8% and precision 85.7% (κ = 0.77); full-text agreement 94.1–100% (κ = 0.82–1); data extraction accuracy 87.5–99.7%. From 18 RCTs (1668 women), sacrocolpopexy (SC) showed higher anatomic success than sacrospinous fixation (SSF) (OR 1.42, 95% CI 0.71–2.84). Transvaginal mesh improved 3-year objective success compared with SSF (OR 1.84, 95% CI 1.13–2.99) but had higher reoperation rates (5–16% vs 2–4%) than SC. We did not find conclusive evidence that any single technique is superior; most comparisons were underpowered, with wide confidence intervals and substantial heterogeneity. All LLM-derived statistical results were identical to those from conventional R analyses, confirming robustness. Validated LLM workflows can enable more efficient and scalable evidence synthesis.

Globally, pelvic organ prolapse (POP) has emerged as an important health issue in women, largely driven by population aging. POP occurs when the pelvic organs descend from their normal position due to weakness of the pelvic floor muscles and connective tissues[1]. The number of women seeking medical care for symptomatic POP continues to rise, with a lifetime risk of requiring surgical intervention estimated at approximately 12–19%[2]. Large population-based data also show that the risk of subsequent surgery is substantial; for example, in a Danish registry study, the overall reoperation rate was 11.5%[3].

The pathogenesis of pelvic organ prolapse (POP) primarily reflects failure of the pelvic floor muscles or connective tissues[4]. Post-hysterectomy vaginal vault prolapse is particularly challenging because critical apical supports, the uterosacral and cardinal ligaments, are weakened or absent. Symptoms may appear even in very early disease (Pelvic Organ Prolapse Quantification [POP-Q] stage I; symptom threshold; point C − 5 cm), often earlier than in other compartments[5]. Vaginal vault prolapse also commonly coexists with anterior or posterior compartment defects, further complicating surgical decision-making. Consequently, consensus on standardized management remains limited[6–8].

In evidence-based medicine, well-designed randomized controlled trials (RCTs), followed by systematic reviews and meta-analyses, play a crucial role in developing clinical practice guidelines and diagnostic or therapeutic recommendations. However, the human-led Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) process has inherent limitations, such as difficulty adapting to different types of reviews, subjectivity in judging study eligibility, and considerable time required for screening and selection[9,10]. Indeed, according to previous research, it takes an average of 67.3 weeks for systematic reviews registered in the International Prospective Register of Systematic Reviews (PROSPERO) to be completed[11].

Researchers have therefore explored using artificial intelligence (AI), notably large language models (LLMs), to assist multiple stages of evidence synthesis, from question formulation and screening to data extraction, bias assessment, code generation, and drafting[12]. While full end-to-end automation remains a challenge, and risks such as hallucinations, bias, and methodological unreliability persist, the potential for AI to significantly accelerate and enhance evidence synthesis is increasingly acknowledged.

This study has two primary objectives. First, aimed to compare the effectiveness and safety of surgical options for posthysterectomy vaginal vault prolapse. Second, we aimed to prospectively evaluate an AI-augmented review workflow alongside expert review.

[1]Department of Obstetrics and Gynecology, Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea. [2]Institute of Women's Life Medical Science, Yonsei University College of Medicine, Seoul, Republic of Korea. [3]Biostatistics Collaboration Unit, Department of Biomedical Informatics, College of Medicine, Yonsei University, Seoul, Republic of Korea. e-mail: swbai@yuhs.ac

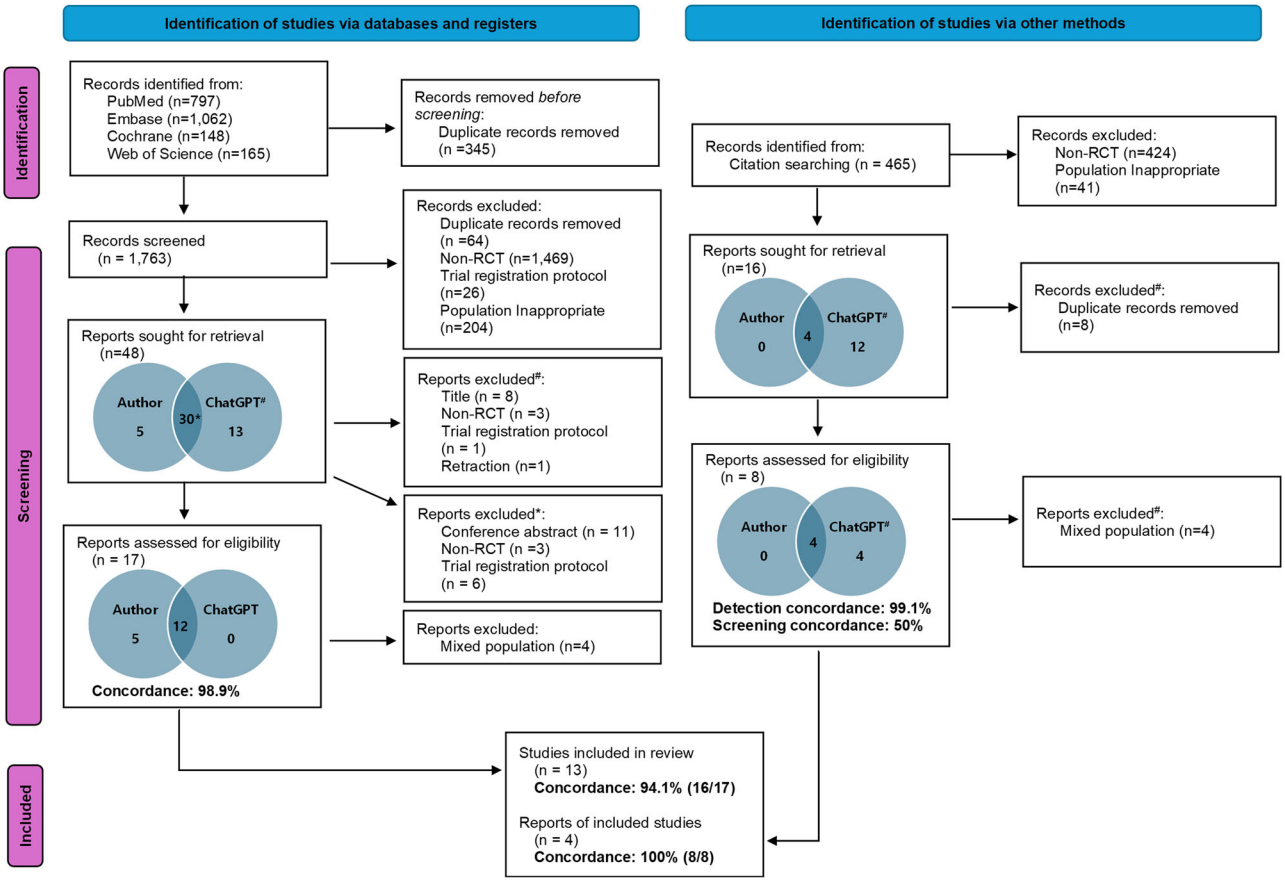


**Fig. 1 | PRISMA 2020 flow diagram.** Study selection process for randomized controlled trials of surgical treatments for posthysterectomy vaginal vault prolapse. Screening was conducted in parallel by two human reviewers and ChatGPT (version 5.0), with discrepancies resolved by human consensus.

## Results

### Title & abstract screening performance

Out of 1763 records screened after duplicate removal, ChatGPT correctly identified 30 true positives and 1,715 true negatives, while yielding 5 false positives and 13 false negatives. This corresponded to an overall accuracy of 98.7%, with a precision of 85.7%, a recall of 69.8%, and an F1-score of 0.77. The inter-rater reliability between ChatGPT and the expert reviewer, quantified using Cohen's κ, was 0.77, indicating substantial agreement beyond chance. (Table S1).

### Full text review performance

At the stage of full-text review, a total of 17 articles were independently assessed by both ChatGPT and the authors. Of these, 16 judgments were concordant and 1 was discordant, yielding an overall percent agreement of 94.1% (16/17). The inter-rater reliability, quantified using Cohen's κ, was 0.82, indicating almost perfect agreement between ChatGPT and the authors. (Table S2).

### Snowballing performance

During the detection stage of the snowballing process ($n = 465$ after duplicate removal), ChatGPT correctly identified 4 true positives and 449 true negatives, with 4 false negatives and no false positives. This corresponded to an overall accuracy of 99.1% (95% CI, 97.8–99.7), a precision of 100% (95% CI, 51.0–100.0), and a recall of 50.0% (95% CI, 21.5–78.5). The resulting F1-score was 0.67, and Cohen's κ indicated substantial agreement ($\kappa = 0.66$).

At the subsequent screening stage ($n = 8$ overlapping records), ChatGPT identified 3 true positives and 1 true negative, with 1 false positive and 3 false negatives. This yielded an accuracy of 50.0% (95% CI, 21.5–78.5), a precision of 75.0% (95% CI, 30.1–95.4), and a recall of 50.0% (95% CI, 18.8–81.2). The corresponding F1-score was 0.60. However, Cohen's κ was 0.00, which should be interpreted with caution given the very small sample size at this stage. (Table S3).

Finally, during the full-text review stage ($n = 8$ candidate studies), ChatGPT and the authors reached complete concordance (8/8, 100%), including 4 correct exclusions and 4 correct inclusions. This yielded a percent agreement of 100% and a Cohen's κ of 1.00, indicating perfect agreement beyond chance. (Table S4, Fig. 1).

### Data extraction performance

Across 291 extracted data points for objective outcomes, ChatGPT demonstrated very high agreement with human reviewers, achieving an accuracy of 99.7% (289/291). Performance metrics were uniformly strong, with accuracy, precision, recall, and F1-score all at 99.7%. Only two discrepancies were observed, one attributable to human error and one to ChatGPT error. (Table S5, Supplementary Data 2).

For subjective outcomes, performance was somewhat lower but remained high. Across 72 extracted data points, ChatGPT achieved an accuracy of 87.5% (63/72), precision of 94.0% (63/67), recall (sensitivity) of 92.6% (63/68), and an F1-score of 93.3%. Specificity was calculated as 0%, though this metric has limited interpretability in this context given the absence of true negatives in the dataset. (Table S5).

### Risk of bias assessment performance

Across 18 randomized controlled trials (RCTs) and 6 RoB 2.0 domains (108 domain-level judgments in total), agreement between the authors and the ChatGPT-assisted assessment was generally high. Of the 108 judgments, 89 were concordant between both assessments, while 17 were correctly classified only by the authors and 2 only by ChatGPT. This corresponds to an overall percent agreement of 82.4%. In contrast, Cohen's κ was –0.03, a

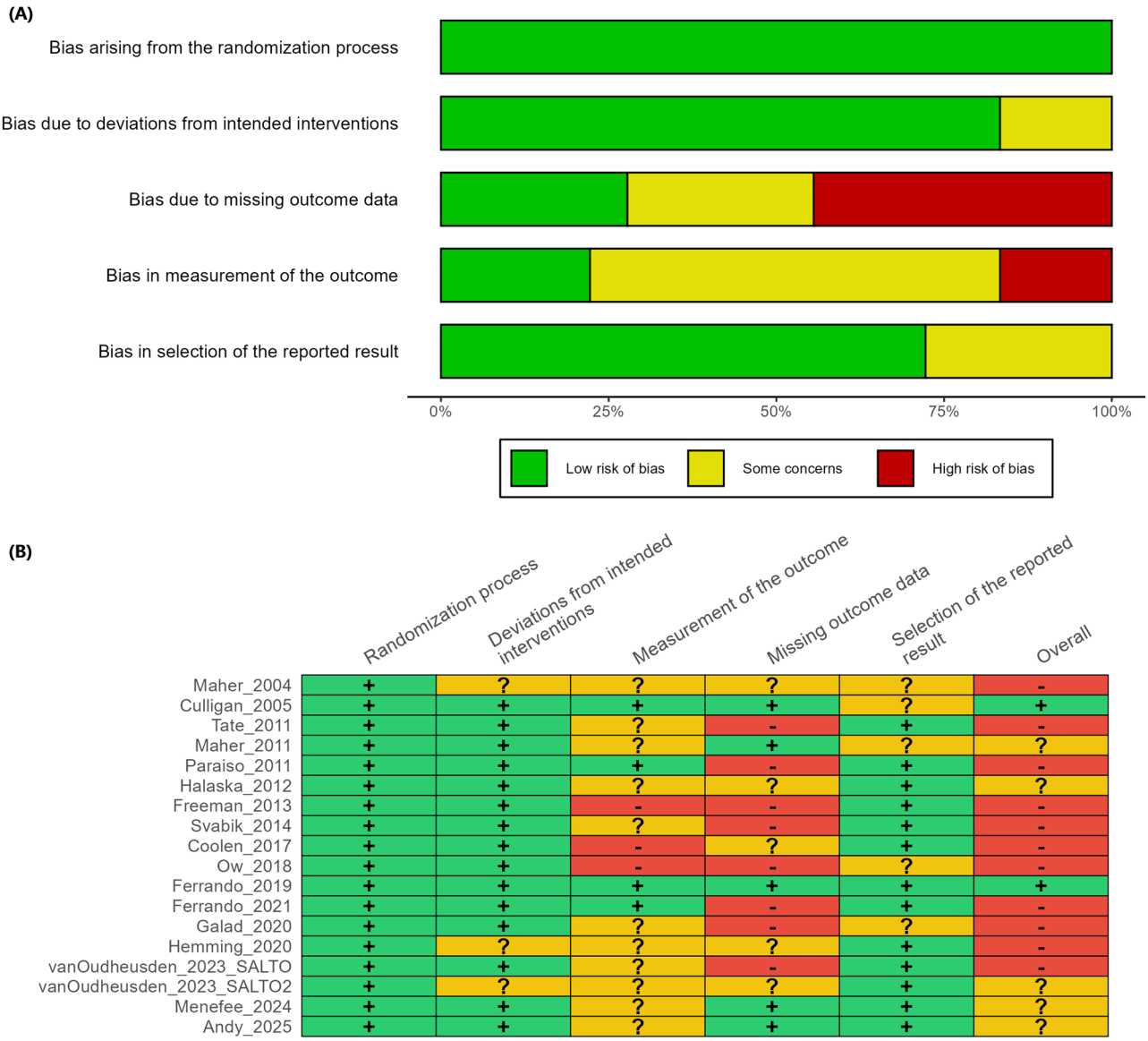


**Fig. 2 | Risk of bias assessment. A** Risk of bias graph: Review authors' judgements about each risk of bias item presented as percentages across all the studies included. **B** Risk of bias summary: Review authors' judgements about each risk of bias item for each study included.

paradoxical finding explained by the highly unbalanced distribution of disagreements (the majority of discordant ratings arose from cases where only the authors provided the correct classification). Because κ is sensitive to prevalence and marginal imbalance, percent agreement may better reflect practical concordance in this specific setting. In cases where the judgments of the authors and ChatGPT differed, the final decision was indicated as "*following the authors' judgment" or "#following the ChatGPT assessment". (Fig. 2, Table S6).

### Consistency with conventional statistical software

All AI-assisted meta-analyses produced results that were identical to those obtained using conventional analyses in R, including effect sizes, 95% CIs, and heterogeneity statistics (Q, p-value, $I^2$, $\tau^2$). This confirms the robustness and reproducibility of the findings. Turning to the clinical evidence base, we next summarize the included trials and their comparative outcomes.

### Included studies, interventions, and participants

A total of 18 RCTs involving 1668 women were included, with a mean age of 64 years[13–30]. Details of the included interventions are summarized in Table S7. Figure 3 illustrates the evidence network, with edges labeled by the number of RCTs per comparison. Across the included RCTs, reported follow-up durations ranged from 1 to 9 years (median 1 year, mean 2.2 years).

### Analytic considerations

Small-study effects/publication bias were not assessed because few trials were available per comparison. Certainty of evidence was not graded. Most contrasts included ≤3 trials, limiting the value of subgroup or sensitivity analyses. Subgroup and meta-regression analyses were not prespecified; therefore, sources of heterogeneity were not formally explored.

### POP-Q point C

At 1 year, two trials comparing TVM with SSF showed no significant difference in POP-Q point C (WMD −1.65 cm, 95% CI −4.16 to 0.87; $I^2 = 92.6\%$). These findings should be interpreted with caution due to very high between-study heterogeneity and imprecision. (Fig. 4) Similarly, ASC and LSC demonstrated equivalent anatomical outcomes (WMD −0.06 cm, 95% CI −0.61 to 0.49; $I^2 = 0\%$). (Fig. 5).

### Objective success

For ASC/LSC versus SSF, pooled data from two trials indicated no significant difference in objective success at 1 year (OR 2.40, 95% CI 0.83–6.94; $I^2 = 0\%$). (Fig. 6) In contrast, three trials comparing TVM with SSF at 1 year suggested a possible advantage of TVM, but with very wide confidence intervals that preclude firm conclusions (OR 6.13, 95% CI 0.87–43.07; $I^2 = 75$). (Fig. 7) Accordingly, the absence of statistical significance here should not be interpreted as evidence of equivalence. At 3 years, TVM maintained superior outcomes over SSF (OR 1.84, 95% CI 1.13–2.99; $I^2 = 0\%$). (Fig. 8).

### Reoperation for prolapse

Two RCTs comparing ASC/LSC with SSF demonstrated a lower, though not statistically significant, risk of reoperation after SC (OR 0.54, 95% CI 0.12–2.44; $I^2 = 21\%$). (Fig. 9).

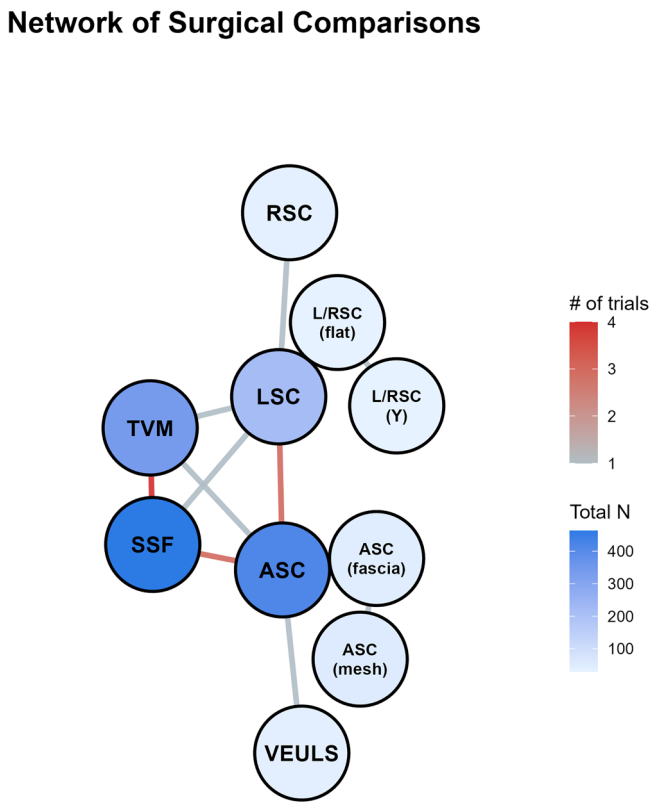


**Fig. 3 | Network of surgical comparisons for vaginal vault prolapse.** Each node represents a surgical intervention; node fill (blue gradient) denotes the cumulative number of randomized participants assigned to that intervention (Total N). Edges connect interventions directly compared within ≥1 randomized trial; edge color (red gradient) indicates the number of trials for that pair. Multi-arm trials contribute all pairwise edges. Node positions are arranged for readability and do not convey effect size or direction. Subtechnique (material) nodes are shown when randomized head-to-head (e.g., ASC fascia vs mesh; L/RSC flat vs Y). ASC (abdominal sacral colpopexy), LSC (laparoscopic sacral colpopexy), RSC (robotic sacral colpopexy), SSF (sacrospinous fixation), TVM (transvaginal mesh), VEULS (vaginal endoscopic unilateral lateral suspension).

### Subjective outcomes

Across RCTs, all surgical interventions were associated with substantial improvement in patient-reported prolapse symptoms and quality of life. Overall, patient-reported improvements paralleled objective findings, although several trials noted that anatomic failure could remain asymptomatic. Detailed study-level data are provided in Table 1.

### Complications

Reoperations for mesh-related complications varied by procedure, occurring in 5–16% of women after TVM and 2–4% after SC. Other complications are provided in Table 2, Table S8, and Figs. S1–4. Perioperative outcomes, including operative time, blood loss, and hospital stay, are summarized in Figs. S5–9.

### LLM replications

ChatGPT-generated forest plots are provided in Figs. S10–S24.

## Discussion

To our knowledge, this review is among the first to prospectively performed an AI-augmented workflow against human review across multiple stages. The model assisted in every stage of the review except the initial database search, which is still beyond its current capability. Importantly, every decision was recorded in a clear log, so that each step of the process can be traced and checked. This approach improved both transparency and reproducibility, which are often lacking in traditional reviews, and represents a key strength of our study.

Traditionally, snowballing requires reviewers to identify candidate references and then screen them against eligibility criteria[31]. In our study, we used ChatGPT to support both steps within a single LLM-assisted workflow, in which the model simultaneously identified potentially relevant references and proposed inclusion and exclusion decisions for subsequent human review. To our knowledge, this is the first demonstration that an LLM can perform snowballing in this manner. This shows that AI can not only understand methodological principles but also deliver practical results, thereby improving both the rigor and efficiency of systematic review workflows. However, the model's performance in this snowballing step had clear limitations: its recall was only about 50% during reference detection, underscoring that human verification was necessary to avoid missing relevant studies. Because prompting was optimized and outputs were verified, this recall should be interpreted as an upper-bound estimate in a controlled setting rather than a generalizable real-world performance level.

Using direct PDF parsing, LLMs markedly improved the efficiency of data extraction. In repeated runs, accuracy exceeded 99%, substantially outperforming prior reports of human error rates of 8–42%[27]. The model automatically generated executable statistical code and structured reports, enabling Python-based analyses with little need for human intervention[32,33]. This lowered the barrier to adoption and saved considerable reviewer time.

Despite these advantages, several important limitations remain. The model could not extract outcomes presented only as figures (Halaska 2012[18])[33], and it struggled with very long reports such as Hemming 2020 (256 pages)[25], where errors persisted until authors manually re-specified key data. In addition, LLMs could not process multiple full-text PDFs in parallel, often resulting in omissions when several documents were handled together[33]. These issues are consistent with prior studies noting that LLMs

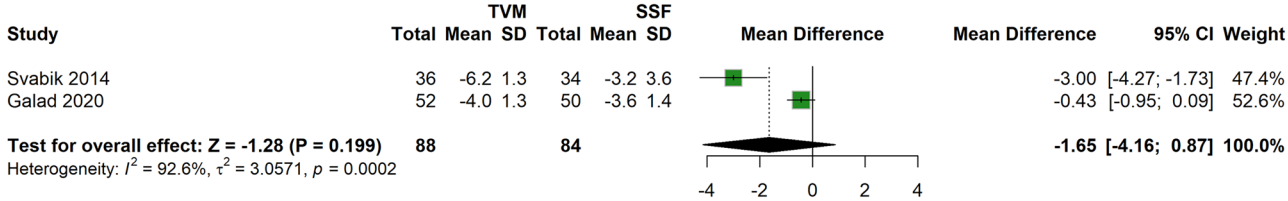


**Fig. 4 | Forest plot of comparison: TVM vs SSF, POP-Q point C (at 1 year).** Caution: high study heterogeneity. TVM, Transvaginal Mesh; SSF, Sacrospinous Fixation; POP-Q, Pelvic Organ Prolapse Quantification system.

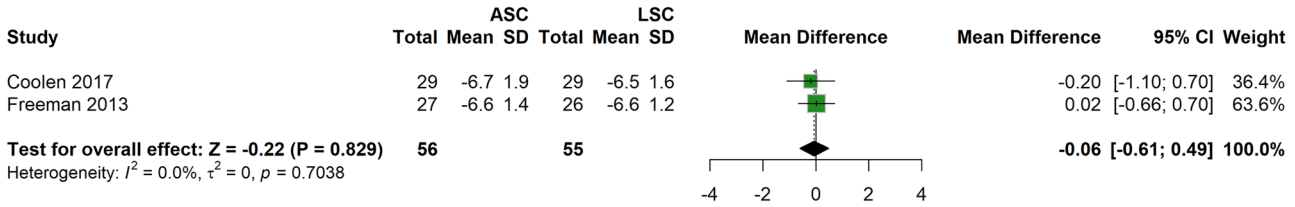


**Fig. 5 | Forest plot of comparison: ASC vs LSC, POP-Q point C (at 1 year).** ASC, Abdominal Sacrocolpopexy; LSC, Laparoscopic Sacrocolpopexy; POP-Q, Pelvic Organ Prolapse Quantification system.

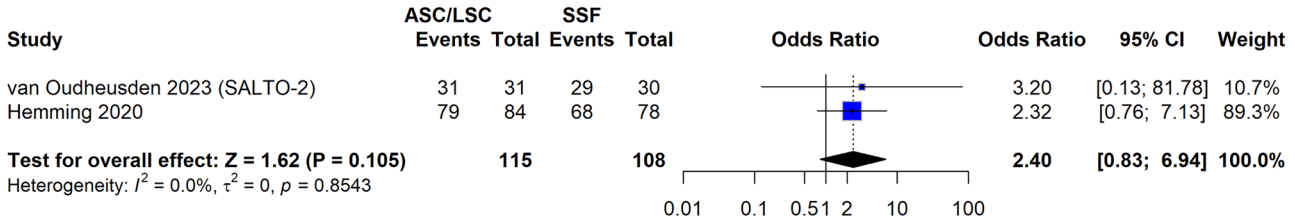


**Fig. 6 | Forest plot of comparison: ASC/LSC vs SSF, Objective Success (at 1 year).** ASC, Abdominal Sacrocolpopexy; LSC, Laparoscopic Sacrocolpopexy; SSF, Sacrospinous Fixation.

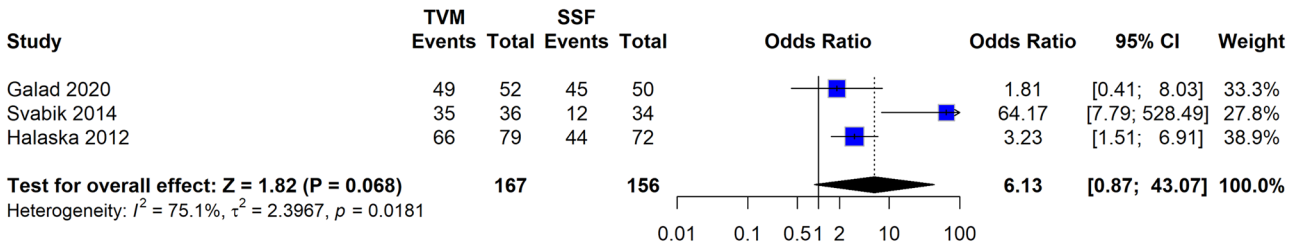


**Fig. 7 | Forest plot of comparison: TVM vs SSF, Objective Success (at 1 year).** Caution: high study heterogeneity. TVM, Transvaginal Mesh; SSF, Sacrospinous Fixation.

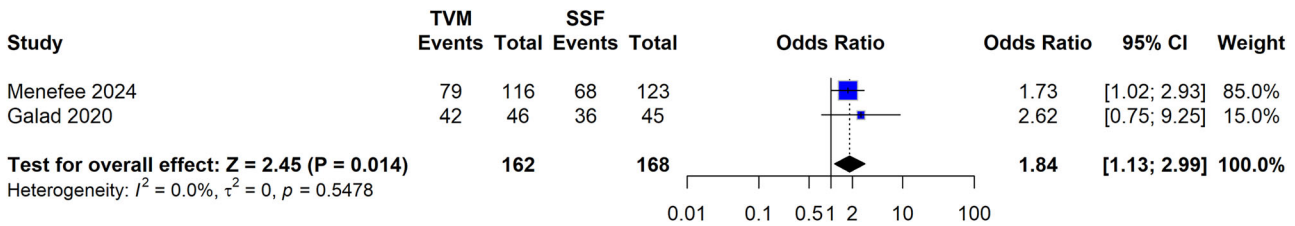


**Fig. 8 | Forest plot of comparison: TVM vs SSF, Objective Success (at 3 years).** TVM, Transvaginal Mesh; SSF, Sacrospinous Fixation.

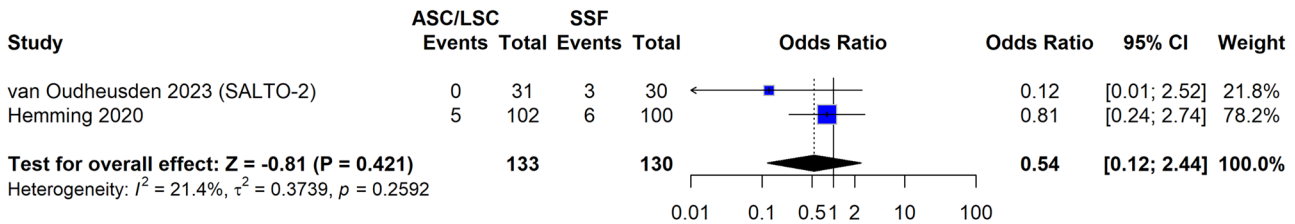


**Fig. 9 | Forest plot of comparison: ASC/LSC vs SSF, Reoperations for POP (at 1 year).** ASC, Abdominal Sacrocolpopexy; LSC, Laparoscopic Sacrocolpopexy; SSF, Sacrospinous Fixation.

struggle with hierarchical tables, figure-based outcomes, and long or irregularly structured documents[27]. Together, these findings underscore the need for continued human oversight and suggest that current models are best suited to sequential, text-based extraction rather than parallel, image- or figure-heavy workflows.

It is essential to recognize the risks of hallucination (plausible yet fabricated outputs) when using LLMs for systematic reviews. A comparative study found that GPT-4 falsely generated non-existent citations in about 28.6% of cases when retrieving references for systematic reviews[34]. Learning from such findings, we incorporated a grounded prompt design and an

**Table 1 | Subjective outcomes**

| Study (1st author year) | Intervention | Subjective outcomes assessment tool | Results | F/U (y) |
|---|---|---|---|---|
| Maher[13] | ASC (polypropylene) vs SSF | SUDI, IIQ, Modified sexual function questionnaires, SF-36, Patient satisfaction (VAS 0–100) | Subjective success rate: 94% (43/46) vs 91% (39/43), $p = 0.19$<br>Patient satisfaction: 85% (39/46) vs 81% (35/43), $p = 0.78$ | 2 |
| Culligan[14] | ASC (fascia lata vs polypropylene) | Not defined | Not defined | 1 |
| Tate[17] | ASC (fascia lata vs polypropylene) | Symptoms of prolapse, or bulge | Clinical (objective and subjective) success: 90% (26/29) vs 97% (28/29), $p = 0.61$<br>Of failures by objective definition, 77% were asymptomatic, suggesting anatomical failures often not clinically relevant[a] | 5 |
| Maher[15] | LSC vs TVM | APFQ, P-QOL, Patient satisfaction (VAS 0–100) | Symptomatic prolapse: 2% (1/53) vs. 7% (4/55), $p = 0.18$<br>Patient satisfaction: 87 ± 21 vs 79 ± 20, $p = 0.002$ | 2 |
| Paraiso[16] | LSC vs RSC | PFDI-20, PFIQ-7, PISQ-12, EG-5D, Activity Assessment Scale | Improvement without differences<br>Robotic group reported higher pain scores at rest and with activity weeks 3–5 (VAS, $p = 0.02$–$0.04$) and required longer NSAID use (median 20 vs 11 days, $p < 0.005$)[a] | 1 |
| Halaska[18] | SSF vs TVM | UIQ, CRAIQ, POPIQ, PISQ short | Improvement without differences<br>CRAIQ (bowel symptoms) improved more in mesh group ($p < 0.05$)[a] | 1 |
| Freeman 2013 (LAS)[19] | ASC vs LSC | PGI-I, P-QOL, SF-36 | PGI-I "very much better", "much better" 90% vs 80%<br>P-QOL "prolapse impact" No difference | 1 |
| Svabik[20] | SSF vs TVM | POPDI, UDI, CRADI, PISQ-12, ICIQ-SF | Improvement without differences<br>De novo SUI higher after Prolift (13 vs 3 cases, $p = 0.023$)[a] | 1 |
| Coolen[21] | ASC vs LSC | UDI, DDI, IIQ, PGI-I | Composite outcome of success: 89.2% (33/37) vs 83.8% (31/37)<br>PGI-I "very much better", "much better" 74% (20/27) vs 71% (22/31), $p = 0.563$<br>No difference | 1 |
| Ow[22] | VEULS with anterior mesh vs ASC | PFDI-20, POPDI, UDI, CRADI, PISQ, bothersome bulge | Bothersome bulge 26.5% (9/34) vs 8.6% (3/35), $p = 0.06$<br>No difference | 4 |
| Ferrando[23] | R/LSC (Dual vs Y mesh) | Bulge | 8% (2/27) vs 4% (1/28), $p = 0.55$ | 0.5 |
| Ferrando[26] | R/LSC (Dual vs Y mesh) | PFDI-20, POPDI, CRADI, UDI, Bulge | Subjective recurrence: 8.3% (2/24) vs 10.0% (2/20), $p = 0.90$<br>No difference | 2 |
| Galad[24] | TVM vs SSF | Patient satisfaction (VAS 0–100), IQoL | Quality of life : 91% (42/46) vs 87% (39/45), $p = 0.898$<br>No difference | 3 |
| Hemming[25] | Vaginal vs Abdominal | POP-SS, QoL, EQ-5D, ICIQ | No difference | 1 |
| van Oudheusden 2023 (SALTO)[27] | LSC vs ASC | UDI, DDI, IIQ, PGI-I, PISQ | Clinical (objective and subjective) outcomes: 78.6% (11/14) vs 84.6% (11/13), $p = 0.686$<br>PGI-I "very much better", "much better" 57.9% (11/19) vs 58.8% (10/17), $p = 0.955$<br>No difference | 9 |
| van Oudheusden 2023 (SALTO-2)[28] | LSC vs SSF | UDI, DDI, IIQ, PGI-I, PISQ, bulge | Clinical (objective and subjective) outcomes: 89.3% (25/28) vs 86.2% (25/29), $p = 0.810$<br>Bothersome bulge 10.3% (3/29) vs 10.0% (3/30), $p = 1.000$<br>PGI-I "satisfaction" 78.6% (22/28) vs 80.0% (24/30), $p = 0.778$<br>No difference | 1 |
| Menefee[29] | NTR vs SC vs TVM | PFDI, POPDI, UDI, CRADI, PFIQ, UIQ, CRAIQ, PISQ-IR, BIPOP, FAS, SF-12 | Symptomatic failure 17/123 vs 17/121 vs 11/115<br>PFDI change −73.1 (−79.3 to −66.9) vs −84.6 (−90.8 to −78.4) vs −85.6 (−91.8 to −79.3), $p = 0.008$[b]<br>UDI ($p = 0.03$), CRADI ($p = 0.08$), CRAIQ ($p = 0.04$)[b]<br>SF-12 mental component $p = 0.03$[b]<br>SC = TVM > NTR | 3 |
| Andy[30] | NTR vs SC vs TVM | BIPOP, PISQ-IR | No difference | 3 |

*APFQ* Australian Pelvic Floor Questionnaire, *BIPOP* Body Image in Pelvic Organ Prolapse, *CRADI* Colorectal-Anal Distress Inventory, *CRAIQ* Colorectal-Anal Impact Questionnaire, *DDI* Defecatory Distress Inventory, *EQ-5D* EuroQol-5 Dimension, *FAS* Female Sexual Function Assessment Scale, *ICIQ-SF* International Consultation on Incontinence Questionnaire–Short Form, *IIQ* Incontinence Impact Questionnaire, *IQoL* Incontinence Quality of Life, *LSC* Laparoscopic Sacrocolpopexy, *NTR* Native Tissue Repair, *PFDI-20* Pelvic Floor Distress Inventory–20, *PFIQ-7* Pelvic Floor Impact Questionnaire–7, *PGI-I* Patient Global Impression of Improvement, *PISQ* Pelvic Organ Prolapse/Urinary Incontinence Sexual Questionnaire, *PISQ-12* Pelvic Organ Prolapse/Urinary Incontinence Sexual Questionnaire–12 items, *PISQ-IR* Pelvic Organ Prolapse/Urinary Incontinence Sexual Questionnaire–IUGA Revised, *POPDI* Pelvic Organ Prolapse Distress Inventory, *POPIQ* Pelvic Organ Prolapse Impact Questionnaire, *POP-SS* Pelvic Organ Prolapse Symptom Score, *P-QOL* Prolapse Quality of Life Questionnaire, *QoL* Quality of Life, *RSC* Robotic Sacrocolpopexy, *SF-12* 12-Item Short Form Survey, *SF-36* 36-Item Short Form Survey, *SUDI* Stress and Urge Incontinence and Daily Impact Questionnaire, *TVM* Transvaginal Mesh, *UDI* Urogenital Distress Inventory, *UIQ* Urinary Impact Questionnaire, *VEULS* Vaginal Endoscopic Unilateral Lateral Suspension, *VAS* Visual Analogue Scale.

[a]It was extracted by ChatGPT only.

[b]It was extracted by the author only.

**Table 2 | Summary of complication and reoperation rates across surgical approaches**

| Surgical procedure | Typical complication profile | Reoperation rate (%) | Clinical implications |
|---|---|---|---|
| SC | Mainly intraoperative injuries (bladder, bowel); mesh exposure 1–4% | 2–4 | Lower mesh-related reoperation risk |
| TVM | Mesh exposure 5–16%; mesh contraction | 10–16 | Highest reoperation risk, largely mesh-related |
| SSF | Bleeding, hematoma, buttock pain | 6–10 | Non-mesh, but higher recurrence risk |
| NTR | Lower severe complication risk | 10-11.5 (mostly recurrence) | No mesh-related events |

*SC* Sacrocolpopexy, *TVM* Transvaginal Mesh, *SSF* Sacrospinous Fixation, *NTR* Native Tissue Repair.

exhaustive audit trail to minimize the risk of hallucinations. Prior work has shown that methods such as retrieval-augmented generation (RAG), task grounding, and formal, concrete prompt styles can significantly reduce hallucination rates[35]. In our protocol, each step was guided by precise instructions (e.g., requiring text snippets to justify exclusions), consistent prompting, and human verification before acceptance.

Beyond hallucination, LLMs in systematic reviews may introduce other critical errors such as omission of eligible studies (as reflected by the model's 69.8% recall in our title/abstract screening stage, where it missed nearly 30% of eligible trials), misclassification of inclusion/exclusion, response inconsistency, and inherent biases. Importantly, these recall values reflect performance under optimized prompting and close human supervision, and may be lower in less controlled settings. In clinical contexts, frameworks that define error types including hallucinations and omissions have shown that refining prompts and workflows can reduce omission error rates to approximately 3.5% and hallucination rates to 1.5%, outperforming human note-taking baselines[36]. To mitigate such risks, we employed repeated runs and prompt diversification to minimize omissions, embedded explicit eligibility criteria into prompts to reduce misclassification, and standardized prompts with full audit trails to limit inconsistency and enable reproducibility checks. Finally, validated database searches and human oversight were used to counteract potential systematic biases (e.g., open-access or geographic). Together, these strategies improved robustness and reduced the risk of error in our LLM-assisted workflow.

In our study, across 108 domain-level judgments, the overall percent agreement between author-led and ChatGPT-assisted RoB 2.0 assessments was 82.4%, a level comparable to that typically reported among human reviewers (κ 0.4–0.7)[37]. By contrast, Cohen's κ was –0.03. This apparent discrepancy reflects the well-known "kappa paradox": when discordant judgments are rare but the marginal distributions of rating categories are highly imbalanced, κ can approach or even fall below zero despite high observed agreement. Accordingly, the negative κ in our study should not be interpreted as true disagreement between human and LLM assessments. A key strength of our study is that we went beyond reporting raw agreement by identifying where discrepancies were concentrated. Most arose in domains requiring nuanced judgment, missing outcome data, outcome measurement, and selective reporting, which are also recognized as the weakest domains even among expert reviewers[38]. These findings suggest that while LLMs cannot replace human expertise, it can approach human-level reliability and, importantly, help highlight domain-specific challenges where additional reviewer oversight is most needed.

Generalisability is also limited. Our performance metrics were obtained in a single, RCT-focused urogynecology review using optimized prompting and intensive human verification. Performance may differ in other clinical domains, non-RCT evidence bases, non-English literature, and reviews with more complex outcome hierarchies.

Robust safeguards are essential: all AI-generated outputs must be verified against original sources before acceptance. Prior reviews also stress that large language models should not be used in isolation for systematic reviews, as AI text may appear authoritative even when incorrect[39]. Accordingly, LLMs should be considered an adjunct rather than a replacement for human reviewers. With proper oversight and ethical standards, combining human and AI strengths can improve the speed and quality of evidence synthesis[40].

Importantly, this was a validation-focused study rather than a time-and-motion study. We did not systematically record person-time at each step of the review process, and all AI-generated outputs (screening, data extraction, and risk-of-bias assessments) were fully double-checked by human reviewers. Thus, although individual LLM runs were completed within minutes[41], we cannot empirically demonstrate that the overall human time required was lower than for a conventional manual review in this setting.

In future implementations, once prompts and workflows have been prospectively validated and the extent of necessary double checking is reduced, the wall-clock time for LLM processing (typically seconds to minutes) may more closely reflect the true human effort required for certain steps. However, our current results do not yet support such a claim, and we therefore regard any efficiency statements as conceptual and hypothesis-generating rather than empirically proven. Turning to the clinical outcomes, we summarize the comparative effectiveness and safety findings from the included randomized trials below.

In this systematic review and meta-analysis of 18 randomized trials, SC provided durable anatomical support, with comparable outcomes between abdominal and laparoscopic approaches. Compared with SSF, SC showed a non-significant trend toward fewer reoperations, consistent with prior reports that no surgical approach is clearly superior overall. Transvaginal mesh was associated with higher objective success than SSF, but at the cost of increased mesh-related complications.

SC attaches the vaginal apex to the anterior longitudinal ligament, offering biomechanically stronger support than SSF, which anchors laterally to the SSF[42]. Although SC is widely regarded as the reference standard for apical suspension, our meta-analysis did not demonstrate consistent superiority across outcomes, highlighting the need for individualized surgical decision-making.

With respect to subjective success, most patient-reported outcomes improve substantially after all procedures. Our findings highlight that anatomical success does not necessarily equate to patient-centered outcomes. Recent IUGA guidance emphasized that the most meaningful definition of success is the absence of vaginal bulge symptoms, given the frequent disconnect between anatomical results and patient experience[43]. Because POP is fundamentally a disorder of function, subjective outcomes such as quality of life and symptom relief are as important as anatomical correction. However, heterogeneity in patient-reported outcome measures across studies limits quantitative synthesis and often necessitates a narrative approach. In our review, we addressed this limitation by systematically summarizing subjective outcomes alongside anatomical findings.

This study represents the most up-to-date systematic review and meta-analysis on surgical approaches for vaginal vault prolapse. Including studies on TVM, despite its withdrawal by the FDA[44], may be considered a methodological strength, as it provides historical and comparative context for understanding why SC has become the current gold standard and contributes detailed data on mesh-related complications. While these findings enrich our understanding of the evolution of prolapse surgery, their clinical applicability is limited because TVM to

contemporary U.S. practice is limited because TVM kits have been withdrawn from the commercial market; however, these findings remain relevant as historical context and for regions where TVM procedures are still performed. Results should therefore be interpreted with caution, acknowledging their historical relevance and role in informing the safety profile of contemporary surgical approaches.

Due to heterogeneity across studies, some important long-term evidence could not be incorporated into our meta-analysis. For example, the SALTO trial, which reported outcomes after a median 9-year follow-up of LSC and ASC, found comparable mean times to surgical reintervention (41.2 months for LSC vs. 55.8 months for ASC, $p = 0.814$). Notably, rare but serious mesh-related complications were observed, including mesh infection requiring extensive surgery 5.6 years after ASC and mesh removal for vaginal exposure and infection 10.2 years after LSC[27]. These findings underscore the need for more long-term follow-up studies to better evaluate durability and late complications after vaginal vault prolapse surgery.

In conclusion, this meta-analysis demonstrates that SC provides durable anatomical support, with similar outcomes between abdominal and laparoscopic approaches, and a possible reduction in reoperation risk compared with SSF that did not reach statistical significance. TVM was associated with higher objective success but also greater mesh-related complications. Consistent with prior evidence, we did not identify definitive superiority of any single surgical technique; most contrasts relied on few trials with imprecise estimates, and further adequately powered RCTs with long-term follow-up are needed. Importantly, absence of evidence of superiority should not be interpreted as evidence of equivalence, particularly for comparisons informed by one or two trials with wide confidence intervals and substantial heterogeneity (e.g., TVM vs SSF at 1 year).

Beyond these clinical insights, this review also highlights the potential of LLMs to transform evidence synthesis. With appropriate human oversight, AI can enhance efficiency, transparency, and reproducibility, offering a promising paradigm for systematic reviews and meta-analyses not only in urogynecology but across digital medicine and healthcare research more broadly.

## Methods

### Study registration

We conducted a PRISMA-guided systematic review and meta-analysis, prospectively registered in PROSPERO (CRD420251039219). The protocol is available in the PROSPERO record, and no post-registration amendments were made.

### Ethical considerations, and use of artificial intelligence

Ethics approval was waived as this study used publicly available data. ChatGPT (OpenAI, GPT-5, San Francisco, CA, USA) assisted screening, data extraction, Cochrane Risk of Bias 2.0 (RoB 2.0) visualization, and meta-analysis under author supervision; it did not make autonomous decisions on study inclusion or interpretation, and all AI outputs were verified by the authors.

### Participants

Only women with symptomatic primary post-hysterectomy vaginal vault prolapse were included. Studies were excluded if they involved uterine prolapse, cervical prolapse, or recurrent vaginal vault prolapse. Mixed-population studies were included only if separate data for vaginal vault prolapse were available. Studies with concomitant hysterectomy procedures were excluded.

### Type of studies

Only RCTs comparing different treatments for vaginal vault prolapse were eligible. Non-RCT designs (quasi-randomized, crossover, retrospective), protocols, abstracts, reviews, case reports, and animal studies were excluded. If multiple articles reported on the same cohort, only the longest follow-up study was included.

### Intervention

No restrictions were applied regarding the type of surgical intervention for vaginal vault prolapse at the eligibility stage. Among the various surgical options described in the literature, sacrocolpopexy (SC), sacrospinous ligament fixation (SSF), and transvaginal mesh (TVM) are the most commonly used procedures for vaginal vault prolapse[45]. SC approaches included abdominal (ASC), laparoscopic (LSC), and robotic (RSC) techniques; vaginal endoscopic unilateral lateral suspension (VEULS) was also considered. Although the U.S. FDA ordered the remaining commercially available transvaginal mesh devices for POP repair to be withdrawn from the U.S. market in April 2019 because of safety concerns[44], TVM studies evaluating transvaginal mesh were still included in this review in order to assess the safety and complication rates of mesh-based procedures compared with native tissue repairs.

### Outcomes

Primary outcomes were POP-Q point C, objective success, and reoperation. Secondary outcomes included patient-reported measures and perioperative parameters (e.g., blood loss, operative time, hospital stay, complications)[46]. Objective (anatomical) success was defined in accordance with each trial's criteria, generally as the absence of significant apical prolapse (typically POP-Q stage 0 or I at follow-up), with only minor variation in the exact cut-offs across studies. Even modest heterogeneity in outcome definitions may contribute to imprecision and limit direct comparability across pooled estimates.

### ChatGPT performance and workflow evaluation

This systematic review was conducted in parallel by two authors (Y.P., S.W.B.) and an AI-augmented workflow, with accuracy compared at each stage using confusion matrices[47]. Dialogue context and conversation history were preserved to leverage LLM contextual learning[9].

### Literature search and study selection

A professional medical librarian searched PubMed, Embase, Cochrane CENTRAL, and Web of Science from inception to April 2025 with no language restrictions (strategy in Supplementary Data 1). LLMs cannot directly execute database queries; human experts performed searches, and ChatGPT was applied in subsequent stages.

### Title and abstract screening

Two authors (Y.P. and S.W.B.) independently screened titles and abstracts for eligibility based on the above criteria, with discrepancies resolved by consensus. The EndNote library file was converted to XML format and uploaded to ChatGPT for the title and abstract screening phase. ChatGPT was provided with the study title, study objective, and explicit inclusion/exclusion criteria, and was instructed to adopt a sensitive screening approach to avoid premature exclusion of potentially relevant studies. To enhance specificity and reduce unnecessary reviews, tailored prompts were applied, with additional instructions designed to minimize hallucinations and omissions[36]. (Table S9).

A confusion matrix was constructed to compare ChatGPT's classifications against those of the human expert (ground truth). Following this evaluation, studies classified as true positives (selected by both) and false negatives (missed by ChatGPT but selected by the expert) were advanced to full-text review. Studies classified as false positives (selected by ChatGPT but excluded by the expert) underwent brief reassessment by the authors to confirm appropriate exclusion, ensuring that no potentially relevant studies were overlooked. All true negatives (excluded by both) were removed from further review[9,47–49].

### Study selection after full-text review

The full-text review was conducted independently and in duplicate by the two authors (Y.P. and S.W.B.) using the articles selected during the previous stage, with disagreements resolved by consensus. Concurrently, ChatGPT

independently reviewed the same set of articles. One author (Y.P.) uploaded the full-text PDF files of the RCTs to ChatGPT.

ChatGPT was instructed to carefully assess each article against the predefined eligibility criteria. If an article was excluded, the reason was recorded according to predefined categories: (1) study design (non-randomized or quasi-randomized trials, trial registration only, or conference abstract only), (2) population (inappropriate population or mixed population without separable data), (3) publication issues (duplicate publication), and (4) data and follow-up (no full text available or follow-up <6 months). To minimize hallucinations, ChatGPT was explicitly instructed to base all judgments only on information contained in the provided full-text documents[36]. (Table S9).

After ChatGPT completed the review, all articles categorized as exclusions were subsequently re-examined by the two authors (Y.P. and S.W.B.) to verify appropriate classification and confirm no relevant studies had been inadvertently excluded.

### Snowballing

To ensure comprehensive identification of relevant studies, snowballing was conducted following the full-text review. Backward and forward snowballing were first performed using the reference lists and citation records of previously published systematic reviews on vaginal vault prolapse. The process was then extended to include the individual RCTs identified during the current review, with both backward (reference lists) and forward (citation lists) searches performed for each RCT[50].

To reduce the risk of hallucinations or omissions inherent to the AI-assisted process, ChatGPT was guided with accurate reference lists including DOI and PMID information. All candidate articles identified through snowballing subsequently underwent full-text review, and their validity was independently verified by the authors (Y.P. and S.W.B.) using external databases (PubMed, Embase, Cochrane Library, and Web of Science). (Table S9).

### Data extraction

Data extraction was independently and in duplicate performed by two authors (Y.P. and S.W.B.) using a standardized data extraction form. Extracted variables included study characteristics (author, year, design, sample size, and follow-up duration), intervention details, primary and secondary outcomes, complication rates, and perioperative outcomes.

Hierarchical tables with characterized by multi-level headers, merged cells, and irregular layouts are known to pose challenges for LLMs and table-recognition systems; therefore, a tidy-format extraction template was designed (Table S9), and all data were recorded in this format. RCT articles in PDF format were uploaded individually, and extracted data were reviewed in real time to minimize omissions[51]. All ChatGPT-generated outputs were cross-validated by the authors to ensure accuracy and completeness, and any discrepancies were resolved by consensus discussion.

### Statistical analysis

Standard performance metrics, including accuracy, precision, recall, specificity, F1-score, and Cohen's κ, were calculated. For accuracy and other proportion-based measures, Wilson score confidence intervals were applied. For the F1-score, confidence intervals were obtained using bootstrap resampling with 1000 iterations. Confidence intervals (CIs) for Cohen's κ were derived from the asymptotic standard error method[9,47,48].

Pairwise meta-analyses were conducted when at least two RCTs reported the same outcome for comparable surgical interventions. For binary outcomes (objective success, reoperation), pooled odds ratios (ORs) with 95% Cis were calculated using random-effects models. For continuous outcomes (e.g., POP-Q point C), pooled weighted mean differences (WMDs) with 95% CIs were calculated.

Between-study heterogeneity was assessed with the $I^2$ statistic and Cochran's Q test, with thresholds of 25%, 50%, and 75% considered low, moderate, and high heterogeneity, respectively. Analyses were performed using R (version 4.4.3). When substantial heterogeneity was identified, results were interpreted with caution. Risk of bias was independently assessed using the Cochrane RoB 2.0 tool, and results are presented graphically.

Zero-event cells were handled with a 0.5 continuity correction. Outcome scales were harmonized so that effects pointed in the same clinical direction, and unit conversions followed pre-specified rules. No subgroup or meta-regression analyses were pre-specified; accordingly, we did not formally investigate sources of heterogeneity. Sensitivity analyses were not pre-specified. Small-study effects (publication bias) were not assessed because the number of studies per comparison was insufficient. Certainty of evidence was not graded; a GRADE summary-of-findings table is planned for a future update.

### Assessment of Risk of Bias

The revised Cochrane Risk of Bias tool for randomized trials (RoB 2.0) was used to assess each study, covering five key domains: randomization, deviations from intended interventions, missing data, outcome measurement, and selective reporting. Each domain was rated as "low risk," "some concerns," or "high risk," and the overall study risk was based on the most serious domain-level rating[52]. Two reviewers (Y.P. and S.W.B.) independently assessed risk of bias, resolving any discrepancies by consensus. Traffic-light and summary plots for risk-of-bias assessments were also generated using R (version 4.4.3). A structured prompt was also entered into ChatGPT to perform RoB 2.0 assessments, and RCT PDF files were uploaded one by one to obtain the results. (Table S9).

## Data availability

All data analyzed in this study were derived from published randomized controlled trials included in the systematic review and meta-analysis. No new raw patient-level data were generated. The datasets supporting screening decisions, extracted variables, and ChatGPT-assisted workflow outputs are available from the corresponding author upon reasonable request.

Received: 21 August 2025; Accepted: 1 February 2026;
Published online: 19 February 2026

## References

1. Wang, B. et al. Global burden and trends of pelvic organ prolapse associated with aging women: An observational trend study from 1990 to 2019. *Front. Pub. Health* 10 - 2022, https://doi.org/10.3389/fpubh.2022.975829 (2022).
2. Nüssler, E., Granåsen, G., Bixo, M. & Löfgren, M. Long-term outcome after routine surgery for pelvic organ prolapse—A national register-based cohort study. *Int. Urogynecology J.* **33**, 1863–1873 (2022).
3. Løwenstein, E., Møller, L. A., Laigaard, J. & Gimbel, H. Reoperation for pelvic organ prolapse: a Danish cohort study with 15-20 years' follow-up. *Int Urogynecol J.* **29**, 119–124 (2018).
4. DeLancey, J. O. What's new in the functional anatomy of pelvic organ prolapse?. *Curr. Opin. Obstet. Gynecol.* **28**, 420–429 (2016).
5. Trutnovsky, G., Robledo, K. P., Shek, K. L. & Dietz, H. P. Definition of apical descent in women with and without previous hysterectomy: A retrospective analysis. *PLoS One* **14**, e0213617 (2019).
6. Brunes, M. et al. Vaginal vault prolapse and recurrent surgery: A nationwide observational cohort study. *Acta Obstet. Gynecol. Scand.* **101**, 542–549 (2022).
7. Woodruff, A. J., Roth, C. C. & Winters, J. C. Abdominal sacral colpopexy: Surgical pearls and outcomes. *Curr. Urol. Rep.* **8**, 399–404 (2007).
8. Chaliha, C. & Khullar, V. Management of vaginal prolapse. *Women's. Health (Lond.)* **2**, 279–287 (2006).
9. Alshami, A., Elsayed, M., Ali, E., Eltoukhy, A. E. E. & Zayed, T. Harnessing the Power of ChatGPT for Automating Systematic Review Process: Methodology, Case Study, Limitations, and Future Directions. *Systems* **11**, 351 (2023).

10. Sarkis-Onofre, R., Catalá-López, F., Aromataris, E. & Lockwood, C. How to properly use the PRISMA Statement. *Syst. Rev.* **10**, 117 (2021).
11. Borah, R., Brown, A. W., Capers, P. L. & Kaiser, K. A. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open* **7**, e012545 (2017).
12. Luo, X. et al. Potential Roles of Large Language Models in the Production of Systematic Reviews and Meta-Analyses. *J. Med. Internet Res.* **26**, e56780 (2024).
13. Maher, C. F. et al. Abdominal sacral colpopexy or vaginal sacrospinous colpopexy for vaginal vault prolapse: a prospective randomized study. *Am. J. Obstet. Gynecol.* **190**, 20–26 (2004).
14. Culligan, P. J. et al. A randomized controlled trial comparing fascia lata and synthetic mesh for sacral colpopexy. *Obstet. Gynecol.* **106**, 29–37 (2005).
15. Maher, C. F. et al. Laparoscopic sacral colpopexy versus total vaginal mesh for vaginal vault prolapse: a randomized trial. *Am. J. Obstet. Gynecol.* **204**, 360.e361–367 (2011).
16. Paraiso, M. F. R., Jelovsek, J. E., Frick, A., Chen, C. C. G. & Barber, M. D. Laparoscopic compared with robotic sacrocolpopexy for vaginal prolapse: a randomized controlled trial. *Obstet. Gynecol.* **118**, 1005–1013 (2011).
17. Tate, S. B., Blackwell, L., Lorenz, D. J., Steptoe, M. M. & Culligan, P. J. Randomized trial of fascia lata and polypropylene mesh for abdominal sacrocolpopexy: 5-year follow-up. *Int Urogynecol J.* **22**, 137–143 (2011).
18. Halaska, M. et al. A multicenter, randomized, prospective, controlled study comparing sacrospinous fixation and transvaginal mesh in the treatment of posthysterectomy vaginal vault prolapse. *Am. J. Obstet. Gynecol.* **207**, 301.e301–301.e307 (2012).
19. Freeman, R. M. et al. A randomised controlled trial of abdominal versus laparoscopic sacrocolpopexy for the treatment of post-hysterectomy vaginal vault prolapse: LAS study. *Int. Urogynecology J. Pelvic Floor Dysfunct.* **24**, 377–384 (2013).
20. Svabik, K., Martan, A., Masata, J., El-Haddad, R. & Hubka, P. Comparison of vaginal mesh repair with sacrospinous vaginal colpopexy in the management of vaginal vault prolapse after hysterectomy in patients with levator ani avulsion: a randomized controlled trial. *Ultrasound Obstet. Gynecol. : Off. J. Int. Soc. Ultrasound Obstet. Gynecol.* **43**, 365–371 (2014).
21. Coolen, A. W. M. et al. Laparoscopic sacrocolpopexy compared with open abdominal sacrocolpopexy for vault prolapse repair: a randomised controlled trial. *Int Urogynecol J.* **28**, 1469–1479 (2017).
22. Ow, L. L. et al. RCT of vaginal extraperitoneal uterosacral ligament suspension (VEULS) with anterior mesh versus sacrocolpopexy: 4-year outcome. *Int Urogynecol J.* **29**, 1607–1614 (2018).
23. Ferrando, C. A. & Paraiso, M. F. R. A Prospective Randomized Trial Comparing Restorelle Y Mesh and Flat Mesh for Laparoscopic and Robotic-Assisted Laparoscopic Sacrocolpopexy. *Female Pelvic Med. Reconstructive Surg.* **25**, 83–87 (2019).
24. Galad, J., Papcun, P., Dudic, R. & Urdzik, P. Single-incision mesh vs sacrospinous ligament fixation in posthysterectomy women at a three-year follow-up: a randomized trial. *Bratisl. lekarske listy* **121**, 640–647 (2020).
25. Hemming, C. et al. Surgical interventions for uterine prolapse and for vault prolapse: The two VUE RCTs. *Health Technol. Assess.* **24**, 1–219 (2020).
26. Ferrando, C. A. & Paraiso, M. F. R. A prospective randomized trial comparing Restorelle® Y mesh and flat mesh for laparoscopic and robotic-assisted laparoscopic sacrocolpopexy: 24-month outcomes. *Int Urogynecol J.* **32**, 1565–1570 (2021).
27. van Oudheusden, A. M. J. et al. Laparoscopic sacrocolpopexy versus abdominal sacrocolpopexy for vaginal vault prolapse: long-term follow-up of a randomized controlled trial. *Int Urogynecol J.* **34**, 93–104 (2023).
28. van Oudheusden, A. M. J. et al. Laparoscopic sacrocolpopexy versus vaginal sacrospinous fixation for vaginal vault prolapse: a randomised controlled trial and prospective cohort (SALTO-2 trial). *Bjog* **130**, 1542–1551 (2023).
29. Menefee, S. A. et al. Apical Suspension Repair for Vaginal Vault Prolapse A Randomized Clinical Trial. *JAMA Surg.* **159**, 845–855 (2024).
30. Andy, U. U. et al. Body Image and sexual function improve following prolapse repair. *American J. Obstetrics Gynecology* https://doi.org/10.1016/j.ajog.2025.01.042 (2025).
31. Choong, M. K., Galgani, F., Dunn, A. G. & Tsafnat, G. Automatic evidence retrieval for systematic reviews. *J. Med Internet Res* **16**, e223 (2014).
32. Kim, J. K., Chua, M. E., Li, T. G., Rickard, M. & Lorenzo, A. J. Novel AI applications in systematic review: GPT-4 assisted data extraction, analysis, review of bias. *BMJ Evidence-Based Med.*, bmjebm-2024-113066, https://doi.org/10.1136/bmjebm-2024-113066 (2025).
33. Reason, T. et al. Artificial Intelligence to Automate Network Meta-Analyses: Four Case Studies to Evaluate the Potential Application of Large Language Models. *PharmacoEconomics - Open* **8**, 205–220 (2024).
34. Chelli, M. et al. Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. *J. Med. Internet Res.* **26**, e53164 (2024).
35. Myers, S. et al. Lessons learned on information retrieval in electronic health records: a comparison of embedding models and pooling strategies. *J. Am. Med. Inf. Assoc.* **32**, 357–364 (2025).
36. Asgari, E. et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *NPJ Digit Med.* **8**, 274 (2025).
37. Minozzi, S., Cinquini, M., Gianola, S., Gonzalez-Lorenzo, M. & Banzi, R. The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. *J. Clin. Epidemiol.* **126**, 37–44 (2020).
38. Lai, H. et al. Assessing the Risk of Bias in Randomized Clinical Trials With Large Language Models. *JAMA Netw. Open* **7**, e2412687–e2412687 (2024).
39. Haltaufderheide, J. & Ranisch, R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). *npj Digital Med.* **7**, 183 (2024).
40. Hu, D., Guo, Y., Zhou, Y., Flores, L. & Zheng, K. A systematic review of early evidence on generative AI for drafting responses to patient messages. *npj Health Syst.* **2**, 27 (2025).
41. Mitchell, E., Are, E. B., Colijn, C. & Earn, D. J. D. Using artificial intelligence tools to automate data extraction for living evidence syntheses. *PLOS ONE* **20**, e0320151 (2025).
42. Cosson, M. et al. A study of pelvic ligament strength. *Eur. J. Obstet. Gynecol. Reprod. Biol.* **109**, 80–87 (2003).
43. Torosis, M. & Ackerman, A. L. Patient-reported outcomes in pelvic organ prolapse repair: the missing information needed to inform our understanding of what matters most to patients. *Gynecology Pelvic Med.* **7**, 1–3 (2024).
44. Dyer, O. Transvaginal mesh: FDA orders remaining products off US market. *Bmj* **365**, l1839 (2019).
45. Coolen, A.-L. W. M. et al. The treatment of post-hysterectomy vaginal vault prolapse: a systematic review and meta-analysis. *Int. Urogynecology J.* **28**, 1767–1783 (2017).
46. Dindo, D., Demartines, N. & Clavien, P. A. Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey. *Ann. Surg.* **240**, 205–213 (2004).
47. Zeng, G. On the confusion matrix in credit scoring and its analytical properties. *Commun. Stat. - Theory Methods* **49**, 2080–2093 (2020).
48. Li, M., Gao, Q. & Yu, T. Kappa statistic considerations in evaluating inter-rater reliability between two raters: which, when and context matters. *BMC Cancer* **23**, 799 (2023).

49. Issaiy, M. et al. Methodological insights into ChatGPT's screening performance in systematic reviews. *BMC Med. Res. Methodol.* **24**, 78 (2024).
50. Wohlin, C. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering* Article 38 (Association for Computing Machinery, London, England, United Kingdom, 2014).
51. Lai, H. et al. Language models for data extraction and risk of bias assessment in complementary medicine. *npj Digital Med.* **8**, 74 (2025).
52. Sterne, J. A. C. et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *Bmj* **366**, l4898 (2019).

## Acknowledgements

The authors acknowledge the Yonsei University Medical Library for support with the literature search. ChatGPT (OpenAI, San Francisco, CA, USA) was used to assist with English language refinement, under the authors' supervision.

## Author contributions

Y.P. - Conceptualization, literature search, study selection, data extraction, quality assessment, preparation of figures and tables, drafting of the manuscript H.Z. - Statistical analysis, data synthesis, preparation of figures and tables, methodological consultation S.W.B. - Literature search, data extraction, validation, interpretation of findings, Project administration.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41746-026-02431-w.

**Correspondence** and requests for materials should be addressed to Sang Wook Bai.