

Systematic Review

Open Access



Artificial intelligence versus human researcher performance for systematic literature searches: a study focusing on the surgical management of base of thumb arthritis

Ishith Seth¹ , Bryan Lim¹ , Yi Xie¹, Richard J. Ross^{1,2}, Roberto Cuomo³ , Warren M. Rozen¹

¹Department of Plastic and Reconstructive Surgery, Peninsula Health, Victoria 3199, Australia.

²Faculty of Medicine and Surgery, Monash University, Victoria 3004, Australia.

³Department of Plastic and Reconstructive Surgery, University of Siena, Siena 53100, Italy.

Correspondence to: Dr. Ishith Seth, Department of Plastic and Reconstructive Surgery, Peninsula Health, Victoria 3199, Australia. E-mail: ishithseth1@gmail.com

How to cite this article: Seth I, Lim B, Xie Y, Ross RJ, Cuomo R, Rozen WM. Artificial intelligence versus human researcher performance for systematic literature searches: a study focusing on the surgical management of base of thumb arthritis. *Plast Aesthet Res*. 2025;12:1. <https://dx.doi.org/10.20517/2347-9264.2024.99>

Received: 28 Jul 2024 **First Decision:** 9 Oct 2024 **Revised:** 12 Dec 2024 **Accepted:** 20 Dec 2024 **Published:** 6 Jan 2025

Academic Editor: Elias Polykandriotis **Copy Editor:** Ting-Ting Hu **Production Editor:** Ting-Ting Hu

Abstract

Aim: In the digital age, artificial intelligence (AI) platforms have gradually replaced traditional manual techniques for information retrieval. However, their effectiveness in conducting academic literature searches remains unclear, necessitating a comparative assessment. This study examined the efficacy of AI search engines (Elicit, Consensus, ChatGPT) vs. manual search for literature retrieval, focusing on the surgical management of trapeziometacarpal osteoarthritis.

Methods: The study was executed per the Cochrane Handbook for Systematic Reviews and PRISMA guidelines. AI platforms were given relevant keywords and prompts, while manual searches used PubMed, Cochrane CENTRAL, Web of Science, and Scopus databases from January 1901 to April 2024. The study focused on English-language randomized controlled trials (RCTs) comparing surgical management of trapeziometacarpal osteoarthritis (TMCJ OA). Two independent evaluators screened and extracted data from the studies. Primary outcomes involved the quality and relevancy of studies chosen by both search methods, evaluated by false positive rates and number of studies, including outcomes of interest.



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.



Results: The manual search yielded the most results (6,018), followed by Elicit (4,980), Consensus (3,436), and ChatGPT (6). Elicit identified the highest number of RCTs (205) but also had the greatest false positive rate (94%). Ultimately, the manual search identified 23 suitable studies, Elicit found 10, Consensus found 9, and ChatGPT identified only 1. No additional studies were found by AI search engines that were not discovered in the manual search.

Conclusion: The findings highlight the potential advantages and drawbacks of AI search engines for literature searches. While Elicit was prone to error, Consensus and ChatGPT were less comprehensive. Significant enhancements in the precision and thoroughness of AI search engines are required before they can be effectively utilized in academia.

Keywords: Artificial intelligence, human, researcher, systematic review, searches

INTRODUCTION

In an era of digital transformation, traditional literature search methods are being supplemented and replaced by artificial intelligence (AI)-based platforms^[1,2]. These include software such as Elicit, Consensus, and ChatGPT, which have been proposed as valuable tools for expediting information retrieval and facilitating the dissemination of medical information. In this domain, ChatGPT has received considerable commentary on its potential in academia across a range of topics, from osteoarthritis to cosmetic surgery, with major concerns about its ability to correctly identify the source of its knowledge, albeit surprisingly accurately. Although an interesting avenue to explore, the comparative efficiency and accuracy of different chatbots in locating and sourcing information compared with traditional human-initiated searches have not been explored^[3-5].

Trapeziometacarpal osteoarthritis (TMCJ OA) is a common condition among the elderly that significantly limits thumb movement and functionality necessary for everyday tasks^[6]. Management of TMCJ OA begins medically in mild cases, progressing to operative intervention only when anti-inflammatory and pain relief prove insufficient. Multiple surgical and non-surgical treatment modalities are available, but their comparative effectiveness is unclear, especially surgical ones^[7]. This gap in the literature leaves healthcare professionals and patients in a predicament during the decision-making process, and a systematic review and meta-analysis is likely an effective means to summarize information and facilitate a consensus in the plastics and orthopedics community.

With this in mind, we carried out a comparative study that scrutinized the performance of Elicit, Consensus, and ChatGPT with manual human literature search methods for the management of TMCJ OA. The primary outcomes were the ability to identify publications with higher-level evidence, as well as the number of publications and their relevance. The outcomes of interest specific to TMCJ OA were also investigated to inform the potential role and value of AI for conducting systematic reviews.

METHODS

The current study adhered to the Cochrane Handbook of Systematic Reviews of Interventions and the preferred reporting items for systematic reviews and meta-analyses (PRISMA) statement guidelines throughout all stages^[8]. The study was registered on PROSPERO, the International Prospective Systematic Review (CRD420431089). The primary objective of this study was to evaluate the performance of AI-based platforms (Elicit, Consensus and ChatGPT) against human experts for conducting a literature search for a systematic review on the base of thumb arthritis treatments. Institutional ethical approval was not required since this study did not involve human subjects. These three AI platforms were selected for their

prominence and widespread adoption in the research community at the time of the study. Elicit, developed by Ought, was chosen for its specialized focus on scientific literature search and summarization. Consensus, created by Consensus.app, was selected for its ability to aggregate and analyze scientific papers. ChatGPT, an advanced language model by OpenAI, was included due to its versatility in understanding and generating human-like text across various domains, including scientific literature. The standard ChatGPT version was used to minimize any potential bias and maintain methodological consistency with other studies.

Literature search strategy

To ensure consistency and comparability between AI and human-based searches, a uniform search strategy was employed. AI-based platforms were prompted with different arrays of keywords and prompted ten times, and all pages were screened for any potential studies [Supplementary Figures 1-3]. The authors (IS and GB) validated the suitability and relevance of the studies initially sourced by the AI tools, and the total search results are shown in Table 1. This entailed identifying randomized controlled trials (RCTs), false positive RCTs, prospective studies, and deciding whether a study was included or excluded without any assistance from the AI tools. The manual search strategy encompassed a combination of pertinent keywords and MeSH terms associated with TMCJ OA, which included thumb OR trapezio-metacarpal OR trapeziometacarpal OR trapezoid-metacarpal OR trapezoidmetacarpal OR trapezium OR carpal OR metacarp OR carpo-metacarpal OR “metacarpophalangeal joint” OR “carpometacarpal joint” OR trapezium) AND (osteoarthritis OR osteoarth OR “joint disease” OR arthropathy) AND (“basal joint arthroplasty” OR “Arthroscopic Resection Arthroplasty” OR “resection arthroplasty” OR trapeziectomy OR “trapezio-metacarpal arthrodesis. The manual literature search was conducted using Medline (via PubMed), Cochrane Library, Web of Science, and Scopus, covering the period from January 1901 to April 2024. Additionally, the reference lists of relevant articles were manually reviewed. Supplementary Materials includes a comprehensive overview of the search strategies employed.

Eligibility criteria

Studies were included if they met the following criteria: (1) RCTs, which compared surgical management of TMCJ OA; (2) they were conducted on human subjects; (3) they were written in the English language. There were no restrictions on the minimum number of cases or duration of follow-up. Studies were excluded if they were noncomparative, included other joints, or did not report the outcomes of interest. Animal studies, review articles, case reports, conference abstracts, non-English language studies, and duplicate references from the analysis were excluded.

Study selection

Titles and abstracts of studies identified during the search were imported into Endnote X20 for preliminary screening. Full texts of potentially relevant papers were further screened using the eligibility criteria. Two independent reviewers (IS and GB) did this, and any disparity in either selecting eligible studies or assessing findings between the two reviewers was resolved through consultation with the rest of the authors.

Data extraction

Two independent authors (IS and GB) extracted data into an Excel spreadsheet with the following parameters: treatment modalities, age, gender, follow-up, level of evidence, inclusion criteria of studies, primary outcomes, and conclusion. A false positive analysis considered cases where AI included RCTs outside the scope of surgical management.

Risk of bias assessment

The methodological quality of each study was assessed using Cochrane risk-of-bias (ROB) tool for

Study ID	Study arms, N	Age, mean (SD)	Male, N (%)	Surgical intervention	Follow up	Level of evidence	Inclusion criteria	Primary outcomes	Conclusion
Belcher 2000	Trapeziectomy, 19	63 (2)	1 (5.26%)	Trapeziectomy by posterior approach vs. T + LRTI (APL-FCR-APL)	14 months	I	1. Adults undergoing trapeziectomy for osteoarthritis of the thumb TMJ were entered into this study between March 1996 and July 1998	1. Pain 2. Physical function 3. Adverse events	Both groups expressed equal satisfaction with the operation and there were no significant differences between the two treatment groups. Simple trapeziectomy is an effective operation for osteoarthritis at the base of the thumb and the addition of a ligament reconstruction was not shown to confer any additional benefit
	Trapeziectomy and LRTI, 23	58 (1)	4 (17.39%)						
Belcher 2001	Trapeziectomy, 13	59 (8)	7 (53.8%)	Trapeziectomy by posterior approach vs. Trapeziectomy + Permacol porcine xenograft	6 months	I	1. Patients undergoing trapeziectomy for osteoarthritis of the thumb trapeziometacarpal joint were entered into the study between April and December 1999	1. Pain 2. Physical function 3. Satisfaction 4. Adverse events	Permacol patients reported greater pain and were less satisfied with their operations than control patients. We conclude that interposition of Permacol is detrimental to the results of trapeziectomy
	Trapeziectomy and Permacol porcine xenograft, 13	59 (9)	7 (53.8%)						
Brennan 2020	Trapeziectomy, 14	75 (6)	3 (21.43%)	Trapeziectomy by posterior approach vs. Trapeziectomy + LRTI (½FCR-MT)	17 years	I	1. Patients with osteoarthritis of the CMCJ of the thumb were recruited	1. Pain 2. Physical function 3. Satisfaction	Even at 17 years, there is no significant benefit of LRTI over trapeziectomy alone for thumb carpometacarpal joint osteoarthritis
	Trapeziectomy and LRTI, 20	75 (6)	5 (25%)						
Corain 2016	Trapeziectomy and HAD, 64	63 (12)	-	Trapeziectomy + HDA vs. Trapeziectomy + LR (APL-MT-FCR)	6.6 years	I	1. No previous surgeries affecting the same arm 2. No diabetes or connective tissue disorders; symptomatic stage 3 or 4 osteoarthritis according to the Eaton classification	1. Pain 2. Physical function 3. Adverse events	We demonstrate that the trapezium excision and bone space distraction technique require a smaller incision, a shorter surgical time, an easier surgical technique, and a less painful recovery, maintaining overlapping levels of functional restore
	Trapeziectomy and LR (APL-MT-FCR), 56	-	-						
De smet 2004	Trapeziectomy	61.5 (10.2)	0	Trapeziectomy vs. Trapeziectomy + LRTI (FCR-MT)	26 months	I	1. Patients suffered from painful primary osteoarthritis of the carpometacarpal joint of the thumb not responding to conservative treatment	1. Pain 2. Physical function	Simple trapeziectomy is a good procedure, especially for elderly patients requiring not much force
	Trapeziectomy and LRTI	58 (6.3)	0						
Field 2007	Trapeziectomy, 32	-	4 (12.5%)	Trapeziectomy by posterior approach vs. Trapeziectomy + LRTI (½FCR-MT)	1 year	I	1. Patients with osteoarthritis of the carpometacarpal joint of the thumb of Eaton and Glickel Grade III or IV 2. Who had not responded to conservative treatment were recruited into the study between 2001 and 2003	1. Pain 2. Physical function 3. Adverse events	In conclusion, this study suggests that there is no benefit to suspension with an FCR sling after trapeziectomy
	Trapeziectomy and LRTI, 33	-	5 (15.15%)						
Gangopdhyay 2012	Trapeziectomy, 53	57 (6)	0	Trapeziectomy by posterior approach vs. Trapeziectomy +	6 years	I	1. Women with painful trapeziometacarpal	1. Pain 2. Adverse	The outcomes of these 3 variations of trapeziectomy were similar after a minimum follow-up of 5 years.
	Trapeziectomy with	57 (6)	0						

	palmaris longus interposition, 46			tendon interposition (PL)			osteoarthritis who had failed to respond to the nonoperative treatment were recruited between 1992 and 2001	events	There appears to be no benefit to tendon interposition or ligament reconstruction in the longer term
	Trapeziectomy with LRTI, 54	57 (6)	0						
Gerwin 1997	Trapeziectomy with Ligament Reconstruction, 11	-	-	Trapeziectomy + LR (½FCR-MT-Minimitek) vs. Trapeziectomy + LRTI (½FCRMT-Minimitek)	23 months	II	1. Patients undergoing trapeziectomy for osteoarthritis of the thumb trapeziometacarpal joint	1. Physical function 2. Satisfaction	Tendon interposition after ligament reconstruction basal joint arthroplasty does not improve the function of the thumb and necessitates a longer surgical incision and a technically more difficult operation
	Trapeziectomy with LRTI, 9	-	-						
Hansen 2013	DLC all-poly cup, 14	56 (11)	2 (14.29%)	Elektra uncemented cup vs. Elektra cemented cup	2 years	I	1. Eaton-Glickel stage-2 or -3 TM joint OA in patients over 18 years of age where nonoperative treatment had failed. 2. OA staging was based on a combination of conventional radiographs and CT scans evaluated by one observer	1. Adverse events	Early implant fixation and clinical outcome were equally good with both cup designs. This is the first clinical RSA study on trapezium cups, and the method appears to be clinically useful for the detection of loose implants
	Electra screw cup, 10	60 (12)	1 (7.69%)						
Hart 2006	trapeziometacarpal arthrodesis	59 (8)	13 (35.14%)	Arthrodesis (K-wire) vs. T + LRTI (½FCR-MT -K-wire)	-	I	1. Patients with primary osteoarthritis of stage 4 according to Eaton and Littler of the first carpometacarpal joint	1. Adverse events	The after-treatment in patients undergoing arthroplasty lasted longer than in patients after the arthrodesis. It is caused by more complex surgery during Epping's procedure. But the outcomes become similar over a longer period. At the final follow-up control after arthroplasty, only older patients subjectively appreciated better functional performance. After this experience, we reserve the arthrodesis for younger active and arthroplasty for older patients
	Trapeziectomy and LRTI	59 (8)							
Kriegs-au 2005	Trapeziectomy with LR, 26	-	-	Trapeziectomy + LR (½FCR-MT) vs. Trapeziectomy + LRTI (½FCR-MT)	4 years	II	1. Patients undergoing trapeziectomy for osteoarthritis of the thumb trapeziometacarpal joint	1. Pain 2. Physical function 3. Satisfaction 4. Adverse events	Tendon interposition does not affect the outcome after the ligament reconstruction for the treatment of osteoarthritis of the thumb carpometacarpal joint. Furthermore, proximal migration of the thumb metacarpal does not appear to influence the functional outcome
	Trapeziectomy with LRTI, 26	-	-						
Marks 2017	Trapeziectomy with LRTI, 29	64 (8)	3 (10%)	Trapeziectomy + LRTI (½FCR-APL-½FCR) vs. Trapeziectomy + Graft Jacket allograft	1 year	I	1. If they were diagnosed with CMC I OA and met indications for trapeziectomy with suspension-interposition arthroplasty	1. Pain 2. Quality of life 3. Adverse events	The use of the FCR tendon or allograft for trapeziectomy with suspension interposition arthroplasty in patients with CMC I OA leads to similar outcomes with more complications, mainly tendon irritations, associated with the latter. Therefore, we only use the allograft in cases of severe instability requiring a larger amount of suspension-interposition material or for revision procedures after failed suspension interposition with the FCR tendon
	Trapeziectomy with Graft Jacket allograft, 31	65 (8)	6 (19%)						

Morais 2021	Trapeziectomy with suture-button suspensionplasty, 37 Ligament reconstruction and tendon interposition, 39	61.8 (7.8) 61.1 (7.4)	4 (10.8%) 2 (5.2%)	Trapeziectomy with suture-button suspensionplasty vs. ligament reconstruction and tendon interposition	40 months	I	1. Patients with TMC arthritis	1. Pain 2. Physical function 3. Range of movement 4. Quality of life 5. Adverse events	The results are related to the hypothesis suggested by biomechanical studies that revealed better initial load-bearing profile and maintenance of trapezial space following serial loading in cadaver models
Ritchie 2008	Trapeziectomy by anterior approach, 20 Trapeziectomy by posterior approach, 20	59 (7) 64 (9)	6 (30%) 5 (25%)	Trapeziectomy by anterior approach vs. Trapeziectomy by posterior approach	33 months	I	1. Adults undergoing trapeziectomy for osteoarthritis of the TMJ were entered into this study between January 2001 and October 2002	1. Pain 2. Physical function 3. Satisfaction 4. Adverse events	Trapeziectomy is a good method of treating osteoarthritis of the thumb base, but outcomes for the anterior approach are equally good or better than with the posterior
Salem 2012	Trapeziectomy, 59 Trapeziectomy and LRTI, 55	- -	8 (13.56%) 9 (16.36%)	Trapeziectomy + LR (½FCR-MT) vs. Trapeziectomy + LRTI (½FCR-MT)	6 years	I	1. Patients with painful trapeziometacarpal joint osteoarthritis who had not responded to nonoperative treatment were recruited during 2002-2005	1. Pain 2. Physical function 3. Adverse events	This study does not provide evidence to support the use of LRTI and temporary K-wire stabilization after trapeziectomy
Salibi 2019	Trapeziectomy, 10 Carpometacarpal denervation, 35	61 (9) 58 (13)	5 (50%) 6 (17.14%)	Trapeziectomy vs. carpometacarpal denervation	5 years	II	1. A diagnosis of CMC arthritis as well as the failure of nonsurgical management with antiinflammatories, bracing, or corticosteroid injections	1. Pain 2. Quality of life 3. Satisfaction 4. Physical function	There was no difference between the two treatments. First CMCJ denervation does not appear to be superior to trapeziectomy. However, the advantage of rapid rehabilitation makes it more favoured by patients but at the expense of a 30% reoperation rate
Sanchez-Flo 2020	Partial Trapeziectomy, 17 Total Trapeziectomy, 17	60.5 (9.8) 61 (8.9)	4 (23.5%) 2 (11.8%)	Partial vs. Total trapeziectomy with interposition arthroplasty	1 year	III	1. Patients with isolated TMOA grade II to III (Eaton-Littler) with articular pain and loss of hand function	1. Physical function 2. Pain 3. Quality of life 4. Adverse events	We cannot conclude that partial trapeziectomy provides an advantage over total trapeziectomy at 1 year after surgery. Although trapeziometacarpal space was substantially preserved in the partial trapeziectomy group at 12 months, this difference was not statistically or clinically significant
Spekreijse 2015	Burton-Pellegrini technique, 36 Weilby technique, 36	65 (9) 64 (9)	- -	Trapeziectomy + LRTI (½FCR-MT) vs. Trapeziectomy + LRTI (½FCR-APL-½FCR)	5 years	I	1. If they had symptoms of stage IV OA of both TMC and STT joints with functional impairment of daily activities after the failure of conservative therapy	1. Pain 2. Physical function 3. Satisfaction 4. Adverse events	This study showed that improved function, strength, and satisfaction obtained at 1 year after trapeziectomy with LRTI with or without the use of a bone tunnel for stage IV TMC thumb osteoarthritis was maintained after 5 years
Spekreijse 2016	Trapeziectomy and LRTI, 21 Arthrodesis, 17	59.5 (6.3) 59.7	0 0	Arthrodesis (plate/screws) vs. T + LRTI (½FCR-APL-½FCR)	5 years	IV	1. Women older than 40 years with primary, symptomatic OA of the thumb TMC joint, stage II or III by the Eaton and	1. Pain 2. Physical function 3.	Trapeziectomy with LRTI leads to better pain reduction and functional outcome after between 1 and 5 years compared with trapeziometacarpal arthrodesis in women over 40 years old with OA

(6)					Glickel classification	Satisfaction 4. Adverse events	stages II to III		
Tagil 2002	Trapeziectomy with LRTI, 13	62 (13.5)	-	Trapeziectomy + LRTI (APL-FCR-APL) vs. Trapeziectomy + Swanson silastic implant	4 years	I	1. Patients with radiographic osteoarthritis and disabling pain agreed to participate in the study and were operated on between 1991 and 1995. 2. All had undergone failed conservative treatment including an orthosis	1. Pain 2. Satisfaction 3. Adverse events	Both methods gave good, but not complete, pain relief and neither produced better results than the other in the short term
	Trapeziectomy with Swanson silastic implant, 13	62 (13)	-						
Thorkildsen 2019	Uncemented joint replacement (Elektra), 20	64 (5)	6 (30%)	Uncemented joint replacement (Elektra) vs. trapeziectomy (with ligament reconstruction and tendon interposition, LRTI)	2 years	I	1. Symptomatic idiopathic osteoarthritis of the CMC1 joint 2. Patients over 18 years of age with general good health	1. Physical function 2. Quality of life 3. Adverse events 4. Time to revision	The place for joint replacements in the treatment of symptomatic CMC1 osteoarthritis is still not clear, whereas trapeziectomy with LRTI was a reliable procedure in this trial. Further comparative studies using implants with documented good long-term function and longer follow-up will be required to finally ascertain whether, or which, joint replacement is superior
	trapeziectomy with LRTI, 20	61 (6)	6 (30%)						
Vermeulen 2014	Trapeziectomy and LRTI, 21	59 (6.3)	-	Arthrodesis (plate/screws) vs. T + LRTI (½FCR-APL- ½FCR)	1 year	I	1. Patients with impaired function who failed to improve after nonsurgical treatment 2. Who had stage-II or III primary osteoarthritis of the trapeziometacarpal joint according to the classification system of Eaton and Glickel	1. Satisfaction 2. Adverse events	Women who are forty years or older with trapeziometacarpal osteoarthritis have fewer moderate and severe complications after trapeziectomy with ligament reconstruction and tendon interposition and are more likely to consider the surgery again under the same circumstances than are those who undergo arthrodesis. Twelve months after surgery, the PRWHE and DASH scores were similar in both groups. We do not recommend routine use of arthrodesis with plate and screws in the treatment of women who are forty years or older with stage-II or III trapeziometacarpal osteoarthritis
	Arthrodesis, 17	59 (6)	-						
Vereulen 2014 (1)	Burton-Pellegrini technique, 36	64.7 (9.1)	-	Trapeziectomy + LRTI (½FCR-MT) vs. Trapeziectomy + LRTI (½FCR-APL-½FCR)	1 year	I	1. Women aged 40 years or older 2. With stage IV osteoarthritis	1. Pain 2. Satisfaction 3. Range of motion 4. Physical function 5. Adverse events	After the bone tunnel technique, patients have better function and less pain 3 months after surgery than do those in the none bone tunnel group, which indicates faster recovery. However, 12 months after surgery, the functional outcome was similar. Because of faster recovery, we prefer the bone tunnel technique in the treatment of stage IV osteoarthritis
	Weilby technique, 36	63.5 (8.5)	-						

N: Number; SD: standard deviation; LRTI: ligament reconstruction and tendon interposition; CT: computed tomography; DASH: Disabilities of the Arm, Shoulder, and Hand; MT: metacarpal tunnel; APL: abductor pollicis longus; CMCJ: carpometacarpal joint; DLC: De la Caffinière; FCR: flexor carpi radialis; T: trapeziectomy; LR: ligament reconstruction; K-wire: Kirschner wire; PL: palmaris longus; HDA: hematoma distraction arthroplasty; CMC: carpal metacarpal; OA: osteoarthritis; PRWHE: Patient-Related Wrist/Hand Evaluation; RSA: radiostereometric analysis; TMJ: trapeziometacarpal joint; TMC: trapeziometacarpal; TMAO: trapeziometacarpal osteoarthritis.

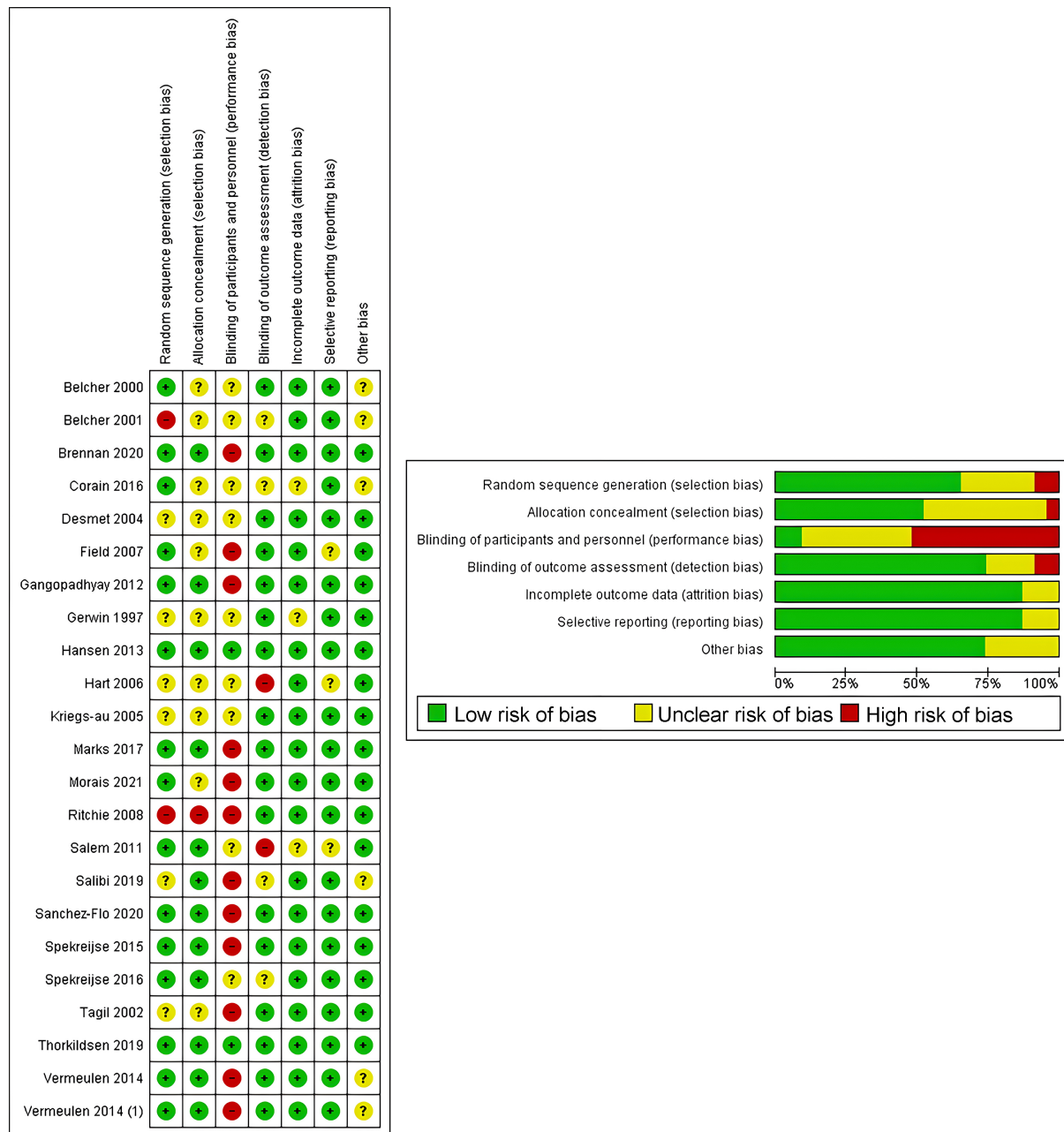


Figure 1. Risk of bias of all included studies.

randomized trials [Figure 1]. The RoB tool addresses the following biases: random sequence generation, bias due to deviations from intended interventions, bias due to incomplete outcome data, bias in the measurement of the outcome, and selective reporting. The items were assessed as “low risk”, “high risk”, or “some concerns”. We used the original RoB tool rather than the updated RoB 2 tool, as our research team had extensive experience with the original tool, ensuring consistent and accurate assessments, and wanted to maintain comparability with other systematic reviews in our area of research that predominantly used the original tool. We acknowledge that the RoB 2 tool offers a more nuanced approach, particularly for assessing bias in subjective outcomes and open-label studies. However, our use of the original RoB tool may have resulted in slightly more conservative bias assessments. This conservative approach strengthens the

reliability of our findings, as it is less likely to underestimate potential biases in the included studies.

RESULTS

The manual search executed by human authors yielded 6,018 initial results, followed by 4,980 results from Elicit, 3,436 from Consensus, and lastly, only 6 from ChatGPT, [Table 1](#). Elicit found 205 RCTs, while the manual search found 63, Consensus returned 42, and ChatGPT identified one [[Figures 2-4](#)]. For prospective studies, the manual search yielded 1,852 results, followed by 1,123 from Elicit, 963 from Consensus, and one from ChatGPT. Elicit's broader selection of RCTs stems from its indiscriminate inclusion of all studies discussing base of thumb arthritis regardless of comparison with surgical management strategies, and its search focused largely on non-surgical management. Lastly, Elicit had the highest false positives at 94%, followed by consensus at 76%, human researchers at 43%, and ChatGPT at 0%, [Table 1](#).

Characteristics of included studies

A total of 23 RCTs^[9-31] from all searches were eligible for inclusion in this study, as shown in [Table 2](#) and [Figure 4](#). The manual search method covered all 23 studies, followed by Elicit, which found 10 studies, then Consensus, which uncovered 9, while ChatGPT identified only one study. The manual search identified all the studies found by the AI search engine searches, and there was no additional benefit from other searches. By the end of the screening process, manual search led to 5,994 excluded papers, followed by Elicit with 4,969, Consensus with 3,427, and ChatGPT excluding 5.

[Table 2](#) summarizes the characteristics of the included studies, including the application of intraoperative adjuvants, the specific muscles implicated, and the surgical approach adopted. In total, 1,335 procedures occurred across 23 studies, 489 of which were trapeziectomies with ligament reconstruction tendon interposition (LRTI) [[Table 2](#)]. Participants were, on average, 49.83 years old and were followed up for an average duration of 3.31 years.

Comparison of AI search engines

Compared with manual search, Consensus and Elicit overlooked studies by Marks, Morais, Sanchez-Flo, and Thorkildsen, which exhibited evidence levels I, I, III, and I, respectively. Elicit displayed similar levels of omission, failing to include studies by Gerwin, Ritchie, and Salem at evidence levels II, I, and I, respectively. While ChatGPT struggled to locate most of the literature, it succeeded in identifying the study by Field *et al.*, a level I evidence study that was neglected by the other AI search engines^[14].

Number of studies included by each AI search engine

While Elicit and Consensus demonstrated analogous capacities for identifying studies with comparable levels of evidence, Elicit displayed superior capability for identifying a greater number of included studies (totaling 522 patients) compared to Consensus (totaling 438 patients).

Outcomes

Pain

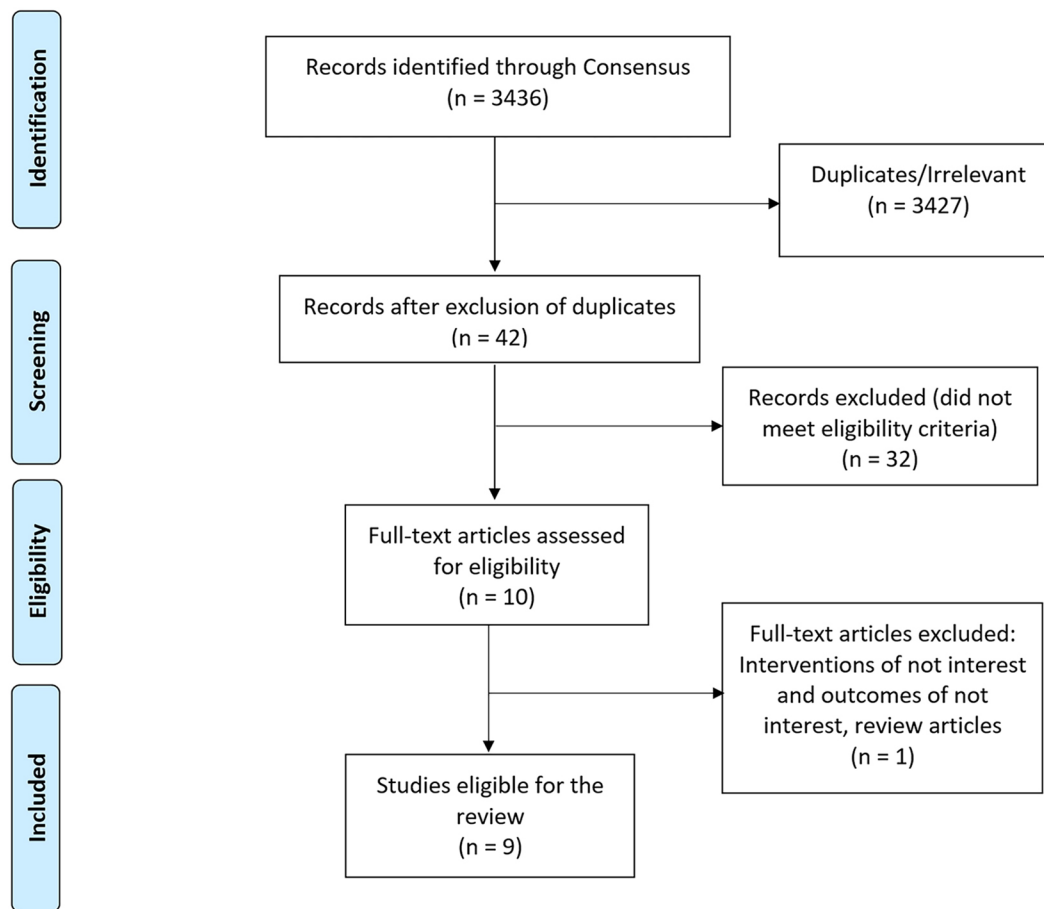
Of the eighteen studies identified by manual searches evaluating pain as an outcome, seven were included by Consensus, and eight by Elicit. Consensus and Elicit found five of the same studies, while ChatGPT found none.

Physical function

Of seventeen studies identified by manual searches that explored physical function as an outcome, eight were found by Consensus and Elicit, although just five were common to both Consensus and Elicit. ChatGPT identified none.

Table 2. Comparison of artificial intelligence and human in literature search

	Consensus	Elicit	ChatGPT	Human manually
Total Search results	3,436	4,980	6	6,018
Randomized controlled trials in search	42	205	1	63
False positive randomized controlled trials, N (%)	32 (76%)	193 (94%)	0	27 (43%)
Prospective studies in search	963	1,123	1	1,852
Included studies	9 (1-9)	10 (1-3, 5, 8-13)	1 (17)	23 (1-23)
Excluded studies	3,427	4,970	5	5,995

**Figure 2.** PRISMA figure of consensus platform search. PRISMA: Preferred reporting items for systematic reviews and meta-analyses.

Adverse Events

Nineteen studies identified by the manual search reported adverse events as an outcome. Seven of these were found by Consensus, and nine by Elicit. Once more, five studies were common between Elicit and Consensus, while ChatGPT found none.

Quality of Life

Five studies identified by manual searches reported the quality of life as an outcome. None of these were included by Consensus, while Elicit identified four. ChatGPT identified none.

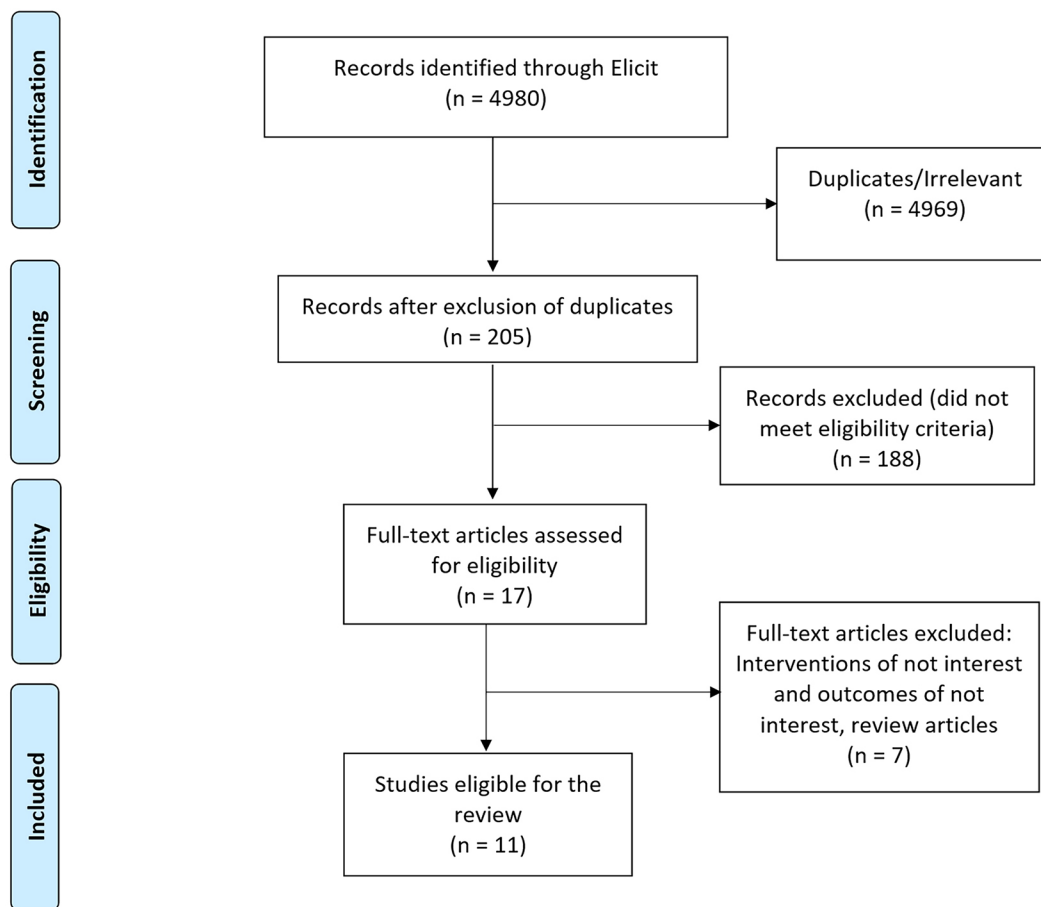


Figure 3. PRISMA figure of Elicit platform search. PRISMA: Preferred reporting items for systematic reviews and meta-analyses.

Satisfaction

Eleven studies identified on manual searches reported satisfaction as an outcome. Seven of these were found by Consensus and five were identified by Elicit. Five studies were common between Elicit and Consensus.

Range of movement

Two included studies addressed the range of motion as an outcome of manual searching. Consensus and Elicit each identified one, but none were common, and ChatGPT found one.

DISCUSSION

This case study is the first to explore the comparative performance between human-initiated and AI-initiated literature searches. These findings demonstrate AI platforms currently have poor proficiency for use in academia, especially ChatGPT, which performed poorly across all domains and outcomes. Although Elicit came the closest to mimicking human precision of the initial search, manual searches were far superior to all AI literature search engines in terms of the number of studies identified and their specificity to the subject of TMCJ OA. AI engines also overlooked studies extracted from the manual search and lacked precision in the subject of the search, evidenced by high false positive identification rates.

Interestingly, the average age of participants across the 23 included studies was 49.83 years, notably younger than the typical patient population seen in most CMC1 (first carpometacarpal joint) osteoarthritis

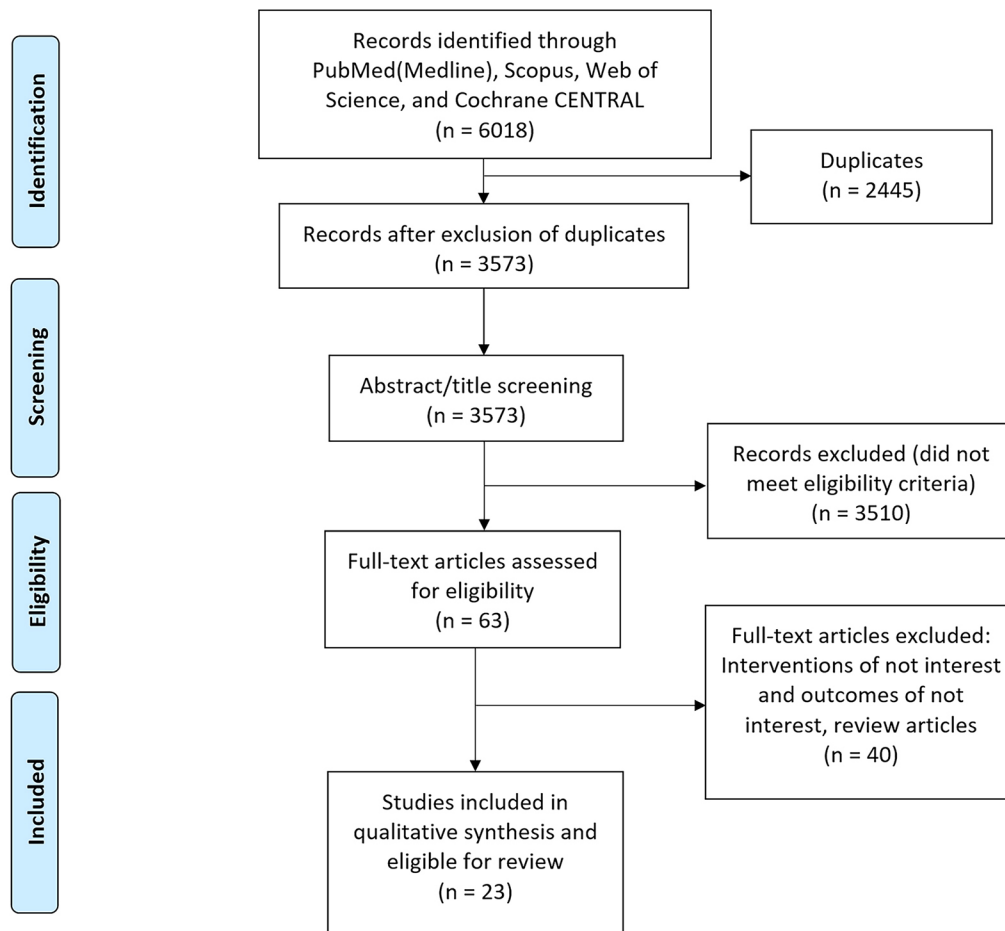


Figure 4. PRISMA figure of manual search. PRISMA: Preferred reporting items for systematic reviews and meta-analyses.

publications. This relatively young cohort raises essential questions about the generalisability of the study results to the broader TMCJ OA population, which typically presents in older adults. Including younger patients may reflect a trend toward earlier surgical intervention, possibly due to increased awareness or changes in treatment paradigms. Further investigation is warranted to understand the implications of this age discrepancy for treatment outcomes and long-term prognosis for TMCJ OA patients.

Upon inspecting the number of relevant studies produced, Elicit was the most comprehensive AI search engine, albeit only surpassing Consensus by a single article. However, most RCTs identified by Elicit were tangential and addressed various topics beyond management strategies. As this methodology has never been applied since the inception of large language models (LLM), these findings cannot be discussed and contextualized in other studies. Despite the promise of AI to replicate laborious manual tasks, the results herein are disappointing and suggest that LLMs currently have no applicability in relieving the burdensome process of literature searching and screening. This study shows that LLMs could do a disservice to the scientific community by excluding publications typically deemed important and including irrelevant ones in initial searches. This misalignment with the topic of discussion led to a 94% false positive rate within the search, compared to a human false positive rate of 43%. While Consensus elicited nearly 1,500 fewer studies than Elicit, it included nine of the ten studies identified by Elicit, yielding a significantly lower false positive rate of 73%. Although Elicit identified the most publications overall, its search was the least precise and

most inefficient of all AI search engines. No AI-driven engine could identify studies not included in the human search, indicating that human searches were the most precise and had a very low false negative rate.

Concerning primary outcomes, Elicit emerged as the sole AI search engine capable of identifying RCTs addressing all relevant primary outcomes. However, Consensus failed to uncover any studies focused on quality of life, although AI search engines could identify more than one study, each discussing the range of motion. ChatGPT exhibited the least effective performance, locating only one study addressing two of the six primary outcomes, and finding a volume of studies that was small in comparison to manual searches by authors and AI search engines^[32,33]. Overall, AI search engines were inferior to manual searching, highlighting a shortcoming in their algorithms for sourcing comprehensive, high-quality literature relevant to the research topic^[2]. The indiscriminate data retrieval by AI search engines in this study points to a potential for them to produce erroneous information outputs, due to a lack of precision and hierarchical structure during information gathering and organization^[2]. Peering into the mind of an algorithm, it is clear from these results that these deficits could account for the erroneous or outdated responses sometimes reported in previous studies. Therefore, for AI to be a viable tool in academic literature searches, substantial improvements are needed in categorization, publication filtering, bias detection, database integration, and ethical data handling.

This study explores the use of AI for literature searches, highlighting the significant improvements required for AI tools to be feasibly incorporated into literature searches for the creation of academic content. These improvements may be grouped into a few main broad categories that should be considered. Paramount among these is ensuring “reproducibility”, which is the cornerstone of academic research and literature searches, as exemplified by the dual-reviewer approach outlined in the PRISMA guidelines. Current AI tools fall short in accuracy and comprehensiveness. Additionally, users may ask AI search engines the same question multiple times and receive different answers informed by different sources^[34,35,36]. A future AI system must be able to recognize and understand context, academic language, and abbreviations to meet the reproducibility standard. Moreover, AI must develop a nuanced understanding of academic context, hierarchy, and the goals of a literature search to match or surpass human researchers in precision and thoroughness^[37].

Secondly, AI should transcend simple keyword identification to gain a deeper semantic understanding of academic papers. This includes comprehending study objectives, methodologies employed, and resultant conclusions. Such advancements would enhance AI’s ability to categorize, filter, and rank search results based on criteria such as relevance, currency, citation frequency, and the publishing journal’s impact factor. Thirdly, improvements are needed to discern and neutralize potential biases, including those related to geographic location, authorship, or publication prestige, to ensure fair data representation. Given the dynamic nature of academia, with its continuous generation of novel knowledge and methodologies, AI platforms must be equipped for accessible, ongoing learning and enhancement.

In conclusion, This study found AI tools such as Elicit, Consensus, and ChatGPT were inaccurate and lacked comprehension compared with human-initiated literature searches. These tools need to evolve beyond simple keyword identification toward a nuanced understanding of academic hierarchy and context. Therefore, AI’s integration into academic literature searches demands substantial enhancements in its understanding of academic context and hierarchy, fulfilling the crucial reproducibility criterion and aligning it with the rigorous standards of human-conducted research.

DECLARATIONS

Authors' contributions

Conceptualization: Seth I, Lim B, Xie Y, Rozen WM

Methodology: Seth I, Ross R, Rozen WM

Data analysis: Seth I, Lim B, Xie Y

Manuscript writing: Seth I, Lim B, Xie Y, Ross RJ, Cuomo R, Rozen WM

Manuscript editing: Seth I, Lim B, Xie Y, Ross RJ, Cuomo R, Rozen WM

All authors edited and approved the final manuscript.

Availability of data and materials

Data supporting the findings of this manuscript are available from the corresponding author upon reasonable request.

Financial support and sponsorship

None.

Conflicts of interest

Rozen WM is the Editorial Board Member of the journal *Plastic and Aesthetic Research*. Rozen WM was not involved in any steps of editorial processing, notably including reviewers' selection, manuscript handling and decision making. The other authors declare that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2025.

REFERENCES

1. Yuniarthe Y. Application of artificial intelligence (AI) in search engine optimization (SEO). *IEEE*. 2017;96-101. DOI
2. Wagner G, Lukyanenko R, Paré G. Artificial intelligence and the conduct of literature reviews. *J Inf Technol*. 2022;37:209-26. DOI
3. Kung J. Elicit (product review). *J Can Health Libr Assoc*. 2023;44:15. DOI
4. Schiermeier Q. Pirate research-paper sites play hide-and-seek with publishers. *Nature*. 2015. Available from: <https://www.nature.com/articles/nature.2015.18876>. [Last accessed on 6 Jan 2025].
5. Ma J, Wu X, Huang L. The use of artificial intelligence in literature search and selection of the PubMed database. *Sci Program*. 2015. Available from: <https://onlinelibrary.wiley.com/doi/10.1155/2022/8855307>. [Last accessed on 6 Jan 2025].
6. Cinquini M, Rocco N, Catanuto G, et al. Should acellular dermal matrices be used for implant-based breast reconstruction after mastectomy? *Plast Reconstr Surg Glob Open*. 2023;11:e4821. DOI PubMed PMC
7. Bowers MR, Pulos N, Pulos BP, Shin AY. Opioid-sparing pain management in upper extremity surgery: part 2: surgeon as prescriber. *J Hand Surg Am*. 2019;44:878-82. DOI PubMed
8. Higgins JPT, Thomas J, Chandler J, et al. Cochrane handbook for systematic reviews of interventions. John Wiley & Sons; 2019. DOI
9. Belcher HJ, Nicholl JE. A comparison of trapeziectomy with and without ligament reconstruction and tendon interposition. *J Hand Surg Br*. 2000;25:350-6. DOI PubMed
10. Belcher HJ, Zic R. Adverse effect of porcine collagen interposition after trapeziectomy: a comparative study. *J Hand Surg Br*. 2001;26:159-64. DOI PubMed
11. Brennan A, Blackburn J, Thomson J, Field J. Simple trapeziectomy versus trapeziectomy with flexor carpi radialis suspension: a 17-year follow-up of a randomized blind trial. *J Hand Surg Eur Vol*. 2021;46:120-4. DOI PubMed
12. Corain M, Zampieri N, Mugnai R, Adani R. Interposition arthroplasty versus hematoma and distraction for the treatment of osteoarthritis of the trapeziometacarpal joint. *J Hand Surg Asian Pac Vol*. 2016;21:85-91. DOI PubMed
13. Smet L, Sioen W, Spaepen D, van Ransbeeck H. Treatment of basal joint arthritis of the thumb: trapeziectomy with or without tendon

- interposition/ligament reconstruction. *Hand Surg*. 2004;9:5-9. DOI PubMed
14. Field J, Buchanan D. To suspend or not to suspend: a randomised single blind trial of simple trapeziectomy versus trapeziectomy and flexor carpi radialis suspension. *J Hand Surg Eur Vol*. 2007;32:462-6. DOI PubMed
 15. Gangopadhyay S, McKenna H, Burke FD, Davis TR. Five- to 18-year follow-up for treatment of trapeziometacarpal osteoarthritis: a prospective comparison of excision, tendon interposition, and ligament reconstruction and tendon interposition. *J Hand Surg Am*. 2012;37:411-7. DOI PubMed
 16. Black BE, Griffin PP. The cerebral palsied hip. *Clin Orthop Relat Res*. 1997;338:42-51. DOI PubMed
 17. Hansen TB, Stilling M. Equally good fixation of cemented and uncemented cups in total trapeziometacarpal joint prostheses. A randomized clinical RSA study with 2-year follow-up. *Acta Orthop*. 2013;84:98-105. DOI PubMed PMC
 18. Hart R, Janeček M, Šiška V, Kučera B, Štěpáček V. Interposition suspension arthroplasty according to Epping versus arthrodesis for trapeziometacarpal osteoarthritis. *Eur Surg*. 2006;38:433-8. DOI
 19. Kriegs-Au G, Petje G, Fojtl E, Ganger R, Zachs I. Ligament reconstruction with or without tendon interposition to treat primary thumb carpometacarpal osteoarthritis. Surgical technique. *J Bone Joint Surg Am*. 2005;87 Suppl 1:78-85. DOI PubMed
 20. Marks M, Hensler S, Wehrli M, Scheibler AG, Schindele S, Herren DB. Trapeziectomy with suspension-interposition arthroplasty for thumb carpometacarpal osteoarthritis: a randomized controlled trial comparing the use of allograft versus flexor carpi radialis tendon. *J Hand Surg Am*. 2017;42:978-86. DOI PubMed
 21. Morais B, Botelho T, Marques N, et al. Trapeziectomy with suture-button suspensionplasty versus ligament reconstruction and tendon interposition: a randomized controlled trial. *Hand Surg Rehabil*. 2022;41:59-64. DOI PubMed
 22. Ritchie JF, Belcher HJ. A comparison of trapeziectomy via anterior and posterior approaches. *J Hand Surg Eur Vol*. 2008;33:137-43. DOI PubMed
 23. Salem H, Davis TR. Six year outcome excision of the trapezium for trapeziometacarpal joint osteoarthritis: is it improved by ligament reconstruction and temporary Kirschner wire insertion? *J Hand Surg Eur Vol*. 2012;37:211-9. DOI PubMed
 24. Salibi A, Hilliam R, Burke FD, Heras-Palou C. Prospective clinical trial comparing trapezial denervation with trapeziectomy for the surgical treatment of arthritis at the base of the thumb. *J Surg Res*. 2019;238:144-51. DOI PubMed
 25. Sánchez-Flò R, Fillat-Gomà F, Marcano-Fernández FA, Berenguer-Sánchez A, Balcells-Nolla P, Torner P. Partial versus total trapeziectomy with interposition arthroplasty for trapeziometacarpal osteoarthritis grade II to III Eaton-Littler: a clinical trial. *J Hand Surg Glob Online*. 2020;2:133-7. DOI PubMed PMC
 26. Spekrijse KR, Selles RW, Kedilioglu MA, et al. Trapeziometacarpal arthrodesis or trapeziectomy with ligament reconstruction in primary trapeziometacarpal osteoarthritis: a 5-year follow-up. *J Hand Surg Am*. 2016;41:910-6. DOI PubMed
 27. Spekrijse KR, Vermeulen GM, Kedilioglu MA, et al. The effect of a bone tunnel during ligament reconstruction for trapeziometacarpal osteoarthritis: a 5-year follow-up. *J Hand Surg Am*. 2015;40:2214-22. DOI PubMed
 28. Tägil M, Kopylov P. Swanson versus APL arthroplasty in the treatment of osteoarthritis of the trapeziometacarpal joint: a prospective and randomized study in 26 patients. *J Hand Surg Br*. 2002;27:452-6. DOI PubMed
 29. Thorkildsen RD, Røkkum M. Trapeziectomy with LRTI or joint replacement for CMC1 arthritis, a randomised controlled trial. *J Plast Surg Hand Surg*. 2019;53:361-9. DOI PubMed
 30. Vermeulen GM, Brink SM, Slijper H, et al. Trapeziometacarpal arthrodesis or trapeziectomy with ligament reconstruction in primary trapeziometacarpal osteoarthritis: a randomized controlled trial. *J Bone Joint Surg Am*. 2014;96:726-33. DOI PubMed
 31. Vermeulen GM, Spekrijse KR, Slijper H, Feitz R, Hovius SE, Selles RW. Comparison of arthroplasties with or without bone tunnel creation for thumb basal joint arthritis: a randomized controlled trial. *J Hand Surg Am*. 2014;39:1692-8. DOI PubMed
 32. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Ross R, Lee M. Aesthetic surgery advice and counseling from artificial intelligence: a rhinoplasty consultation with ChatGPT. *Aesthetic Plastic Surgery*. 2023;47:1985-93. DOI PubMed PMC
 33. Seth I, Cox A, Xie Y, et al. Evaluating Chatbot efficacy for answering frequently asked questions in plastic surgery: a ChatGPT case study focused on breast augmentation. *Aesthet Surg J*. 2023;43:1126-35. DOI PubMed
 34. Journal CME Questions. *J Hand Surg*. 2023;48:699. Available from: <https://www.sciencedirect.com/science/article/pii/S0363502323002617>. [Last accessed on 26 Dec 2024].
 35. Seth I, Lim B, Xie Y, Hunter-Smith DJ, Rozen WM. Exploring the role of artificial intelligence chatbot on the management of scaphoid fractures. *J Hand Surg Eur Vol*. 2023;48:814-8. DOI PubMed
 36. Seth I, Lim B, Xie Y, et al. Comparing the efficacy of large language models ChatGPT, BARD, and Bing AI in providing information on rhinoplasty: an observational study. *Aesthet Surg J Open Forum*. 2023;5:ojad084. DOI PubMed PMC
 37. Shridharani SM, Dayan S, Biesman B, et al. Efficacy and safety of tapencarium (RZL-012) in submental fat reduction. *Aesthet Surg J*. 2023;43:NP797-806. DOI PubMed PMC