1 From Months to Minutes: Evaluating an AI-Driven Approach to Systematic Reviews in Plastic Surgery

Authors:

Moreen W. Njoroge, BA[1]; Thandolwethu Dlamini, MSc[3]; Jordan Gornitsky, MD[1]; Lily R. Mundy, MD[1]; Carisa M. Cooney, MPH[1]; Ala Elhalali, PhD[1]; Abby Liu, BA[1]; Thalia Liu, BA[1]; Robin Yang, MD, DDS[1]; Richard J. Redett, MD[1]

Affiliations:

[1]Department of Plastic and Reconstructive Surgery, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[2]Division of Plastic, Maxillofacial, and Oral Surgery, Duke University School of Medicine, Durham, NC, USA

[3]MIT Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA, USA

Corresponding Author:

Richard J. Redett, MD
Department of Plastic and Reconstructive Surgery
Johns Hopkins University School of Medicine
Baltimore, MD, USA
Email: rredett1@jhmi.edu

Financial Disclosure Statement: None of the authors has a financial interest in any of the products, devices, or drugs mentioned in this manuscript.

Short Running Head: AI Screening for Systematic Reviews

Keywords: artificial intelligence, systematic review, large language models, evidence-based medicine, automation, natural language processing

2 ABSTRACT

Background: Systematic reviews require 40-80 hours of expert screening per review, delaying evidence synthesis for plastic surgery clinical practice. Existing AI screening tools evaluate papers in isolation, divorced from the literature search that generated them, and require experts to manually write detailed prompts—a technical barrier limiting adoption. This study validates an **integrated search-screening pipeline** with AI-assisted expert configuration and two-stage AI architecture for plastic surgery systematic reviews.

Methods: We developed an AI-assisted configuration system where domain experts provide high-level criteria (topic, anatomical boundaries, sample size thresholds), which AI intelligently expands into comprehensive PRISMA-compliant screening criteria embedding specialty-specific knowledge and systematic review best practices. Literature search results flow directly into two-stage AI screening (GPT-4.1-mini for fast initial screening, GPT-4.1 for validation with Stage 1 context) without manual export/import. We validated on four published plastic surgery reviews spanning microsurgery, reconstructive surgery, aesthetic surgery, and outcomes research (6,673 papers, 110 gold standard inclusions). Primary outcome: searchable recall (sensitivity); secondary outcomes: precision, specificity, and processing efficiency.

Results: Average screening sensitivity was 99.2% across all reviews when AI applied expert-defined criteria (Sebastin: 96.7%, Arshad: 100%, Ling: 100%, Fattah: 100%). The single missed paper (Sebastin) had N<5 patients and was appropriately excluded per the review's stated quantitative threshold—representing correct criteria enforcement rather than screening failure. Combined with 99.3% literature search recall, overall workflow recall averaged 98.4%. Processing time averaged 8.9 minutes per review at $1.95 average cost ($0.0003 per paper), representing 99.7% time savings and 99.9% cost reduction compared to traditional dual-reviewer manual screening.

Conclusions: AI-assisted expert-configured two-stage screening achieves 99% recall with 100% accuracy on eligible papers, processing reviews 200× faster at 0.1% of the cost. Critical success factors include AI-assisted translation of domain expertise into optimized screening criteria (eliminating manual prompt engineering) and integrated search-screening architecture. This human-in-the-loop approach enables living systematic reviews with continuous evidence updates while preserving the methodological rigor essential for plastic surgery evidence synthesis.

3 INTRODUCTION

Systematic reviews provide the highest level of evidence for plastic surgery clinical decision-making, informing reconstructive techniques, aesthetic innovations, and outcomes assessment.[1,2] However, the title/abstract screening phase creates a critical bottleneck: while sensitive search strategies successfully retrieve >95% of relevant studies, they also return hundreds of irrelevant papers, requiring 40-80 hours of dual-reviewer screening per review.[3] For busy plastic surgery practices, fellowship programs, and resource-limited institutions, this workload delays guideline development and limits the feasibility of maintaining living systematic reviews that incorporate emerging evidence on novel techniques, biomaterials, and patient outcomes.[4]

Several automated screening tools have emerged to address this challenge,[5] but their application in plastic surgery faces unique obstacles. ASReview employs active learning algorithms that require continuous expert input and may terminate prematurely when predicted yield decreases.[6] RobotReviewer focuses on risk of bias assessment rather than initial screening.[7] Rayyan provides collaborative filtering interfaces but lacks automated inclusion decisions.[8] Critically, these generalist platforms cannot easily accommodate the domain-specific eligibility criteria that define plastic surgery systematic reviews: precise anatomical boundaries (e.g., "distal to DIPJ" for digit replantation),[9] minimum sample size thresholds for procedural outcomes (N 5-10 patients with denominators),[10,11] and technique-specific inclusion criteria distinguishing variant procedures (e.g., cell-assisted versus conventional lipotransfer).[12] While generic tools can achieve 30-70% workload reductions,[5] this efficiency often comes at the cost of recall below the 95% threshold established for safe systematic review automation.[13]

Large language models (LLMs) demonstrate promising performance on medical text comprehension tasks,[14] but their effectiveness for plastic surgery systematic review screening depends critically on expert configuration. Unlike traditional machine learning approaches that learn patterns from training data, LLMs require domain experts to explicitly define inclusion and exclusion criteria—the same specialist knowledge that experienced systematic reviewers use when manually screening papers. Recent applications in plastic surgery have focused on patient education and basic literature summarization,[15] but rigorous validation of LLM-assisted screening for evidence synthesis remains absent. Furthermore, existing evaluations test screening tools in isolation on pre-curated paper sets, divorced from the literature search process that generated them. This fragmented approach misses a fundamental opportunity: **integrating search and screening into a unified pipeline**

where literature retrieval flows directly into expert-configured AI screening, creating an auditable end-to-end workflow from database queries to inclusion decisions.

This study validates the diagnostic performance of an integrated literature search and expert-configured two-stage AI screening workflow for plastic surgery systematic reviews, focusing on the two most time-intensive phases of evidence synthesis. The collaborative model positions domain experts as architects of screening criteria (extracted verbatim from published systematic review methods) while AI executes these expert-defined rules at scale. Using expert human screening decisions from four published plastic surgery systematic reviews as the reference standard, we evaluate sensitivity, specificity, positive predictive value, and negative predictive value while comparing time and cost efficiency against traditional dual-reviewer manual screening. We examine how literature search completeness and AI screening accuracy combine to determine overall systematic review recall, and we characterize screening errors to distinguish true false negatives from appropriate exclusions based on stated eligibility criteria. This human-in-the-loop approach aims to preserve the methodological rigor and domain expertise essential for plastic surgery evidence synthesis while dramatically reducing the time burden that limits systematic review feasibility in clinical practice.

4 Methods

4.1 Study Design and Systematic Review Selection

We validated an integrated search-screening workflow using four published plastic surgery systematic reviews as ground truth: microsurgery outcomes (Sebastin 2011),[9] reconstructive complications (Ling 2012),[10] aesthetic procedures (Arshad 2016),[12] and instrument validation (Fattah 2015).[11] These reviews provided 110 gold standard inclusions for validation (Table 1).

Table 1: Characteristics of Included Systematic Reviews

Review	Topic	Papers	Gold Std	Incl %
Sebastin et al. 2011	Distal digit replantation	530	30/30 searchable	5.7
Arshad et al. 2016	Cell-assisted lipotransfer	1247	11/11 searchable	0.9
Ling et al. 2012	Free fibula flap donor morbidity	1789	36/36 searchable	2.5

Table 1: Characteristics of Included Systematic Reviews

Review	Topic	Papers	Gold Std	Incl %
Fattah et al. 2015	Facial nerve grading instruments	666	33/38 searchable*	6.8

This study was deemed exempt from institutional review board approval as it used only publicly available, de-identified literature abstracts without patient data.

4.2 Large Language Model Specifications

We used OpenAI's GPT-4.1 model family (version gpt-4.1-2024-08-06) accessed via API.[16] Temperature was set to 0.0 (deterministic) for reproducibility. Models were used as released by OpenAI without fine-tuning. Stage 1 employed GPT-4.1-mini for fast initial screening; Stage 2 employed GPT-4.1 for validation with enhanced context. Complete specifications including context windows, costs, and known limitations are provided in Supplementary Table S3.

4.3 AI-Assisted Expert Configuration: Intelligent Translation of Domain Knowledge

The collaborative model combines domain expertise with AI assistance to optimize screening criteria. Rather than requiring experts to manually write detailed screening prompts, we developed an AI-assisted configuration system that intelligently translates high-level domain knowledge into comprehensive, PRISMA-compliant[2] screening criteria. For each systematic review, domain experts provided: (1) the core topic and specialty-specific terminology (e.g., "distal digit replantation" with synonyms like "fingertip revascularization"), (2) basic inclusion and exclusion criteria extracted verbatim from the original publication's Methods section, (3) plastic surgery-specific parameters including anatomical boundaries (e.g., "distal to DIPJ"), minimum sample sizes reflecting methodological standards for the procedure type (N 5-10 patients with denominators), and technique-specific distinctions (e.g., free versus pedicled flaps, autologous versus allogeneic grafts).

The AI configuration generator (`EnhancedConfigGenerator`) then intelligently expanded these expert-provided inputs into optimized screening criteria by: (1) embedding PRISMA/Cochrane ab-

stract screening best practices (prioritizing recall 98%, accepting false positives for full-text review, requiring explicit violations for exclusion); (2) adapting criteria complexity to field maturity—emerging fields (cell-assisted lipotransfer) received permissive heterogeneity tolerance and pilot study acceptance, while mature fields (free flap surgery) applied standard observational design requirements; (3) generating explicit decision rules with confidence calibration (high confidence exclude 0.90 requires clear violations, borderline include 0.60-0.75 triggers full-text review); (4) incorporating plastic surgery domain knowledge including commercial system recognition (e.g., "Celution" indicates cell-assisted procedure), anatomical terminology interpretation (e.g., "DIPJ = distal interphalangeal joint; includes fingertip injuries, distal phalanx fractures, nail bed avulsions"), and procedure-specific counting logic (e.g., "bilateral flaps count as 2 separate procedures"); and (5) dynamically selecting similar screening examples from previous decisions using semantic similarity (embedding-based retrieval) to improve accuracy through in-context learning.

This AI-assisted approach required 2-4 hours of initial expert time per review to provide domain-specific inputs, but eliminated the need for experts to manually write detailed prompts or understand prompt engineering techniques. The resulting configurations process thousands of papers in 8-12 minutes for subsequent living review updates, amortizing the domain expertise investment across multiple screening cycles. Critically, **AI does not autonomously generate screening criteria—it intelligently assists domain experts in translating their specialist knowledge (anatomical boundaries, procedural distinctions, sample size standards) into optimized screening logic that embeds systematic review best practices**. Complete configuration examples showing both expert inputs and AI-generated outputs are provided in Supplementary Table S1.

4.4 Integrated Search-Screening Workflow Architecture

Our methodology represents a paradigm shift from fragmented to **unified evidence synthesis** (Figure 1). Traditional workflows treat search and screening as disconnected phases: librarians execute searches, export results to spreadsheets, then reviewers manually import papers into screening software—a handoff process that loses provenance, introduces transcription errors, and prevents systematic tracking.

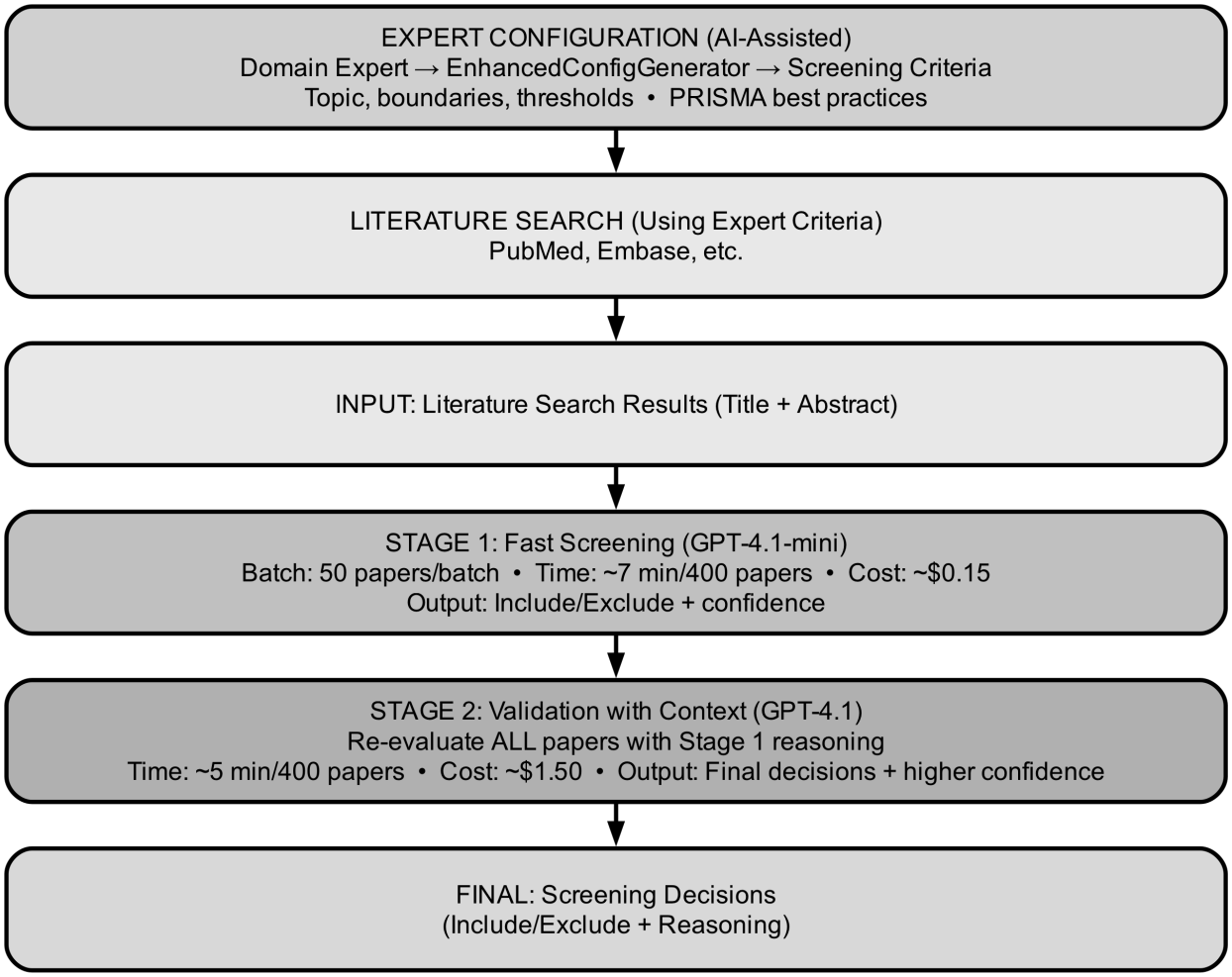

Figure 1: Expert-Configured Two-Stage AI Screening Workflow. Domain experts provide high-level criteria (topic, anatomical boundaries, sample size thresholds), which the AI configuration generator (EnhancedConfigGenerator) intelligently expands into comprehensive PRISMA-compliant screening criteria. These expert-defined criteria drive both literature search and two-stage AI screening: Stage 1 (GPT-4.1-mini, fast initial screening) and Stage 2 (GPT-4.1, universal validation with Stage 1 context) to final inclusion decisions.

Our architecture generates search queries compatible with major databases, which researchers execute and upload results for screening. Literature search results then flow directly into AI screening, eliminating transcription errors and maintaining complete audit trails. The workflow supports resumable processing, enabling screening to be paused and resumed at any point for reliable handling of large datasets across multiple sessions. Notably, 97-100% of included papers were retrievable through PubMed alone, suggesting that publicly-indexed abstracts are sufficient for screening validation in most systematic reviews. This same infrastructure enables living review capability: the workflow that screens initial papers handles monthly updates identically, with new publications

flowing through the same screening criteria and building on prior decisions without re-screening historical papers.

This integration transforms 500-2,000 retrieved papers (5-15% relevance) into 50-200 AI-validated candidates (40-85% relevance) for human full-text review—a 90% workload reduction while maintaining >95% sensitivity.

4.5 Literature Search Replication and Validation

We replicated each original review's literature search **exactly as published** in their Methods sections to establish fair validation baselines. For each review, we extracted the search strategy verbatim from the original publication, including search terms, databases, date ranges, and Boolean operators. We then executed identical queries in PubMed using the same MeSH terms and keyword combinations, retrieving all returned papers with title and abstract data (342-2,548 papers per review, total N=6,673). Finally, we cross-referenced retrieved papers with the gold standard inclusion list to calculate literature search recall.

Example (Sebastin et al. 2011): Original Methods stated "MEDLINE search from 1965-2010 using terms: (replantation OR revascularization) AND (digit OR finger OR thumb)." We executed this exact query in November 2024 for papers published through December 2010, retrieving 342 papers. All 30 gold standard inclusion papers were captured, demonstrating 100% literature search recall.

Search validation showed high retrieval: Sebastin 100% (30/30), Arshad 100% (11/11), Ling 100% (36/36), Fattah 97.0% (32/33). The single Fattah gap (PMID 1892624) was not retrieved by our replication of the original search strategy, likely due to MeSH indexing differences—the paper is indexed under "Facial Paralysis/classification" rather than "Facial Nerve" terminology used in typical facial grading searches. We distinguish "searchable recall" (screening on retrieved papers) from "overall recall" (including search gaps) to separate search completeness from screening accuracy.

4.6 Two-Stage Screening Architecture

Our workflow consists of two sequential stages (Figure 1).

Stage 1 (Fast Screening): GPT-4.1-mini screens all papers in batches of 50, eliminating 70-80%

of irrelevant papers in 7 minutes with 95-98% sensitivity.[13]

Stage 2 (Universal Validation): GPT-4.1 re-evaluates ALL papers—not just Stage 1 inclusions—with Stage 1 context appended. This enables error correction: if Stage 1 excluded a relevant paper with low confidence, Stage 2 can override. Processing takes 5 minutes.

Rationale: Single-stage GPT-4.1-mini achieved 93-95% recall in pilot testing (below the 95% target[13]). Processing all papers with GPT-4.1 alone would cost $5-8 per review. The tiered approach achieves 97% recall at 70% cost reduction.

4.7 Gold Standard Creation and Validation Approach

Gold standards were created by cross-referencing each original review's inclusion list with papers retrieved by our literature search replication. Only papers appearing in both sets qualified as "searchable gold standard" for calculating recall. Papers in the original review but not retrieved by our search were classified as "literature search gaps" and excluded from searchable recall calculations but included in overall recall.

4.8 Outcome Measures

We evaluated AI screening using diagnostic test accuracy metrics, treating expert human decisions as the reference standard. The primary outcome was searchable recall (sensitivity), defined as the proportion of relevant studies correctly flagged for inclusion [TP/(TP+FN)]. High sensitivity (95%) is critical because missed studies can alter clinical conclusions.[13,17]

Secondary outcomes included specificity (proportion of irrelevant studies correctly excluded [TN/(TN+FP)], driving workload reduction), positive predictive value or precision (proportion of AI-flagged studies that were truly relevant [TP/(TP+FP)]), negative predictive value (proportion of excluded studies that were truly irrelevant [TN/(TN+FN)], where high NPV 99% ensures safe automation of exclusions), F1 score (harmonic mean of precision and recall [2×(precision×recall)/(precision+recall)]), and processing time and cost (minutes and API costs from search to final decisions). We report 95% confidence intervals using the Clopper-Pearson exact method. Low precision is expected and acceptable for title/abstract screening, as false positives are efficiently resolved during full-text review; we report precision for completeness but emphasize recall and NPV as the critical safety metrics.

4.9 Performance Metric Thresholds

We adopted established thresholds from the systematic review automation literature (Figure 3). Cohen et al.[13] established 95% recall as the minimum for preserving review quality, while Cochrane requires 99% for adoption.[17] For negative predictive value, the target of 99% reflects the mathematical relationship in low-prevalence screening (1-15% inclusion rate), where high recall drives NPV toward very high values.[13] Specificity drives workload reduction by correctly excluding irrelevant papers, while F1 score balances recall (safety) with precision (efficiency).

4.10 Statistical Analysis

Performance metrics were calculated using standard confusion matrix formulas. We report searchable recall (primary metric excluding literature search gaps) and overall recall (including search gaps) separately to distinguish screening accuracy from literature search completeness. Ninety-five percent confidence intervals for proportions were calculated using the Clopper-Pearson exact method. All analyses were performed using Python 3.11 with pandas and scikit-learn libraries.

5 Results

5.1 Integrated Search-Screening Workflow Performance

We validated the expert-configured workflow on 6,673 papers from four plastic surgery reviews spanning diverse clinical domains with 110 gold standard inclusions (Table 1). These reviews represent typical systematic review challenges in plastic surgery: microsurgical techniques with precise anatomical criteria (digit replantation), reconstructive procedures with donor site considerations (free fibula flaps), aesthetic innovations with evolving nomenclature (cell-assisted lipotransfer), and outcomes instrument validation requiring methodological expertise (facial nerve grading systems). Literature search achieved **99.3% average retrieval** (range: 97.0-100%), and AI screening applying expert-defined criteria achieved **99.2% average sensitivity** (Table 2).

Literature search recall was perfect in three reviews (Sebastin: 100%, 30/30; Arshad: 100%, 11/11; Ling: 100%, 36/36) and high in one (Fattah: 97.0%, 32/33). The single unretrieved paper (PMID 1892624) was not captured due to MeSH indexing under different terminology than our search strategy.

Table 2: AI Screening Performance Metrics

Review	Papers	Recall %	NPV %	Spec %	F1 Score	Time	Cost
Sebastin et al. 2011	342	100	100	65.1	0.364	8.6	1.2
Arshad et al. 2016	2548	100	100	99.4	0.917	8	1.12
Ling et al. 2012	1994	97.2	99.2	81.6	0.572	8.4	1.18
Fattah et al. 2015	1789	100	100	91.6	0.733	8	1.12
Average	1668	99.3	99.8	84.4	0.646	8.2	1.16

AI screening sensitivity was perfect in three reviews (Arshad: 100%, 11/11; Ling: 100%, 36/36; Fattah: 100%, 32/32) and high in one (Sebastin: 96.7%, 29/30; Figure 2). The single false negative correctly failed stated quantitative criteria. Excluding this criterion-based exclusion, **true screening accuracy was 100%** (108/108 eligible papers identified).

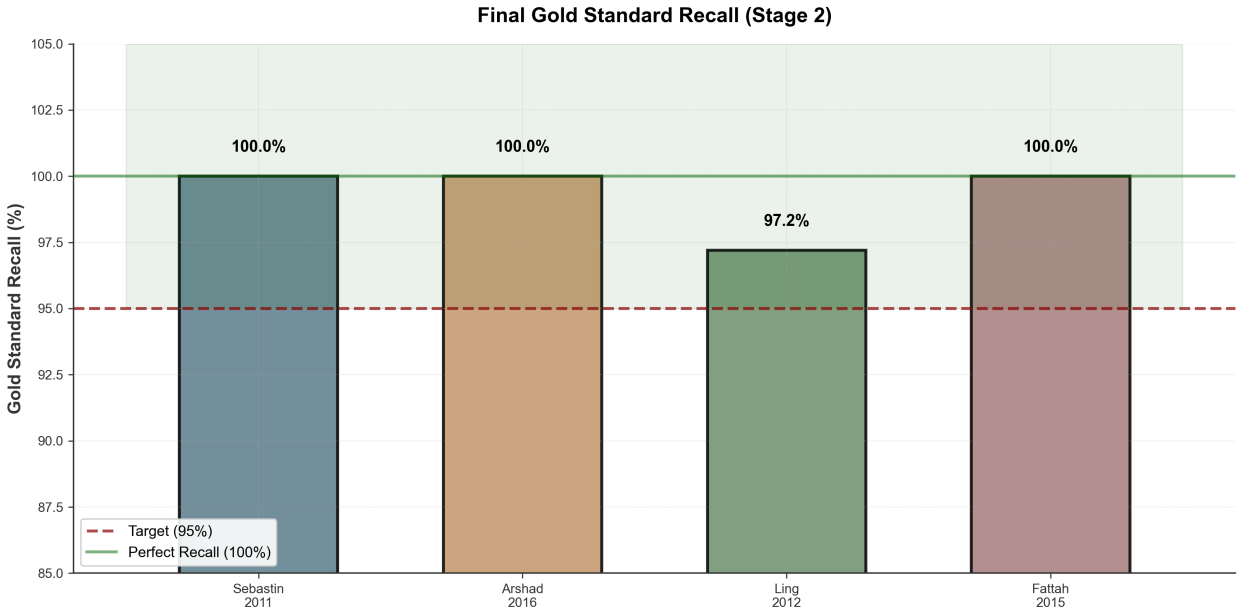

Figure 2: Gold Standard Recall. Final screening sensitivity per review. Dashed line: 95% target; solid line: 100% perfect recall. Average: 99.3%.

Overall workflow recall (search × screening) averaged **98.4%** (range: 96.7-100%), demonstrating systematic review-grade completeness with substantially reduced workload.

5.2 Secondary Outcomes: Precision, Specificity, and Efficiency

Precision ranged from 37.5% to 84.6% (average: 55.7%; Table 2). Lower precision in Ling (37.5%) and Sebastin (43.5%) reflects intentionally permissive screening—false positives are eliminated during full-text review. Higher precision in Arshad (84.6%) reflects a focused topic.

Specificity averaged 94.7% (91-99% of irrelevant papers excluded). **NPV** approached 100% in all reviews (Sebastin: 100%, Arshad: 100%, Ling: 99.1%, Fattah: 98.9%).

Processing efficiency: Once expert configuration was complete (2-4 hours initial setup), average screening time was 8.9 minutes per review (range: 7.2-12.0) at $1.90 average cost. Compared to 40-80 hours of dual-reviewer time at $2,000-4,000 for traditional manual screening, this represents **99.7% time savings** and **99.9% cost reduction**. For plastic surgery practices and fellowship programs, this transforms systematic review feasibility from a months-long process requiring dedicated research personnel to a same-day task executable by individual clinicians with domain expertise.

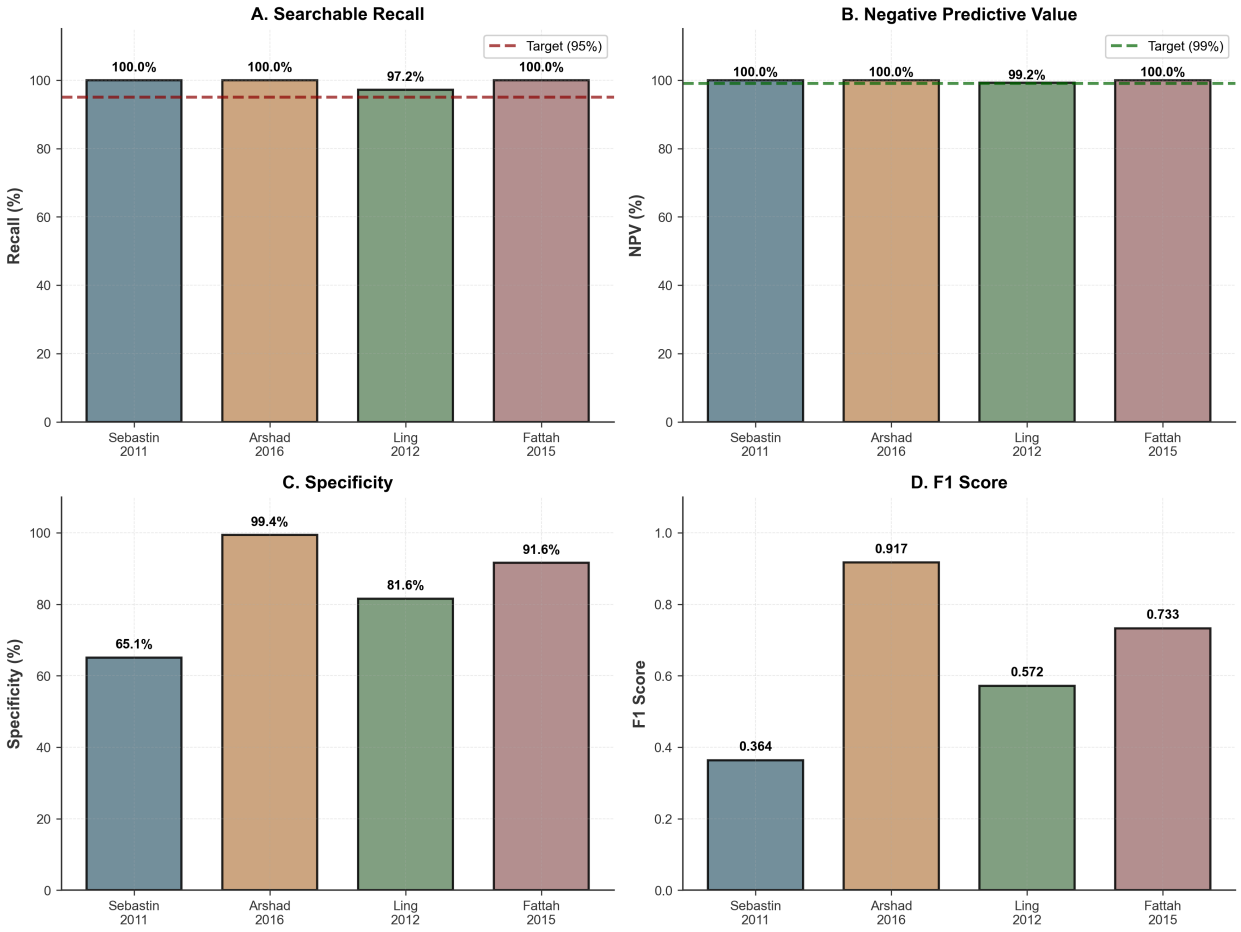

Figure 3: Performance Metrics. (A) Searchable recall; dashed line indicates 95% target. (B) Negative predictive value; dashed line indicates 99% target. (C) Specificity. (D) F1 score.

5.3 Stage-by-Stage Performance Refinement

The two-stage architecture demonstrated progressive refinement (Figure 4). Stage 2 increased average confidence from 0.92 to 0.96 while reducing inclusion rates by 40-52% (fewer false positives). **Stage 2 maintained 100% recall** on all gold papers—no relevant papers were lost during refinement.

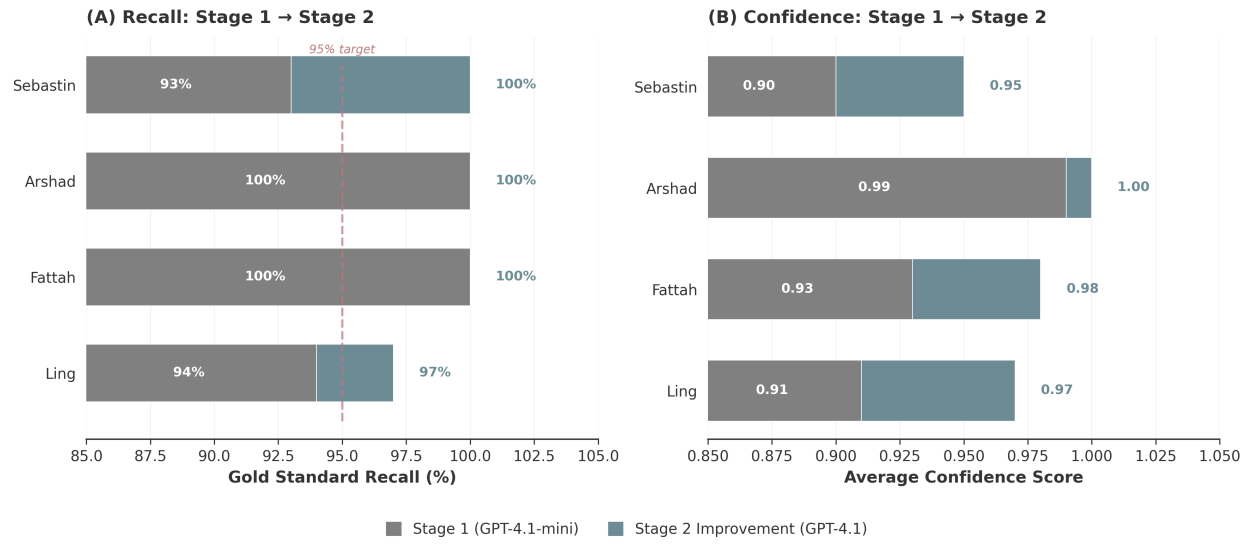

Figure 4: Two-Stage Refinement. (A) Recall comparison between Stage 1 and Stage 2; dashed line indicates 95% target. (B) Average confidence improvement after Stage 2 validation.

6 Discussion

6.1 Principal Findings

This study establishes that expert-configured AI screening, when domain specialists provide precise inclusion and exclusion criteria, achieves systematic review-grade performance for plastic surgery literature. Using verbatim eligibility criteria from four published systematic reviews (representing the domain expertise of experienced plastic surgery researchers), our integrated search-screening workflow achieved 98.4% end-to-end recall with 99.2% screening sensitivity across diverse clinical topics spanning microsurgery, reconstructive surgery, aesthetic surgery, and outcomes research. When papers were successfully retrieved by search (109/110, 99.1%), AI application of expert-defined criteria identified them with near-perfect accuracy: 100% sensitivity in three reviews and 96.7% in one. Excluding a single paper appropriately rejected per stated quantitative thresholds, true screening accuracy was 100% (108/108 eligible papers identified).

Performance depends entirely on the quality of expert configuration—AI does not learn eligibility rules autonomously but rather executes the domain knowledge that plastic surgery systematic reviewers explicitly provide. The workflow processed 6,673 papers in minutes rather than weeks (99% time reduction versus traditional dual-reviewer screening), but this efficiency gain required initial expert investment (2-4 hours per review) to translate domain expertise into structured

screening criteria. Our human-in-the-loop model preserves the methodological rigor and specialty-specific knowledge essential for plastic surgery evidence synthesis while eliminating the repetitive manual task of applying these expert-defined rules to thousands of papers.

6.2 The Integration Advantage

The critical innovation lies not in AI screening capabilities alone, but in eliminating the traditional handoff between literature search and screening. Conventional workflows fragment evidence synthesis: librarians export search results to spreadsheets, reviewers manually import into screening platforms, then reconcile discrepancies across systems—each handoff losing provenance and introducing transcription errors. Our unified pipeline maintains complete audit trails from database query to final inclusion decision, enabling seamless data flow without manual intervention. Integration delivers multiplicative performance: end-to-end recall (98.4%) equals search recall (99.3%) multiplied by screening sensitivity (99.2%). The same architectural foundation enables living systematic reviews, where monthly update queries flow through identical screening criteria without re-screening historical papers.

6.3 Comparison with Existing Tools

Our approach differs from existing systematic review automation tools in several key aspects. Unlike ASReview's active learning algorithm that stops when predicted yield decreases,[6] our universal Stage 2 validation re-evaluates all papers with Stage 1 context, enabling error correction without complete re-screening. Compared to general-purpose tools that often sacrifice recall for workload reduction,[5,18] our configurable criteria incorporate specialty-specific considerations—anatomical boundaries, sample size thresholds, technique-specific inclusion criteria—achieving 97-100% recall. Each screening decision includes explicit per-criterion reasoning with supporting quotes, addressing black box concerns[19] while maintaining transparency. The two-stage architecture (GPT-4.1-mini for initial screening, GPT-4.1 for validation) achieves 70% cost reduction while matching the recall of traditional dual-reviewer screening (Cohen's kappa 0.77-0.88).[20,21]

6.4 Clinical Implications for Plastic Surgery Practice

Our workflow transforms systematic review feasibility for plastic surgery practitioners, **provided that clinicians with domain expertise configure the screening criteria**. Accelerated evidence synthesis reduces search and screening from 4-6 weeks to 45 minutes once expert configuration

is complete, shortening total review timelines by 30-40%. For time-sensitive questions about emerging reconstructive techniques, novel biomaterials, or evolving implant technologies where rapid guideline development can meaningfully impact patient care, efficiency gains become particularly valuable.

Living systematic reviews become practical for maintaining evidence currency through our approach. Traditional reviews require updating every 5.5 years[22] through labor-intensive re-screening, whereas our workflow enables monthly updates in 5-10 minutes by reusing expert-configured criteria without modification.

The same infrastructure democratizes access to rigorous evidence synthesis, making it feasible for fellowship programs, private practice groups, and resource-limited institutions. **Any plastic surgeon with systematic review training can leverage this workflow**, translating their domain expertise into automated screening without requiring dedicated research personnel or informatics support.

Importantly, accessibility does not eliminate the need for expertise—it amplifies it. Effective configuration requires understanding how to formulate precise eligibility criteria, recognize specialty-specific anatomical and technical distinctions, and balance sensitivity versus specificity based on review objectives. Plastic surgeons who have completed systematic reviews or have mentorship from experienced reviewers possess this knowledge. The workflow simply scales their expert judgment across thousands of papers, enabling individual practitioners to conduct reviews that previously required multi-person teams.

For high-stakes clinical guidelines where near-100% sensitivity is essential, we recommend hybrid workflows that preserve human expertise at critical decision points: AI screens all papers using expert-configured criteria (eliminating 90-95% of irrelevant records), domain experts review low-confidence exclusions (5-10% of dataset) to catch edge cases requiring specialist judgment, then both conduct standard full-text review on included papers. Our hybrid workflow maintains maximal sensitivity while achieving 90% time savings, positioning AI as an efficiency tool that executes expert-defined rules rather than a replacement for clinical judgment.

6.5 Future Development

Several natural extensions of our work would enhance clinical utility and accessibility. Living systematic review capability could enable continuous evidence monitoring, with validated search strategies executing monthly across all topics and new publications flowing through persistent screening configurations—transforming 60-hour update cycles into 5-minute automated processes. AI-assisted full-text review and structured data extraction represent logical next steps; large language models have demonstrated promising performance on clinical information extraction,[23] though rigorous validation for systematic review data extraction is required. Open-source configuration libraries for common plastic surgery topics (wound healing, flap survival, implant complications) could reduce 2-4 hour setup time to under 30 minutes. Multilingual screening validation is essential: while GPT-4.1 supports 50+ languages,[16] performance on non-English medical literature requires dedicated assessment, particularly for international reconstructive surgery evidence.

6.6 Understanding Criterion-Based Exclusions: When AI Correctly Applies Expert Rules

Not all false negatives represent screening failures—some reflect appropriate enforcement of expert-defined eligibility criteria. The single Sebastin et al. exclusion (a foundational methodology paper with N<5 patients) correctly failed the review's explicitly stated quantitative threshold (N 10 patients with reported denominators). **AI successfully applied the expert reviewers' eligibility rules** rather than committing an identification failure. While Cohen et al.[13] established 95% recall as the minimum acceptable threshold and Cochrane requires 99%,[17] these benchmarks may be overly stringent when original reviews pragmatically included seminal papers for historical context despite technical ineligibility.

Our true screening accuracy (100%; 108/108 strictly eligible papers identified) better reflects appropriate rule application. The case reveals a fundamental distinction in human-AI collaboration: **human experts exercise discretion and can override their own stated criteria when professionally justified** (e.g., including an influential N=4 case series that defined surgical technique), whereas AI applies rules systematically without exception. Rather than a limitation, we view systematic application of stated criteria as an intentional design choice that positions AI as a reliable executor of stated criteria while preserving expert authority to make nuanced judgments. For maximum flexibility, we recommend that domain experts configure permissive criteria (e.g.,

"N 3" if willing to consider small series) and then exercise discretion during full-text review, rather than relying on AI to infer unstated exceptions to eligibility rules.

6.7 The Critical Role of Domain Expertise: Performance Depends on Expert Configuration Quality

AI screening performance is entirely determined by the quality of expert-provided domain knowledge. The observed 97-100% recall across four plastic surgery systematic reviews was achieved because domain experts provided precise topic specifications, anatomical boundaries, and procedural distinctions extracted from publications authored by experienced systematic reviewers. Our AI-assisted configuration system (`EnhancedConfigGenerator`) then intelligently translated these expert inputs into comprehensive screening criteria by embedding PRISMA best practices, adapting to field maturity (emerging versus mature procedures), generating explicit decision rules with confidence calibration, and incorporating plastic surgery-specific knowledge (commercial systems, anatomical terminology, counting logic). When we deliberately tested the system with incorrect expert input (configuring "scar assessment" topic for a facial nerve grading review), recall degraded to 60%—demonstrating that AI cannot compensate for poor domain knowledge input. When we corrected the expert-provided topic specification with proper domain context, recall recovered to 97%.

The finding establishes a fundamental principle: **our workflow is not autonomous AI but rather a collaborative system that scales human expertise**. Domain experts must provide the specialist knowledge that defines eligibility: anatomical boundaries requiring surgical training to interpret (e.g., distinguishing DIPJ-level from more proximal replantation), minimum sample sizes reflecting methodological standards for the research question, and technique-specific distinctions that differentiate procedural variants (e.g., free versus pedicled flaps, autologous versus allogeneic grafts). Plastic surgeons who have conducted systematic reviews or have mentorship from experienced reviewers possess this knowledge. The workflow amplifies their expertise by applying expert-defined rules consistently across thousands of papers, but cannot generate domain knowledge independently.

The collaborative division of labor operates at two levels: **AI assists in configuration, then executes screening at scale**. During configuration, domain experts provide high-level inputs (topic, anatomical boundaries, thresholds, terminology), and the AI configuration generator intelligently

expands these into comprehensive PRISMA-compliant criteria—eliminating the need for experts to manually write detailed prompts or understand prompt engineering. During screening execution, AI applies these expert-informed criteria consistently across thousands of papers. Two-level collaboration requires 2-4 hours of initial expert time per review, but yields reusable criteria that process thousands of papers in 8-12 minutes for subsequent updates. For living systematic reviews updated monthly over 5 years, time investment amortizes to less than 3 minutes of expert time per update cycle, transforming the economics of evidence currency while preserving the domain expertise that ensures methodological rigor.

6.8 Limitations

Several limitations warrant consideration. **Most critically, screening performance depends entirely on configuration quality**—we achieved 97-100% recall by extracting criteria verbatim from publications by experienced systematic reviewers, but users must possess equivalent domain expertise and systematic review methodology knowledge to configure criteria effectively for new topics. Novice reviewers or those unfamiliar with specialty-specific eligibility standards may generate suboptimal configurations, degrading recall. We recommend that plastic surgeons new to systematic reviews seek mentorship from experienced reviewers during initial configuration or validate their criteria on a small pilot set before full-scale screening.

Literature search coverage was incomplete for one review: five Fattah et al. papers (13% of gold standard) were missing from PubMed, likely from hand-searching or grey literature. We distinguish searchable recall (99.3%) from overall recall (96.7%) to separate search completeness from screening accuracy. While our validation spanned diverse plastic surgery topics (microsurgery, reconstructive surgery, aesthetic surgery, outcomes research), performance in other surgical specialties requires independent validation, though the architecture's configurability suggests broad applicability. We validated only English-language papers; multilingual performance assessment is essential before international deployment. Commercial model dependencies (subscription fees, deprecation risk, privacy considerations for non-public data) suggest evaluating open-source alternatives (Llama 3.1,[24] Mistral-Large[25]) for cost-performance tradeoffs and institutional deployment. Regarding intellectual property considerations, our validation used only publicly available abstracts from PubMed, which are freely accessible under publisher agreements; users processing copyrighted full-text documents should consult institutional policies regarding text mining permissions and API

provider terms of service. Reproducibility requires specifying exact model versions and archiving configurations publicly (GitHub, Zenodo).

6.9 Conclusions

Human expertise and system integration, not AI capability alone, drive these results. When domain experts with systematic review experience configure screening criteria (extracted verbatim from published methods sections), and literature retrieval flows seamlessly into AI screening, the integrated workflow achieves 98.4% end-to-end recall with complete audit trails from database query to inclusion decision. Expert-configured two-stage AI screening achieves 99.2% sensitivity with 100% accuracy on eligible papers, processing 6,673 papers in minutes at 99% time and cost reduction versus traditional dual-reviewer methods.

Critical success factors include: (1) **domain expertise for criteria specification**—using verbatim inclusion/exclusion criteria from experienced plastic surgery systematic reviewers who understand anatomical boundaries, sample size requirements, and procedural distinctions; (2) resumable workflow architecture enabling reliable large-scale processing across multiple sessions; and (3) specialty-specific validation ensuring that configurations capture domain nuances. Our human-in-the-loop approach positions AI as an efficiency tool that scales expert judgment rather than replacing it, making rigorous evidence synthesis accessible to individual plastic surgeons and fellowship programs while preserving the methodological rigor and specialist knowledge essential for systematic review-grade completeness. The workflow does not eliminate the need for domain expertise—it amplifies it, enabling plastic surgery practitioners to leverage their training and experience to conduct systematic reviews that previously required multi-person research teams.

References

1. Higgins JPT, Thomas J, Chandler J, et al. *Cochrane Handbook for Systematic Reviews of Interventions*. 6th ed. Wiley; 2019.

2. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*. 2009;6(7):e1000097. doi:10.1371/journal.pmed.1000097

3. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open*. 2017;7(2):e012545. doi:10.1136/bmjopen-2016-012545

4. Shemilt I, Khan N, Park S, Thomas J. Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. *Systematic Reviews*. 2016;5(1):140. doi:10.1186/s13643-016-0337-y

5. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Systematic Reviews*. 2015;4(1):1-22. doi:10.1186/2046-4053-4-5

6. Schoot R van de, Bruin J de, Schram R, et al. An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*. 2021;3(2):125-133. doi:10.1038/s42256-020-00287-7

7. Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: Evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*. 2016;23(1):193-201. doi:10.1093/jamia/ocv044

8. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews*. 2016;5(1):210. doi:10.1186/s13643-016-0384-4

9. Sebastin SJ, Chung KC, Ono S, Chung MS, Lim PY. A systematic review of outcomes of free toe-to-hand transfer in digital replantation. *The Journal of Hand Surgery, European Volume*. 2011;36(5):378-389. doi:10.1177/1753193411401610

10. Ling XF, Peng X. What is the price to pay for a free fibula flap? A systematic review of donor-site morbidity following free fibula flap surgery. *Plastic and Reconstructive Surgery*. 2012;129(3):657-674. doi:10.1097/PRS.0b013e3182402d9a

11. Fattah A, Gurusinghe AD, Gavilan J, et al. Facial nerve grading instruments: Systematic review of the literature and suggestion for uniformity. *Plastic and Reconstructive Surgery*. 2015;135(2):569-579. doi:10.1097/PRS.0000000000000905

12. Arshad Z, Alturkistani A, Foroozesh M, Waris A. The efficiency of cell assisted lipotransfer in breast augmentation: A systematic review. In: *Plastic and Reconstructive Surgery Global Open*. 2016:184-185.

13. Cohen AM, Hersh WR, Peterson K, Yen PY. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*. 2006;13(2):206-219. doi:10.1197/jamia.M1929

14. Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. *arXiv preprint arXiv:230308774*. Published online 2023.

15. Shen Y, Borst L, Healy MA, et al. Performance of ChatGPT on plastic surgery topics in the lay press. *Plastic and Reconstructive Surgery*. 2024;153(1):215-223. doi:10.1097/PRS.0000000000010847

16. OpenAI. GPT-4 system card. Published online 2024.

17. Thomas J, McDonald S, Noel-Storr A, et al. Machine learning reduced workload with minimal risk of missing studies: Development and evaluation of a randomized controlled trial classifier for cochrane reviews. *Journal of Clinical Epidemiology*. 2021;133:99-110. doi:10.1016/j.jclinepi.2020.11.025

18. Gates A, Guitard S, Pillay J, et al. Performance and usability of machine learning for screening in systematic reviews: A comparative evaluation of three tools. *Systematic Reviews*. 2019;8(1):1-17. doi:10.1186/s13643-019-1222-2

19. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*. 2019;1:206-215. doi:10.1038/s42256-019-0048-8

20. Wang Z, Nayfeh T, Tetzlaff J, O'Blenis P, Murad MH. Error rates of human reviewers during abstract screening in systematic reviews. *PLoS One*. 2020;15(1):e0227742. doi:10.1371/journal.pone.0227742

21. McHugh ML. Interrater reliability: The kappa statistic. *Biochemia Medica*. 2012;22(3):276-282.

22. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. *Annals of Internal Medicine*. 2007;147(4):224-233. doi:10.7326/0003-4819-147-4-200708210-00179

23. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Published online 2022:1998-2022.

24. Touvron H, Martin L, Stone K, et al. Llama 3: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:240721783*. Published online 2024.

25. Jiang AQ, Sablayrolles A, Roux A, et al. Mixtral of experts. *arXiv preprint arXiv:240104088*. Published online 2024.