Surgical Overview
SURGICAL TECHNOLOGY INTERNATIONAL Volume 46
doi: 10.36922/sti.0457

Artificial Intelligence Documentation Tools in Surgery: A Systematic Review

Damien Gibson[1,2,3]*, Victor Yu[1,3], Kate Alexander[1], Kun Yu[4], Scott Leslie[1,2,3], Ruban Thanigasalam[1,5], Nicola Jeffery[1,3], Daniel Steffens[1,3,6]

[1]Surgical Outcomes Research Centre, Royal Prince Alfred Hospital, Sydney, New South Wales, Australia
[2]Faculty of Medicine and Health, Central Clinical School, The University of Sydney, Sydney, New South Wales, Australia
[3]Department of Urology, Royal Prince Alfred Hospital, Sydney, New South Wales, Australia
[4]Data Science Institute, University of Technology Sydney, Sydney, New South Wales, Australia
[5]Department of Urology, Chris O'Brien Lifehouse, Sydney, New South Wales, Australia
[6]NHMRC Clinical Trials Centre, The University of Sydney, Sydney, New South Wales, Australia

***Corresponding author:**
Damien Gibson
(Damien.gibson@health.nsw.gov.au)

Abstract

Background: Clinical documentation burden is a major contributor to burnout in surgery. Artificial intelligence (AI) tools, such as automatic speech recognition (ASR) and large language models (LLMs), may streamline documentation without sacrificing quality.

Objective: We systematically reviewed the performance of ASR- and LLM-based documentation tools in surgical settings.

Methods: Following the Preferred Reporting Items for Systematic Reviews and Meta-analyses, MEDLINE, Embase, CENTRAL, and Scopus (January 2015–October 2025) were searched for studies evaluating AI-enabled documentation (e.g., ambient scribes, advanced ASR, LLM-assisted drafting) in surgical care. Dual reviewers screened, extracted, and assessed risk of bias using the Risk of Bias in Non-randomized Studies of Exposures tool. Heterogeneity of included studies precluded meta-analysis, and results are presented narratively.

Results: Seven studies published between 2023 and 2025 across otolaryngology, neurosurgery, plastic surgery, and urology were included. Tools such as LLM-assisted operative reports, ambient clinic scribes, and ASR dictation were employed. Findings revealed that AI scribes improved documentation efficiency (5.16 min vs. 10.58 min) and reduced documentation time (5–50 s vs. 7.1–7.4 min), with hybrid clinician-in-the-loop workflows achieving the best balance of speed and quality. AI scribe notes were non-inferi-

doi: 10.36922/sti.0457

or to clinician notes on the Physician Documentation Quality Instrument-9 (33.6/45). Operative note quality was highest with hybrid attending-reviewed generative pre-trained transformer drafts (79% as-is approval) and lowest with generative pre-trained transformer-only notes (23%). Whisper ASR was non-inferior to Dragon Medical One for word error rate and superior when linguistic errors were excluded. .

Conclusion: Early evidence suggests clinician-supervised AI documentation may accelerate note generation while maintaining comparable quality, with hybrid use outperforming AI-only approaches. However, the evidence base is early, heterogeneous, and largely non-randomized, and downstream outcomes—including burnout—remain unmeasured. Real-world trials incorporating patient, workflow, safety, and governance outcomes are needed to guide supervised implementation.

***Keywords*:** Artificial intelligence, Surgical documentation, Ambient digital scribes, Large language models, Operative reports, Clinical workflow, Surgeon burnout

INTRODUCTION

Clinical and administrative documentation is a leading contributor to clinician burden and burnout in surgical practice.[1,2] Surgical workflows span multiple environments (e.g., clinic, ward, and theatre), including high-stakes, time-critical records (e.g., operative reports). Errors or omissions can have immediate downstream consequences for perioperative decision-making and patient safety. This requires rapid, accurate documentation of complex encounters, often extending work after hours and increasing cognitive load.[2] These tasks detract from direct patient communication, with each hour of face-to-face care generating approximately two hours of administrative work.[3,4] This creates substantial evening and weekend "pyjama time" linked to emotional exhaustion, discontentment, burnout, and an increased risk of diagnostic error.[5,6]

While clinical and administrative documentation requires clinical judgment, much of its execution is inherently scribe-like and thus amenable to automation.[7] Emerging artificial intelligence (AI) tools, particularly ambient speech-to-text systems that draft clinical documentation during the encounter, demonstrate promise by offloading repetitive tasks and reducing surgeon workload.[8,9]

Contemporary AI documentation tools combine automatic speech recognition (ASR) with large language models (LLMs) to convert clinician–patient dialogue into structured documentation.[1] ASR backbones such as OpenAI's Whisper or Google's Cloud Speech-to-Text enable accurate, multi-speaker transcription.[1,10] LLMs (e.g., ChatGPT-4.0, Claude.ai, and Google's Med-Gemini family) drive summarization, sectioning, and style control.[1,10] On top of these foundations, commercial solutions provide "ambient" drafting and electronic medical record (EMR) handoff, including Nuance DAX Copilot and Amazon HealthScribe, as well as newer vendors such as Freed AI, Nabla Copilot, Tali, Heidi Health, Lyrebird Health, i-scribe, Scribeberry, and ScribeMD.[1,9,10] The capabilities described across this landscape include real-time or near-real-time capture, speaker attribution, entity extraction (e.g., problems, medications, orders), template-aware note assembly (e.g., subjective/objective/assessment/plan, operative notes, letters), and clinician-in-the-loop review with smart edits.[1,10]

Despite these tools aiming to reduce administrative workload while ensuring high-quality patient care, independent evaluations in surgical settings remain limited. The size and direction of these effects in surgical care remain uncertain, including key implementation, safety, and governance considerations specific to high-throughput perioperative workflows. This review aims to synthesize the current evidence on AI documentation tools used in surgical settings, focusing on their impact on documentation time, note quality and accuracy, integration and user perception, and downstream clinician and patient outcomes including burnout.

METHODS

This systematic review was conducted in accordance with the Cochrane Handbook for Systematic Reviews, using the Population, Intervention, Comparison, Outcomes, and Study Type framework.[11] Reporting follows the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines (**Appendix 1**).[12] The protocol was registered postoperatively in the PROSPERO International Prospective Register of Systematic Reviews (CRD420251156704).

Eligibility criteria

Eligible study designs included randomized trials, quasi-experimental, pre–post, or observational cohort/cross-sectional designs conducted in surgical settings (e.g., pre-operative consultation, intra-operative, and post-operative follow-up) that evaluated AI-enabled documentation tools compared to usual documentation workflows or a pre-implementation baseline. Case reports, edi-

doi: 10.14440/sti.0457

torials, protocols, conference abstracts without full data, narrative reviews, purely technical or algorithmic papers without clinical deployment, and studies not centered on surgical documentation outcomes were excluded. We included English-language full texts only due to the reviewers' language proficiency and to minimize misinterpretation of complex methodological and outcome data; this is acknowledged as a limitation.

Population, Intervention, Comparison, Outcome, and Study Type framework

The population of interest comprised surgeons and surgical teams in any specialty and setting (e.g., outpatient, inpatient/theatre, perioperative, postoperative). Mixed-specialty studies were eligible when surgical data were separable or comprised ≥95% of the sample. Interventions were AI-enabled documentation/scribe tools that capture, transcribe, structure, or draft clinical documentation (e.g., ambient digital scribes, LLM-assisted note drafting, advanced ASR/natural-language-processing systems integrated with the EMR). Comparators included usual documentation workflows (e.g., typed/templated notes, dictation, human scribes), alternative digital workflows, or pre–post within-subject designs.

The primary outcome measure was documentation efficiency, defined as: (i) time spent on documentation per clinical encounter (minutes); (ii) after-hours documentation time per session; and (iii) time to note finalization. Time to note finalization was defined as the interval between the patient encounter and completion of the clinical record or correspondence in the EMR system. Secondary outcomes included documentation quality (e.g., Physician Documentation Quality Instrument-9 [PDQI-9] or validated variants), including completeness, organization, accuracy, satisfaction/usability, clinician outcomes such as burnout and cognitive load, patient outcomes including satisfaction, and any safety or adverse events related to documentation, including AI hallucination.

Information sources and search strategy

A systematic search of MEDLINE (Ovid), Embase (Ovid), CENTRAL, and Scopus (January, 1, 2015, to October 9, 2025) was conducted using terms for AI and digital/ambient scribes, surgical settings, and administrative burden, documentation, or workflow (**Appendix 2**). The search strategy was developed with support from a senior medical librarian from the University of Sydney. Reference lists of included studies and relevant systematic reviews were manually searched for additional relevant studies.

Selection process

All search results were imported into EndNote 20 (Clarivate, United Kingdom) and de-duplicated using exact-match fields (e.g., title, author, year, DOI/PubMed ID), followed by a manual sweep of near-duplicates (e.g., minor title, author variants). Two reviewers independently screened titles and abstracts against prespecified eligibility criteria, then independently assessed full texts, with disagreements resolved by a third reviewer. Reasons for full-text exclusion were recorded and presented in the PRISMA flow diagram (**Figure 1**). No automation tools were used for selection; EndNote was used only for citation management and de-duplication.

Data collection process

Using a piloted, standardized data collection form, two reviewers independently extracted data from each included article, and discrepancies were resolved by discussion, with a third reviewer adjudicating if needed. No automation tools were used for data extraction.

Prespecified outcome domains were defined and extracted for all compatible results across measures and time points, prioritizing adjusted estimates. Study characteristics included study design; setting and country; surgical specialty; participant characteristics; AI tool details (e.g., vendor/generation if disclosed, functions, ambient capture, ASR, NLP, LLM drafting); comparator; outcome definitions, instruments, or time points; effect estimates and precision; adjustment variables; and blinding/timing for quality assessments (e.g., PDQI-9).

The primary outcome was documentation efficiency, defined as total clinician time required for review, editing, and sign-off (time-to-finalization) in minutes. Secondary outcomes included documentation time, defined as automated draft generation time (system processing time to produce a first draft), excluding surgeon review or sign-off. Documentation quality was measured using PDQI-9. PDQI-9 rates nine domains (up-to-date, accurate, thorough, useful, organized, comprehensible, succinct, synthesized, and internally consistent) on a 1–5 Likert scale; summed totals range 9–45, with higher scores indicating better note quality. Accuracy included error/amendment rates, while user perception, integration, and clinician burnout were defined by Likert satisfaction scores. Where multiple analyses were reported, all were included and the authors' designated primary analysis was noted.

Study risk of bias and certainty assessment

Risk of bias was assessed with the Risk of Bias in Non-randomized Studies of Exposures (ROBINS-E) tool, as no randomized studies were identified. Two independent reviewers assessed standard domains, with consensus or third-party adjudication as needed.[13] ROBINS-E ratings were assigned per outcome as low, moderate, serious, critical, or no information, with the overall study rating equal to the highest (worst) domain rating. Certainty assessment (Grading of Recommendations Assessment, Development, and Evaluation) was not conducted because results were reported at the single-study level for each outcome, precluding a body-of-evidence rating.

Data synthesis

We prespecified effect measures: mean or standardized mean differences for continuous outcomes (e.g., documentation time, PDQI-9), risk/odds ratios for dichotomous outcomes (e.g., error or amendment rates), and rate ratios for counts. When multiple measures or time points were reported, we extracted all compatible results and prioritized the authors' primary analysis, the longest relevant follow-up, and adjusted estimates.

Owing to heterogeneity in settings, interventions, comparators, outcomes, and designs, we conducted a structured narrative synthesis: study characteristics and effect estimates were tabulated, adjusted results were prioritized, and multiple reports from the same study were treated as one. Meta-analysis, subgroup analyses, and small-study assessments were not undertaken (<10 studies per outcome), and no automation tools were used.

Meta-analysis was not feasible because interventions spanned distinct modalities (LLM-drafted operative

doi: 10.14440/sti.0457

notes, ambient ASR+LLM clinic notes/letters, and ASR dictation), comparators differed (usual care, cross-tool comparisons, or none), and outcome measures were inconsistently reported (seconds vs. minutes, PDQI-9 vs. bespoke Likert/composite scores, and word error rate [WER]), often without compatible variance estimates. With fewer than 2–3 studies per outcome domain, quantitative pooling would be statistically unstable and clinically misleading.

RESULTS

Study characteristics

Seven studies published between 2023 and 2025 met the inclusion criteria (**Figure 1**). Detailed characteristics of the included studies are provided in **Table 1**. Only two studies involved patients,[14,15] while the remainder assessed the tools in simulated or structured scenarios rather than real-world deployment in routine workflows, limiting external validity. Clinical settings spanned otolaryngology, neurosurgery, plastic surgery, and urology. Three of the seven studies assessed the AI-assisted generation of operative reports utilizing prompts with OpenAI's LLM ChatGPT-4.[15-17] The remaining studies were conducted in outpatient clinic settings.[14,18,19]

Across studies, AI documentation combined ASR with LLMs for transcription, summarization, and structuring, typically using tools such as Whisper, Azure Speech+GPT-4o, and commercial ambient scribes. Interventions ranged from in-room ambient capture with auto-drafted clinic notes to AI-augmented operative reports, often in hybrid workflows where clinicians and AI sequentially refined the draft.

Risk of bias

Across seven non-randomized studies, ROBINS-E ratings indicated a moderate to serious risk of bias (moderate = 4, serious = 3; **Table 2**). The main concerns were unaddressed confounding (e.g., clinician experience, case mix), selection issues (e.g., small, single-center, or simulated samples), and outcome measurement (subjective ratings with variable blinding). Classification of interventions and deviations from intended interventions were generally at low to moderate risk, and missing data were minimal. In light of this profile and the heterogeneity across studies, we prioritized adjusted and objective estimates and interpreted effects cautiously.

Operative note generation (large language model drafting)

Documentation efficiency/time

Large-language-model tools consistently reduced the time required to generate operative reports across the included studies (**Table 3**). Abdelhady and Davis[15] reported that unedited ChatGPT-generated plastic surgery operative notes were produced in an average of 5.1 s, compared with 7.1 min for manually written reports (approximately a 99% reduction).

Similarly, Ali *et al.*[17] demonstrated that non-human–validated ChatGPT-generated cranial and spinal operative reports were produced in approximately 50 s, although this was not formally compared with surgeon-authored notes.

Hack *et al.*[16] reported that otolaryngology operative notes were generated fastest by ChatGPT alone (43 s) versus an attending–ChatGPT hybrid workflow (272 s), residents (408 s), and attendings (444 s) (hybrid vs. attending = −172 s, approximately a 39% reduction; ChatGPT vs. attending = −401 s, approximately a 97% reduction). No study reported after-hours documentation burden or time from the clinical encounter to note sign-off.

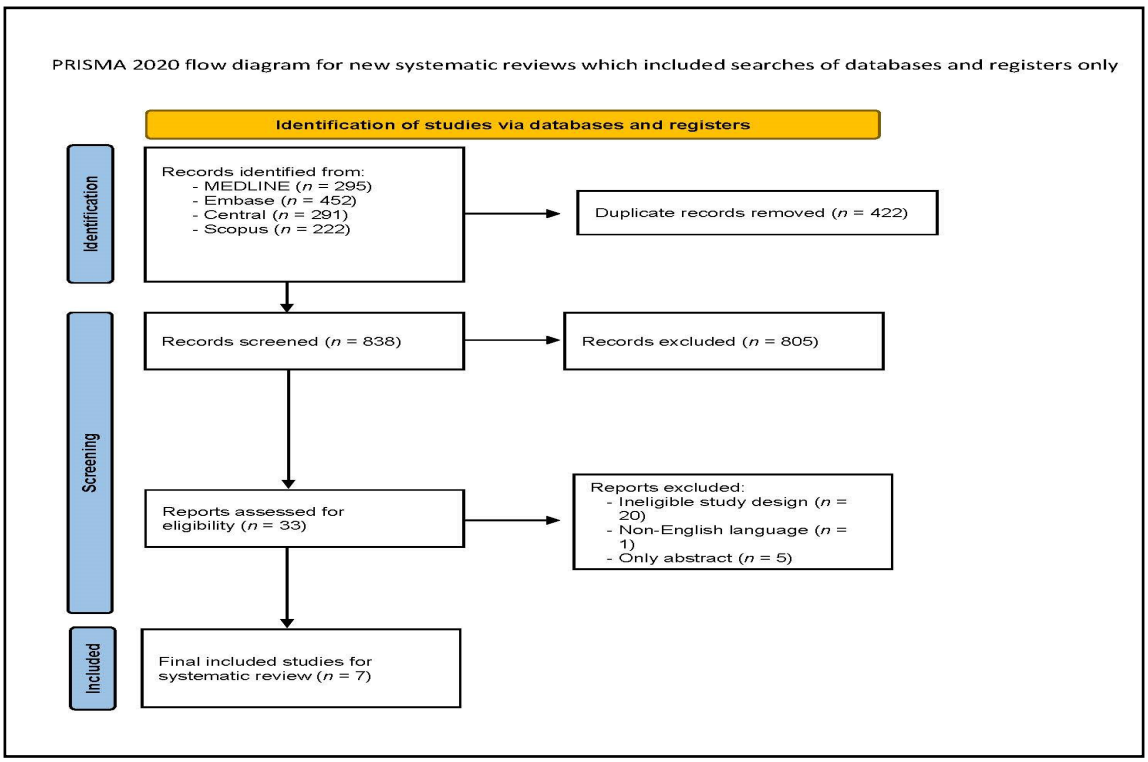

Figure 1. The Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) flow.

doi: 10.14440/sti.0457

Table 1. Characteristics of the included studies (*n* = 7)

Study	Country	Study design	Population	Intervention	Comparator	Sample size
Hack *et al*.[16]	Israel/UK	Prospective double-blind observational	Otolaryngology residents and attendings (perioperative)	ChatGPT 4.0 operative notes (OpenAI)	Attending/resident/hybrid	20 notes; 10 blinded reviewers
Ali *et al*.[17]	USA	Retrospective controlled comparison	Neurosurgery attendings (perioperative)	ChatGPT-4 operative reports (OpenAI)	Surgeon-authored reports	36 notes; 144 ratings by 3 neurosurgeons
Abdelhady and Davis[15]	UK	Prospective single-site observational	Plastic surgery residents and attendings (perioperative)	ChatGPT-4/ DALL-E operative notes	Manual surgeon reports	30 operative notes
Ong *et al*.[14]	Singapore	Prospective single-blind pilot	Urology outpatient clinic (attending urologist)	Custom ambient AI scribe (Azure Speech + GPT-4o)	Clinician-authored clinic letters	5 AI scribe notes, 5 clinician notes; 80 PDQI-9 evaluations
Moryousef *et al*.[19]	Canada	Multi-site observational, survey-based	Urology faculty and trainees (simulated clinics)	Nabla, Tali, Heidi, Scribeberry, and ScribeMD	Cross-tool comparison only	15 AI scribe notes; 20 reviewers
Hopkins *et al*.[18]	USA	Prospective single-blind comparative	Neurosurgeons dictating operative reports	OpenAI Whisper; Dragon Medical One (ASR)	Cross-tool comparison	10 operative reports
Thomson *et al*.[20]	Australia	Non-randomized observational	Urologists (simulated outpatient consults)	Heidi, Lyrebird Health, i-Scribe, Medow, and mAIscribe	Urology doctors	15 AI scribe notes (3 scenarios × 5 scribes)

Abbreviations: AI: Artificial intelligence; ASR: Automatic speech recognition; UK: United Kingdom; USA: United States of America; PDQI-9: Physician Documentation Quality Instrument-9.

Accuracy and documentation quality

Abdelhady and Davis[15] demonstrated that AI-generated operative notes had 100% adherence to preset prompt guidelines. Surgeons rated note quality at 4.2 ± 0.847 on a Likert scale. Edits averaged 2.1 per note, primarily correcting medical terminology, numerical values, patient demographics, and staff names.

Ali *et al*.[17] found that AI-generated neurosurgical notes were comparable to surgeon-written notes in accuracy (Likert score 4.44 vs. 4.33, $p = 0.512$) and organization (4.54 vs. 4.24, $p = 0.064$), but were less detailed (3.73 vs. 4.42, $p < 0.001$) and used more complex language (Flesch–Kincaid Grade Level 13.13 vs. 9.99, $p = 0.001$).

Hack *et al*.[16] reported that hybrid attending–generative pre-trained transformer (GPT) generated notes had the highest composite quality and "as-is" approval rate (79%) compared with attending-only (68%), resident-only (52%), and GPT-only (23%) notes.[16] Hybrid notes outperformed residents on Likert scores assessing completeness ($p < 0.05$), maneuver description ($p < 0.001$), and step sequencing ($p < 0.01$).[16] GPT-only drafts, while fastest, had the lowest approval rates and more omissions and verbosity.

Workflow integration and user perception

Abdelhady and Davis[15] reported high satisfaction among both surgeons (high/very high: 25/30, 83%) and patients (high/very high: 26/30, 86%) when AI-generated operative reports were supplemented with DALL-E–generated visual illustrations.

doi: 10.14440/sti.0457

Table 2. Risk-of-bias assessment: ROBINS-E domain ratings by study

Study	Confounding	Selection of participants	Classification of interventions	Deviations from intended interventions	Missing data	Measurement of outcomes	Selection of reported result	Overall
Hack *et al.*[16]	Moderate	Moderate	Low–moderate	Low–moderate	Low	Moderate	Moderate	Moderate
Ali *et al.*[17]	Serious	Moderate	Low–moderate	Low–moderate	Low	Moderate–serious	Moderate	Serious
Abdelhady and Davis[15]	Serious	Serious	Low–moderate	Low–moderate	Low	Serious	Moderate	Serious
Ong *et al.*[14]	Moderate	Moderate	Low–moderate	Low–moderate	Low	Moderate	Moderate	Moderate
Moryousef *et al.*[19]	Serious	Serious	Low–moderate	Low–moderate	Low	Serious	Moderate	Serious
Thomson *et al.*[20]	Moderate	Moderate	Low–moderate	Low–moderate	Low	Moderate	Moderate	Moderate
Hopkins *et al.*[18]	Moderate	Moderate	Low–moderate	Low–moderate	Low	Low	Moderate	Moderate

Burnout and cognitive load

Hack *et al.*[16] and Ong *et al.*[14] reported time savings and improved focus on patient interaction as potential contributors to reduced stress and cognitive load. Although none of the included studies quantitatively measured burnout, several identified the reduction of administrative burden as a primary rationale for adopting AI documentation tools.

Ambient scribes (automatic speech recognition + large language model)

Documentation efficiency/time

Only Thomson *et al.*[20] reported documentation efficiency measured as in-session drafting time (**Table 3**), showing that commercially available AI scribes could process a full urology consultation and generate both the clinic note and referral letter in a mean of 5.16 min, compared with 10.58 min for clinicians performing the same task (−5.42 min; approximately a 51% reduction).

Accuracy and documentation quality

Ong *et al.*[14] reported that ambient AI scribes (Azure + GPT-4o) produced urology clinic notes comparable to those produced by clinicians (PDQI-9: 33.6/45 vs. 32.9/45, $p = 0.412$; total PDQI-9 range 9–45) (**Table 3**). Domain scores for accuracy, thoroughness, usefulness, organization, comprehensibility, and synthesis were similar, with no significant differences (all $p > 0.05$).[14] AI-generated notes were preferred in 55% of evaluator ratings, suggesting non-inferiority in documentation quality.

Thomson *et al.*[20] reported higher accuracy for AI-generated notes (93.39%) than for clinician notes (85.64%), with fewer errors (AI scribe: 0.40 vs. clinician: 1.48 errors per note).

Moryousef *et al.*[19] assessed commercially available AI scribes using composite evaluation scores based on brevity, accuracy, quality, thoroughness, structure, and readability. Nabla demonstrated the highest overall accuracy (68%) and the lowest critical error score (28%).

Workflow integration and user perception

Moryousef *et al.*[19] conducted a national survey of Canadian urologists to evaluate freely accessible AI scribes, including Nabla, Tali, Heidi, Scribeberry, and ScribeMD. Among the 20 respondents, 75% indicated interest in using AI scribes for clinical documentation, while 90% were open to use if improvements are implemented. Furthermore, 16 of these 18 respondents (89%) believed that AI scribes would enable more patients to be seen and allow additional appointments to be scheduled.

Thomson *et al.*[20] reported that blinded urologists rated AI-scribe–generated letters as easier to understand than clinician-written letters (mean Likert score: 4.33 ± 0.82 vs. 3.10 ± 2.28; $p < 0.05$).[20]

doi: 10.14440/sti.0457

Table 3. Results of the included studies (*n* = 7)

Study	AI tool	Documentation efficiency/time	Accuracy and quality	Integration and perception	Burnout/cognitive load
LLM generation of operative reports					
Hack *et al.*[16]	ChatGPT-4 (otolaryngology operative reports)	Efficiency not reported; per-note time: GPT-only = 43 s; hybrid = 272 s; resident = 408 s; attending = 444 s. (hybrid vs. attending = −172 s; ~39% reduction)	Hybrid had highest as-is approval (79%) vs. attending (68%), resident (52%), and GPT-only (23%); hybrids outperformed residents for completeness, maneuvre description and step sequencing	N/A	Qualitatively described time savings Increased paitent interaction and decreased burden
Ali *et al.*[17]	ChatGPT-4 (neurosurgical operative reports)	Efficiency not reported; AI reports ~50 s (35–80 s); no surgeon time data	AI notes had similar accuracy ($p = 0.512$) and organization ($p = 0.064$) to surgeon reports but were less detailed ($p < 0.001$) and written at a higher reading level ($p < 0.001$)	N/A	N/A
Abdelhady and Davis[15]	ChatGPT-4 + DALL-E (plastic surgery operative reports)	Efficiency not reported; AI notes 5.1 s vs. 7.1 min for human-generated notes (−420.9 s; ~99% reduction)	100% adherence to the AI prompt; surgeon quality rating 4.2/5 with few edits	Surgeon and patient satisfaction high (Likert scale ≈4.2–4.3/5)	N/A
AI scribes (ASR + LLM)					
Ong *et al.*[14]	Azure Speech + GPT-4o ambient scribe	N/A	AI vs. clinician PDQI-9: 33.6/45 vs. 32.9/45 (non-inferior); similar domain scores, AI slightly less succinct (total PDQI-9 range 9–45)	N/A	Identified decreased admin burden as rational for adopting AI tools.
Moryousef *et al.*[19]	Nabla, Tali, Heidi, Scribeberry, and ScribeMD	N/A	Best composite accuracy: Nabla (68%), Tali (64%); most notes had 0–3 errors, but 25–75% contained ≥1 critical error depending on scenario	75% would adopt AI scribes, 90% if improved; 89% anticipated change in practice	75% of respondents described admin as major contributor to burnout. 80% respondents fellows or jr residents
Thomson *et al.*[20]	Heidi, Lyrebird Health, i-Scribe, Medow, and mAIscribe	AI scribes 5.16 min vs. urologists 10.58 min to complete note + letter ($p < 0.001$) (−5.42 min; ~51% reduction)	Higher accuracy (93.39% vs. 85.64%) and fewer errors (0.4 vs. 1.48 per note) than doctors	Blinded urologists rated AI letters easier to understand (4.33 vs. 3.10/5)	N/A
Dictation tools (ASR)					
Hopkins *et al.*[18]	Whisper vs. Dragon Medical One (ASR)	N/A	Overall word error rate: Whisper 1.75% vs. Dragon 1.54% (Δ +0.21% absolute; non-inferior); excluding linguistic errors, Whisper superior (0.50% vs. 1.34%, Δ −0.84% absolute; $p < 0.001$)	N/A	N/A

Abbreviations: AI: Artificial intelligence; ASR: Automatic speech recognition; LLM: Large language model; N/A: Not available; PDQI-9: Physician Documentation Quality Instrument-9.

doi: 10.14440/sti.0457

Burnout and cognitive load

Moryousef *et al.*[19] reported that 75% of respondents identified documentation workload as a major contributor to urologist burnout and expected AI scribes to alleviate this burden. Among those experiencing burnout, 80% were junior residents or fellows (**Table 3**). None of the included studies quantitatively assessed burnout outcomes.

Dictation tools (automatic speech recognition)

Documentation efficiency/time

No studies evaluated documentation efficiency or time.

Accuracy and documentation quality

Hopkins *et al.*[18] reported that OpenAI's Whisper dictation model demonstrated a mean WER of 1.75%, compared with 1.54% for Dragon Medical One, demonstrating non-inferiority (**Table 3**). When linguistic errors were excluded, Whisper outperformed Dragon Medical One, with a significantly lower WER (0.50% vs. 1.34%, $p < 0.001$).

Workflow integration and user perception

No studies evaluated workflow integration and user perception.

Burnout and cognitive load

No studies evaluated burnout and cognitive load.

Hallucination

Across studies, fully automated AI documentation was more error-prone than human-supervised workflows. Despite this, hallucinations (fabricated or misattributed details) were only reported in Hack *et al.*,[16] where GPT-only notes had the lowest "as-is" approval rate (23%) and while eloquent, "frequently contained factual errors, hallucinations, or overgeneralizations." However, hybrid notes required no alterations to "procedural steps," "clinical content", or "hallucinated findings."

In urology, Ong *et al.*[14] identified hallucination errors and automation bias as key barriers, advocating AI use only as a complement to clinician documentation. Other studies mainly reported errors in anatomical terms/proper nouns and subtle omissions or misinterpretations, rather than overt hallucinations.

Implementation and governance considerations

Across studies, implementation factors were frequently identified as key determinants of feasibility and safety, including consent/notification processes, data storage and residency, degree of clinician oversight, workflow/EMR integration, and post-deployment monitoring. To consolidate these recurring domains into a practical checklist for surgical services considering adoption, we summarize implementation readiness domains in **Table 4**.

DISCUSSION

Across contemporary surgical settings, AI-enabled documentation tools accelerate note generation and streamline workflows without significantly reducing overall quality. Hybrid, clinician-in-the-loop approaches (LLM with expert verification) achieved the best balance of speed, completeness, and approval, while AI-only drafts were

Table 4. Implementation readiness domains for AI documentation tools in surgery (practical checklist)

Domain	Key questions (examples)	Minimum safeguards/metrics
Privacy, security, and data residency	Where is audio/text processed and stored?; For how long?; Is patient data used to train models?; Encryption in transit/at rest?	Local/contracted residency as required; encryption; retention limits; explicit no-training clause; breach notification and audit logs
Consent and transparency	How are patients informed?; Opt-out vs. opt-in?; What is recorded (ambient capture scope)?	Plain-language consent/notification; clear scope; signage; ability to pause/stop recording; documentation of consent status
Clinical safety and accountability	Who signs-off?; How are hallucinations/omissions detected?; What is the escalation pathway?	Clinician-in-the-loop finalization; sampling QA; addendum/amendment tracking; incident reporting; automation-bias training
Workflow and EMR integration	How does the tool integrate (copy/paste vs. application programming interface)?; Does it fit templates (SOAP/op note)?; Does it reduce clicks?	Interoperability plan; measurable time-to-sign-off; usability testing; fallback workflow
Monitoring and ongoing governance	How are model/version updates handled?; Are performance drifts monitored?; Vendor reporting and change control?	Version disclosure and change logs; periodic re-validation; KPI dashboard (e.g., time, quality, errors); governance oversight

Abbreviations: EMR: Electronic medical record; KPI: Key performance index; QA: Quality assurance; SOAP: Subjective/objective/assessment/plan.

doi: 10.14440/sti.0457

fastest but more error-prone and less frequently "as-is" acceptable. Outpatient ambient scribes produced notes comparable to clinician documentation on PDQI-type metrics, and modern ASR engines were non-inferior to a leading commercial dictation tool with respect to WER. As these systems adapt to consistent editing patterns and align with a surgeon's preferred structure and language, the quality and speed of outputs are likely to improve further. Despite this, the evidence for downstream benefits (reduced after-hours charting, cognitive load, or burnout) is plausible but remains indirect.

Strengths and weaknesses

We used a prospectively registered protocol, dual independent screening and data extraction, and prespecified outcomes spanning efficiency, quality, safety signals, and governance. The resultant evidence base is early and collates heterogeneous AI modalities (e.g., ambient scribes, dictation, LLM-assisted drafting) relevant to real surgical workflows. Most studies were small, single-center, or simulated, relied on proxy outcomes, had variable comparators and short follow-up, and inconsistently reported model/version, vendor configuration, and conflicts of interest. Few eligible studies and heterogeneity precluded meta-analysis. Restrictions to English-language studies, database limitations, and rapid tool evolution introduce selection and publication bias, and some findings may already be out of date given the fast model iteration and publication lag time, tempering the strength and generalizability of conclusions. The English-language restriction may have excluded relevant studies, biasing toward English-speaking settings. Future updates should incorporate multi-language screening and translation support.

Relevance to other studies

Our findings align with emerging primary-care and general ambulatory literature showing reduced documentation burden and stable note quality with ambient AI scribes.[21-23] Our synthesis extends this to surgical contexts (operative and clinic notes), highlighting a consistent advantage for human-supervised hybrid workflows over AI-only drafting. Differences arise in specialty nuance: operative reports benefit from AI structure but risk omissions or over-documentation without expert review, while clinic notes show near-equivalence on PDQI domains yet still require clinician synthesis for subtle reasoning. Compared with non-surgical reports of productivity gains, surgical studies more often emphasized governance (e.g., privacy, consent, auditability) and terminology accuracy, reflecting higher stakes around procedural detail.[21-23]

Although no included studies directly measured burnout, multiple proxies moved in a favorable direction (e.g., shorter drafting times, smoother downstream tasks, improved perceived workflow). Other non-surgical studies have reported improvements in task load and burnout.[21] We also note that an industry/marketing report from an AI scribe vendor in the National Health Service primary care (grey literature) describes large reductions in documentation time and cognitive load; however, these estimates are not peer-reviewed and should be interpreted cautiously, serving only as contextual background rather than clinical evidence.[23]

Meaning and implications for practice and policy

The commercial success of ambient AI scribes centers on the synergy of each technology's strengths: ASR provides fast, low-latency capture of dialogue and objective data, and LLMs excel at structuring narratives, synthesizing salient problems, and mapping plans to standardized templates. Accuracy gains are most evident in the organization and capture of routine elements, with residual errors concentrated in specialty terminology and nuanced findings that benefit from clinician review. The net effect is a fundamental change to the existing documentation workflow: clinicians spend less time typing and instead validate and personalize the note, thus improving throughput and freeing time for direct patient care. This expert oversight remains the critical safety step to confirm clinical reasoning and ensure safe, context-appropriate finalization.

For clinicians and policymakers, rapid adoption of AI documentation tools hinges on governance that squarely addresses privacy/data governance, consent, transparency, and accountability. Ambient capture of sensitive dialogue demands clear data-flow rules (capture scope, storage location/duration, training use), explicit model and version disclosure (including on-device versus cloud), audit trails, and defined liability. Consent and transparency should be explicit: patients should be informed when AI is used to generate clinical documentation, what data are captured and where they are stored, and who remains accountable for the final record, with clear opt-out pathways where feasible. Because AI outputs can create "automation bias" (over-trust and reduced scrutiny), implementation should mandate clinician verification as a safety step and include monitoring for systematic errors, omissions, and drift over time. Controls should prioritize data minimization, encryption, legal compliance, adversarial resilience, and explicit consent. Contracts must allocate controller/processor roles, specify breach notification and cost-sharing, and bar undisclosed model training on clinical data. Procurement should require local data residency where applicable, security attestations, human-in-the-loop sign-off, and measurable outcomes such as turnaround time, rework, and patient-centered metrics. Implementation should consider all minimum standards and key questions described in **Table 4**, including staff training, patient information/consent, and continuous quality and safety surveillance.

Future directions

Future work should quantify surgeon trust in AI as a primary implementation outcome and link it to measurable effects on clinician well-being using validated instruments (Maslach Burnout Inventory, Copenhagen Burnout Inventory, and cognitive-load measures such as NASA-TLX). Patient-reported outcome measures should be concurrently collected, as efficiency gains must not compromise patient experience. Large, real-world studies are needed to characterize safety signals (hallucination, misattribution) and evaluate safeguards across accents, languages, noise, subspecialties, and multimorbidity. Cost-effectiveness trials should compare ambient scribes against enhanced dictation tools and human scribes, capturing total cost, throughput, and rework. To enable future meta-analyses, we propose minimum reporting standards for AI documentation studies, including model/vendor/version, supervision level (AI-only vs clinician-in-the-loop), reviewer role, integration details (EMR interface/application programming interface, templating, order sets), data-flow and data residency, consent model,

doi: 10.14440/sti.0457

and a core outcomes set (documentation minutes including after-hours time, turnaround to sign-off, PDQI-9 or validated quality metrics, addendum/amendment rate, error/hallucination rate, and patient/clinician satisfaction).

CONCLUSION

Across contemporary surgical settings, early evidence suggests AI-enabled documentation tools can accelerate note generation and streamline workflows while maintaining broadly comparable note quality, particularly when ASR is paired with LLM drafting and clinician verification. However, most available studies are small and/or simulated, at moderate-to-serious risk of bias, and do not directly measure downstream outcomes such as after-hours burden, safety events, or burnout.

Accordingly, supervised clinician-in-the-loop pilots with mature governance may be reasonable, but robust, prospective, real-world evaluations remain the priority.

AUTHORS' DISCLOSURE

Professor Daniel Steffens holds a Cancer Institute NSW Career Development Fellowship. No other authors have received any funding or support.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1. Kunze KN, Bepple J, Bedi A, Ramkumar PN, Pean CA. Commercial Products Using Generative Artificial Intelligence Include Ambient Scribes, Automated Documentation and Scheduling, Revenue Cycle Management, Patient Engagement and Education, and Prior Authorization Platforms. *Arthroscopy*. 2025;41(11):4950-4955. doi: 10.1016/j.arthro.2025.05.021
2. Dimou FM, Eckelbarger D, Riall TS. Surgeon burnout: a systematic review. *J Am Coll Surg*. 2016;222(6):1230-1239. doi: 10.1016/j.jamcollsurg.2016.03.022
3. Kataria S, Ravindran V. Electronic health records: a critical appraisal of strengths and limitations. *J R Coll Physicians Edinb*. 2020;50(3):262-268. doi: 10.4997/jrcpe.2020.309
4. Kroth PJ, Morioka-Douglas N, Veres S, *et al*. Association of electronic health record design and use factors with clinician stress and burnout. *JAMA Netw Open*. 2019;2(8):e199609. doi: 10.1001/jamanetworkopen.2019.9609
5. McPeek-Hinz E, Boazak M, Sexton JB, *et al*. Clinician burnout associated with sex, clinician type, work culture, and use of electronic health records. *JAMA Netw Open*. 2021;4(4):e215686. doi: 10.1001/jamanetworkopen.2021.5686
6. Melnick ER, Dyrbye LN, Sinsky CA, *et al*. The association between perceived electronic health record usability and professional burnout among US physicians. *Mayo Clin Proc*. 2020;95(3):476-487. doi: 10.1016/j.mayocp.2019.09.024
7. Varghese C, Harrison EM, O'Grady G, Topol EJ. Artificial intelligence in surgery. *Nat Med*. 2024;30(5):1257-1268. doi: 10.1038/s41591-024-02970-3
8. Chryssofos S, Ochoa E, Sacks JM. The Digital Scribe: A New Wave of Efficiency and Quality of Life for Plastic Surgeons. *Plast Reconstr Surg Glob Open*. 2025;13(5):e6754. doi: 10.1097/GOX.0000000000006754
9. van Buchem MM, Kant IMJ, King L, Kazmaier J, Steyerberg EW, Bauer MP. Impact of a Digital Scribe System on Clinical Documentation Time and Quality: Usability Study. *JMIR AI*. 2024;3(1):e60020. doi: 10.2196/60020
10. Ormond MJ, Garling EH, Woo JJ, Modi IT, Kunze KN, Ramkumar PN. Artificial Intelligence in Commercial Industry: Serving the End-to-End Patient Experience Across the Digital Ecosystem. *Arthroscopy*. 2025;41(5):1683-1690. doi: 10.1016/j.arthro.2025.01.064
11. Higgins J, Thomas J, Chandler J, *et al*. *Cochrane Handbook for Systematic Reviews of Interventions version 6.3 (updated February 2022)*. Cochrane; 2022.
12. Page MJ, McKenzie JE, Bossuyt PM, *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. doi: 10.1136/bmj.n71
13. Bero L, Chartres N, Diong J, *et al*. The risk of bias in observational studies of exposures (ROBINS-E) tool: concerns arising from application to observational studies of exposures. *Syst Rev*. 2018;7(1):242. doi: 10.1186/s13643-018-0915-2
14. Ong JHW, Tung JYM, Sng GGR, *et al*. A pilot study using ambient artificial intelligence scribes in clinical documentation in a urology outpatient clinic. *BJU Int*. 2025;136(3):417. doi: 10.1111/bju.16784
15. Abdelhady AM, Davis CR. Plastic Surgery and Artificial Intelligence: How ChatGPT Improved Operation Note Accuracy, Time, and Education. *Mayo Clin Proc Digit Health*. 2023;1(3):299-308. doi: 10.1016/j.mcpdig.2023.06.002
16. Hack S, Attal R, Locatelli G, *et al*. Surgeon, Trainee, or GPT? A Blinded Multicentric Study of AI-Augmented Operative Notes. *Laryngoscope*. 2025. doi: 10.1002/lary.70063
17. Ali A, Kumar RP, Polavarapu H, *et al*. Bridging the Gap: Can Large Language Models Match Human Expertise in Writing Neurosurgical Operative Notes? *World Neurosurg*. 2024;192:e34-e41. doi: 10.1016/j.wneu.2024.08.062
18. Hopkins BS, Dallas J, Yu J, *et al*. The use of generative artificial intelligence-based dictation in a neurosurgical practice: a pilot study. *Neurosurg Focus*. 2025;59(1):E8. doi: 10.3171/2025.4.FOCUS24834
19. Moryousef J, Nadesan P, Uy M, Matti D, Guo Y. Assessing the Efficacy and Clinical Utility of Artificial Intelligence Scribes in Urology. *Urology*. 2025;196:12-17. doi: 10.1016/j.urology.2024.11.061
20. Thomson A, Perera M, Murphy D, Lawrentschuk N. Scribe smarter, not harder: how artificial intelligence scribes stack up against human clinicians. *BJU Int*. 2025;137(1):15-17. doi: 10.1111/bju.70037
21. Shah SJ, Crowell T, Jeong Y, *et al*. Physician Perspectives on Ambient AI Scribes. *JAMA Netw Open*. 2025;8(3):e251904. doi: 10.1001/jamanetworkopen.2025.1904
22. Shah SJ, Devon-Sand A, Ma SP, *et al*. Ambient artificial intelligence scribes: physician burnout and perspectives on usability and documentation burden. *J Am Med Inform Assoc*. 2025;32(2):375-380. doi: 10.1093/jamia/ocae295
23. Albrecht M, Shanks D, Shah T, *et al*. Enhancing clinical documentation with ambient artificial intelligence: a quality improvement survey assessing clinician perspectives on work burden, burnout, and job satisfaction. *JAMIA Open*. 2024;8(1):ooaf013. doi: 10.1093/jamiaopen/ooaf013

doi: 10.14440/sti.0457

APPENDIX 1

Section and topic	Item #	Checklist item	Location where item is reported (page)
Title			
Title	1	Identify the report as a systematic review.	1
Abstract			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	1
Introduction			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	2
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	2
Methods			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	3
Information sources	6	Specify all databases, registers, websites, organizations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	3
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	3
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	4
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	4
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	4
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	4
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	4
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	4
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis [item #5])	4
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions	4
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses	4
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used	4
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression)	4
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results	4

doi: 10.14440/sti.0457

Section and topic	Item #	Checklist item	Location where item is reported (page)
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases)	4
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome	4
Results			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram	5
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded	5
Study characteristics	17	Cite each included study and present its characteristics	5
Risk of bias in studies	18	Present assessments of risk of bias for each included study	5
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots	5
Results of syntheses	20a	For each synthesis, briefly summarize the characteristics and risk of bias among contributing studies	5
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect	5
	20c	Present results of all investigations of possible causes of heterogeneity among study results	5
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results	6
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed	6
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed	6
Discussion			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence	7
	23b	Discuss any limitations of the evidence included in the review	7
	23c	Discuss any limitations of the review processes used	7
	23d	Discuss implications of the results for practice, policy, and future research	7
Other information			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered	2
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared	2
	24c	Describe and explain any amendments to information provided at registration or in the protocol	2
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review	1
Competing interests	26	Declare any competing interests of review authors	1
Availability of data, code, and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review	1

Note: Adapted from Page *et al.*[12]

doi: 10.14440/sti.0457

APPENDIX 2

Title:

Search strategy

Review question: What are the artificial intelligence tools and their effect on the administrative burden of surgeons?

Aim: To synthesize and evaluate the current evidence on AI tools in surgical practice

Conduct a systematic review to identify and assess AI tools designed to reduce administrative burden in the surgical setting.

(i) Scoping review of AI tools reducing administrative burden of surgeons in pre-operative setting
(ii) Assessment of the effect of these AI tools on decreasing burnout and increasing clinician satisfaction

Core terms

Artificial intelligence

Pre-operative surgical setting (including urology and broader surgical contexts)

Administrative burden/workload/burnout

Outcomes related to efficiency, satisfaction, or workflow

Search structure

Search set 1: Artificial intelligence and related terms

*Artificial Intelligence/**

*Machine Learning/**

*Neural Networks (Computer)/**

*Deep Learning/**

(artificial intelligence OR AI OR "machine learning" OR "deep learning" OR "expert system*" OR "expert systems" OR "neural network*" OR "natural language processing" OR "predictive analytics" OR "computer assisted" OR "computer-aided" OR "computer simulation").ti,ab,kw.

Search set 2: Pre-operative setting in surgery/urology

Surg* OR Urolog* OR preoperative OR "pre-operative" OR perioperative OR "peri-operative" OR presurgical OR "pre-surgical" OR "pre-surgery"

*Preoperative Care/** OR *Perioperative Care/**

(preoperative OR "pre-operative" OR perioperative OR "peri-operative" OR presurgical OR "pre-surgical" OR "pre-surgery". ti,ab,kw.

*Surgical Procedures, Operative/**

*Urologic Surgical Procedures/**

(surgery OR surgical OR urology OR "urologic*" OR "urological" OR "surgical patient*" OR "operative care").ti,ab,kw. Surg*

Urolog*

Search set 3: Administrative burden, workload, burnout, and related outcomes

administrative adj2 burden OR clerical adj2 burden OR documentation OR paperwork OR "data entry" OR "clerical tasks" OR workload OR "work load" OR "task shifting" OR "efficienc*" OR "workflow OR satisfaction OR "clinician satisfaction" OR "patient satisfaction" OR "staff satisfaction OR wellbeing OR "well-being" OR "quality of work life" OR "work-life balance"

Burden

(administrative adj2 burden).ti,ab,kw.

(clerical adj2 burden).ti,ab,kw.

(documentation OR paperwork OR "data entry" OR "clerical tasks").ti,ab,kw.

(workload OR "work load" OR "task shifting" OR "efficienc*" OR "workflow".ti,ab,kw.

*Burnout, Professional/**

(burnout OR "moral injury").ti,ab,kw.

(satisfaction OR "clinician satisfaction" OR "patient satisfaction" OR "staff satisfaction").ti,ab,kw.

(wellbeing OR "well-being" OR "quality of work life" OR "work-life balance").ti,ab,kw.

*Efficiency, Organizational/**

*Resilience, Psychological/**

4. Limit by Date and Language

Publication date: January 2015 to current date

MEDLINE:

((("Artificial Intelligence"[Mesh] OR "Machine Learning"[Mesh] OR "Neural Networks (Computer)"[Mesh] OR "Deep Learning"[Mesh]) OR (artificial intelligence[tiab] OR AI[tiab] OR "machine learning"[tiab] OR "deep learning"[tiab] OR "expert system*"[tiab] OR "neural network*"[tiab] OR "natural language processing"[tiab] OR "predictive analytics"[tiab] OR "computer assisted"[tiab] OR "computer-aided"[tiab] OR "computer simulation"[tiab]))) AND ((("Preoperative Care"[Mesh] OR "Perioperative Care"[Mesh]) OR (preoperative[tiab] OR "pre-operative"[tiab] OR perioperative[tiab] OR "peri-operative"[tiab] OR presurgical[tiab] OR "pre-surgical"[tiab] OR "pre-surgery"[tiab] OR "preoperative period"[tiab] OR "preoperative management"[tiab])) AND (("Surgical Procedures, Operative"[Mesh] OR "Urologic Surgical Procedures"[Mesh]) OR (surgery[tiab] OR surgical[tiab] OR urology[tiab] OR "urologic*"[tiab] OR "urological"[tiab] OR "surgical patient*"[tiab] OR "urologic patient*"[tiab] OR "urological patient*"[tiab] OR "operative care"[tiab]))) AND ((administrative burden[tiab] OR clerical burden[tiab] OR documentation[tiab] OR paperwork[tiab] OR "data entry"[tiab] OR "clerical tasks"[tiab] OR workload[tiab] OR "work load"[tiab] OR "task shifting"[tiab] OR efficienc*[tiab] OR workflow[tiab] OR "workflow improvement"[tiab] OR "Burnout, Professional"[Mesh] OR burnout[tiab] OR "physician burnout"[tiab] OR "clinical burnout"[tiab] OR "professional burnout"[tiab] OR "moral injury"[tiab] OR satisfaction[tiab] OR "clinician satisfaction"[tiab] OR "patient satisfaction"[tiab] OR "staff satisfaction"[tiab] OR wellbeing[tiab] OR "well-being"[tiab] OR "quality of work life"[tiab] OR "work-life balance"[tiab] OR "Efficiency, Organizational"[Mesh] OR "Resilience, Psychological"[Mesh])) AND ("2015/01/01"[PDAT] : "2025/10/09"[PDAT]) AND english[lang]

Medline by OVID

(artificial intelligence or AI or "machine learning" or "deep learning" or "neural network*" or "natural language processing" or "predictive analytics").mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kf, fx, dq, bt, nm, ox, px, rx, ui, sy, ux, mx]

AND

(preoperative or "pre-operative" or perioperative or "peri-operative" or presurgical or "pre-surgical" or "pre-surgery").mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kf, fx, dq, bt, nm, ox, px, rx, ui, sy, ux, mx]

AND

((((administrative adj2 burden) or clerical) adj2 burden) or documentation or paperwork or workload or "work load" or "efficienc*" or "workflow OR satisfaction OR clinician satisfaction OR patient satisfaction OR staff satisfaction" or wellbeing or "well-being" or "quality of work life" or "work-life balance").mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kf, fx, dq, bt, nm, ox, px, rx, ui, sy, ux, mx]

AND

doi: 10.14440/sti.0457

(Surg* or Urolog*).mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kf, fx, dq, bt, nm, ox, px, rx, ui, sy, ux, mx]

doi: 10.14440/sti.0457