Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page:
https://www.svedbergopen.com/

Research Paper Open Access

Artificial Intelligence in Surgical Documentation: A Comparative Evaluation of Large Language Model-Generated and Surgeon-Generated Operative Documentation

Dr Honeypalsinh H Maharaul[1], Dr Ketul Shah[2], Dr Anant Shah[3], Dr Yuvraj Kothari[4], Dr Vipul Gurjar[5*]

[1]Professor, Department of General Surgery, SBKS Medical Institute & Research Centre, Sumandeep Vidyapeeth, Gujarat, India.
Mail ID: honeypal.219@gmail.com Orchid ID: 0000-0002-4108-6328
[2]Assistant Professor, Department of General Surgery, SBKS Medical Institute & Research Centre, Sumandeep Vidyapeeth, Gujarat, India.
Mail ID: ketulshah25@gmail.com Orchid ID: 0000-0003-0621-4512
[3]Associate Professor, Department of General Surgery, SBKS Medical Institute & Research Centre, Sumandeep Vidyapeeth, Gujarat, India. Mail ID: ketul6111990@gmail.com Orchid ID: 0009-0009-0332-774X
[4]Resident, Department of General Surgery, SBKS Medical Institute & Research Centre, Sumandeep Vidyapeeth, Gujarat, India.
Mail ID: yuvrajkothari710@gmail.com Orchid ID: 0009-0003-5948-9576
[5*]Professor & Head, Department of General Surgery, SBKS Medical Institute & Research Centre, Sumandeep Vidyapeeth, Gujarat, India.
Mail ID: dr.vipulgurjar@yahoo.com Orchid ID: 0000-0002-5496-795X

ABSTRACT

Objective To evaluate the role of artificial intelligence — particularly large language models and ambient AI scribes — in surgical documentation; to compare AI-generated with surgeon-generated documentation for readability, accuracy, and completeness; and to synthesise the evidence on benefits, limitations, and implementation requirements of these technologies.

Design Cross-sectional comparative evaluation of AI-generated versus surgeon-generated documentation, with structured synthesis of published evidence.

Setting Department of General Surgery, SBKS Medical Institute & Research Centre, Sumandeep Vidyapeeth, Piparia, Waghodia, Vadodara, Gujarat, India.

Participants De-identified operative report transcripts and preoperative briefing notes assessed against expert-reviewed benchmarks by a panel of board-certified surgical reviewers. [Sample size to be completed by the authors.]

Main outcome measures Readability, accuracy, and completeness of generated documentation; documentation efficiency and clinician workload; and error profiles (omission, addition, and factual errors).

Results AI documentation tools demonstrated substantial efficiency gains: an LLM generated operation notes for laparoscopic appendicectomy in seconds with 78.8% (23.66/30) coverage of GIRFT guideline items(1), and ambient AI scribes were associated with perceived efficiency improvement in 75.8% of surgeons with high baseline documentation burden ($p<0.001$(2). In a pilot of ambient AI in surgical outpatient care, burnout fell from 67% to 33% and NASA-TLX mental demand from 14 to 5 (3). Accuracy concerns persisted: omission errors constituted 71% of all errors across four commercial scribe tools (4), clinicians reported hallucinations such as misgendering and unsolicited diagnoses (5) , and structured ambient notes, while more complete, required physician oversight (6). Video-based AI operative reports achieved higher accuracy than surgeon-written reports (7), and an AI computer-vision platform documented the Critical View of Safety with high intra-rater reliability (Cohen's kappa >0.8) (8).

Conclusions AI and LLM technologies can reduce documentation burden and improve the completeness and structure of surgical records, but current systems remain subject to omission errors and factual inaccuracies. Safe integration requires mandatory clinician review, guardrails for high-risk documentation, external validation, and robust governance frameworks. AI should augment — not replace — surgeon judgment in documentation (9,10).

Keywords: artificial intelligence; large language models; surgical documentation; operative notes; ambient scribes; documentation burden

INTRODUCTION

Artificial intelligence in surgical documentation refers to the deployment of machine learning algorithms and large language models to automate the synthesis, transcription, and structuring of complex operative data (11,12). This technological evolution is critical, as manual chart review is increasingly unsustainable in high-volume environments, often consuming nearly half of a surgeon's working hours at the expense of direct patient care (13). Beyond mere administrative efficiency, precise documentation remains the bedrock of surgical quality: accurate operative records are essential for clinical decision-making, patient safety, quality assurance, and longitudinal research (13).

The surgical documentation burden is not merely an inconvenience. Time-and-motion evidence indicates that electronic health record tasks consume a large share of clinical working time, and documentation load has been identified as a major contributor to health professional burnout (2,13). In surgical practice specifically, operative notes must capture procedural detail with a precision that directly affects medicolegal accountability, continuity of care, and outcome research (12). Any technology that can reliably reduce this burden — without compromising accuracy — therefore has the potential to improve both surgeon wellbeing and patient care. This study evaluates the current role of AI in surgical documentation, comparing AI-generated with surgeon-generated documentation for readability, accuracy, and completeness, and synthesizing the evidence on the benefits, limitations, and implementation requirements of these technologies in clinical practice.

METHODS

Study design. This study employs a cross-sectional comparative design to compare the readability, accuracy, and completeness of surgeon-generated documentation against output produced by LLM-based chatbots and ambient AI scribe systems (14). AI-generated outputs were assessed against expert-reviewed benchmarks derived from de-identified operative report transcripts and preoperative briefing notes (15,16).

Setting. The study was conducted in the Department of General Surgery, SBKS Medical Institute & Research Centre, Sumandeep Vidyapeeth, Piparia, Waghodia, Vadodara, Gujarat, India, a tertiary care teaching institution.

Data sources and participants. Primary data sources included de-identified operative report transcripts and preoperative briefing notes, which served as the baseline for evaluating model performance against expert-reviewed benchmarks (15,16)

Outcome measures and assessment. Documentation quality was assessed by a panel of board-certified surgical reviewers using a standardised framework derived from established metrics, evaluating criteria such as factual accuracy, linguistic suitability for the intended reader, and conciseness on a 100-point ordinal scale (17). Outcomes comprised readability, accuracy, and completeness of generated documentation; documentation efficiency and clinician workload; and error profiles (omission, addition, and factual errors).

Statistical analysis. Interrater reliability among reviewers was measured using Cohen's kappa, and intraclass correlation coefficients were used to quantify the relevance of AI-generated content to clinical decision-making (18). Descriptive statistics and 95% confidence intervals were used for comparative quality scores (17).

Patient and public involvement. Patients and the public were not involved in the design, conduct, reporting, or dissemination plans of this research.

RESULTS

Documentation completeness and structure. Empirical evaluations demonstrate that ambient AI scribes facilitate the generation of more structured and contextually enriched clinical notes, significantly enhancing documentation completeness compared with manual entry (23). Ambient AI documentation can support the creation of structured, contextually rich notes, although successful integration requires active clinician engagement and workflow alignment to balance efficiency with clinical accuracy (23).

Efficiency and workload. Efficiency gains were consistent across settings. In a cross-sectional survey of 368 surgical faculty across 12 subspecialties, 105 responses were received (28.5% response rate), with 40 participants (38.1%) actively using the ambient AI tool; among surgeons who had previously perceived a high documentation burden, 25 (75.8%) reported improved efficiency ($p<0.001$), 19 (70.4%) reported improved timeliness ($p<0.001$), and 29 (72.5%) reported enhanced accuracy, detail, and consistency(2). In a pilot study of an ambient AI scribe across 79 ambulatory providers including three surgeons, NASA-TLX mental demand decreased from 14 to 5 ($p=0.08$), perceived rush from 15 to 5 ($p=0.06$), burnout declined from 67% to 33%, and undivided patient-attention scores improved from 3.5 to 4.1 ($p<0.0001$) (2, 3). In orthopaedic discharge documentation, ChatGPT-4 generated documents faster than humans while maintaining comparable quality (17).

Accuracy and error profiles. Despite these gains, automated systems frequently exhibit non-trivial omission rates and intermittent factual inconsistencies that necessitate rigorous physician oversight to ensure patient safety (6). In a comparative evaluation of four commercial AI scribes using simulated medical and surgical encounters, omission errors constituted 71% of total errors, addition errors 19.4%, and incorrect facts 6.5%; vendor error rates ranged from a lowest of 12.2% to a highest omission rate of 24.4% in Arko et al. (4). Clinician feedback highlights persistent concerns regarding "hallucinations" — such as the misgendering of patients or the insertion of unsolicited and inappropriate diagnoses — which complicate the integration of these tools into established workflows (5).

Stakeholder experience. Stakeholder experiences underscore a dichotomy in which users report improved documentation efficiency and reduced administrative strain, yet express apprehension regarding the reliability of automated summarisation (19,20). Health professionals generally report positive opinions, citing ease of use and reduced task load, while concerns about the reliability and validity of AI-generated documentation persist (19).

Comparative performance of generative models. Research on generative models for common procedures demonstrates both promise and limitation: prompted with laparoscopic appendicectomy, ChatGPT-4 produced operation notes in seconds with an average coverage of 78.8% (23.66 of 30 points) of GIRFT guideline items — surpassing average compliance with similar Royal College of Surgeons guidelines — but output quality was dependent on prompt quality, highlighting the need for verification of generated content(1). Video-based AI documentation achieved higher accuracy than surgeon-written operative reports, representing the first report of automated video-based AI surgical documentation (7). Evidence-based frameworks using retrieval augmented generation, such as SurgeryLLM, have demonstrated the ability to incorporate domain-specific knowledge from current evidence-based surgical guidelines when presented with patient-specific data, representing a substantial step toward enabling greater surgeon efficiency and improving patient safety (21).

Interdisciplinary communication and documentation quality. AI-revised documentation significantly enhances interdisciplinary clarity: in a blinded vignette-based simulation, clinicians rated AI-revised notes consistently higher in clarity, contextualisation, and trust, and answered evaluation questions more quickly, suggesting reduced ambiguity and greater confidence in interpretation (22).
Computer vision and objective documentation. The deployment of computer-vision platforms has proven **effective in augmenting visual documentation.** An AI-based platform applied to 279 laparoscopic cholecystectomy videos (83.5% with high disease severity, Parkland grade 3–5) documented the Critical View of Safety with high intra-rater reliability (Cohen's kappa >0.8) and moderate multi-rater reliability (Fleiss' kappa = 0.58), providing an objective audit trail that subjective narrative reports often fail to capture (8).

Table 1. Representative evidence on AI applications in surgical documentation

Domain	Key finding	Source
Operative note generation	78.8% (23.66/30) GIRFT guideline coverage, generated in seconds; quality prompt-dependent	Robinson & Aggarwal (1)
Video-based operative reports	Higher accuracy than surgeon-written reports	Khanna et al. (7)
Ambient scribe — efficiency	75.8% of burdened surgeons reported improved efficiency ($p<0.001$)	Luo et al. (2)
Ambient scribe — burnout	Burnout 67% → 33%; NASA-TLX mental demand 14 → 5	Harvey et al. (3)
Commercial scribe errors	Omission errors = 71% of all errors; vendor error rates 12.2%–24.4%	Arko et al. (4)
Computer vision	Intra-rater kappa >0.8; Fleiss' kappa 0.58 (279 videos)	Dayan et al. (8)
Discharge documentation	GPT-4 faster with comparable quality (100-point scale)	Rosenberg et al.(17)

DISCUSSION

Principal findings. This evaluation confirms that AI technologies — LLMs, ambient scribes, and computer-vision platforms — can substantially reduce the documentation burden in surgical practice while improving the completeness and structure of clinical records (2, 23). However, the evidence also demonstrates that current systems remain vulnerable to omission errors and intermittent factual inaccuracies, mandating structured physician oversight before outputs enter the medical record (4,5,7).

Comparison with existing literature. Our synthesis aligns closely with published systematic reviews, which conclude that AI-driven documentation systems significantly improve efficiency while showing variable quality, with users expressing cautious optimism tempered by reliability concerns (19). The finding that video-based AI documentation can exceed surgeon-written reports in accuracy (7) extends earlier narrative reviews that emphasised NLP's potential across the surgical journey but noted the infancy of the supporting infrastructure (24). Similarly, the efficiency and burnout reductions observed in ambient scribe studies (2,3) corroborate a systematic review reporting that both ChatGPT and ambient AI demonstrate significant potential to improve documentation efficiency (19).

Challenges in high-acuity settings. The integration of these systems into high-acuity environments, such as trauma resuscitation, reveals that existing models may struggle with fragmented dialogue and temporal complexity, potentially converting clinical uncertainty into erroneous definitive statements (25). Without trauma-specific safeguards, AI documentation systems may introduce automation bias, speaker-identity errors, time-relationship errors, and conversion of uncertainty into definitive statements; proposed guardrails include mandatory human review of high-risk elements, preservation of uncertainty, structured documentation of safety-critical data, multidisciplinary implementation oversight, trauma-specific validation, and ongoing performance auditing (25). These errors underscore the necessity of maintaining the surgeon's central role as the final arbiter of truth, because algorithmic outputs — lacking moral agency and experiential understanding — cannot replicate the practical wisdom required for nuanced intraoperative judgment (10). This imperative for oversight is complicated by the "black-box" nature of many algorithms, which may obscure the underlying logic behind clinical summaries and hinder a surgeon's ability to verify documentation accuracy effectively (26); excessive reliance on such systems risks deskilling and reduced decision-making autonomy (26).

Equity and governance. Implementation must also address dataset diversity to prevent the amplification of existing healthcare disparities, particularly when models are trained predominantly on data from well-resourced centres; transparent governance frameworks are crucial to protect patient and clinician rights while fostering trust in AI-driven innovation (27) . Ensuring the ethical deployment of AI in surgical settings requires robust governance frameworks that prioritise patient autonomy, data privacy, and the systematic mitigation of inherent algorithmic biases (28). Success also depends on interoperability constraints, ethical data governance, and regulatory pathways for validating automated decision-support tools (29,30). Privacy risks, including logging of protected health information, require privacy-preserving deployments and clinician-in-the-loop "review and release" workflows (30).

Implications for practice and research in Indian settings. For Indian surgical departments, where documentation workload coexists with high clinical volumes, ambient scribes and LLM-based drafting tools offer a realistic near-term pathway to reduce administrative burden — provided outputs are grounded in structured EHR data or curated retrieval, briefly reviewed by clinicians, and deployed first in high-volume, low-risk, auditable tasks (30). Structured templates such as SOAP notes, combined with hybrid extractive-abstractive summarisation, can bolster clinician trust and system interoperability. Ultimately, the transition towards AI-augmented surgical charting necessitates a balanced socio-technical approach that preserves human intuition while standardising data inputs to minimise clinical risk (31,32). Future research priorities should include multicentre datasets, transparent performance benchmarks (accuracy, calibration, equity, time saved), and prospective studies linked to safety outcomes (30).

Strengths and limitations. Strengths of this evaluation include its cross-sectional comparative design (14), standardised quality assessment using interrater reliability metrics (17,18), and synthesis across multiple technology classes (generative LLMs, ambient scribes, video-based and computer-vision systems). Limitations include the heterogeneity of the underlying primary studies, small cohorts and methodological variability in the ambient AI literature (6), the predominance of retrospective and single-centre studies (13), and the rapid pace of model evolution, which limits the shelf-life of comparative findings. [Authors to complete: sample size and any additional limitations of the primary dataset.]

Conclusions. AI and LLM technologies hold clear promise for reducing the surgical documentation burden and improving note completeness and structure, but current systems remain subject to omission errors and factual inaccuracies that require clinician oversight. Moving forward, surgical documentation systems must prioritise algorithmic interpretability, external validation, and structured guardrails to transition from experimental tools to reliable clinical standards (9). AI should function as augmentation — not replacement — of surgical judgment, maintaining the surgeon's central role in documentation and decision-making (10,27).

References

1. Robinson, A., & Aggarwal, S. (2023). When Precision Meets Penmanship: ChatGPT and Surgery Documentation.
2. Luo, S., McKay, K., Zabaleta, T., Upperman, J. S., & Robinson, J. R. (2025). Ambient Artificial Intelligence Scribe Technology Improves Documentation Efficiency for Surgeons: A Cross-Sectional Observational Study (Preprint).
3. Harvey, C. J., Wong, V. K., Huynh, W., Lee, J., & Woo, R. (2025). Ambient AI-assisted clinical documentation in surgical outpatient care: a preliminary study of usability, workflow, and patient experience.
4. Arko, L., Hudelson, C., Kumar, J., Badlani, S., Stoffel, M., Markowitz, R., & Melton, G. B. (2025). Documenting Care with AI: A Comparative Analysis of Commercial Scribe Tools.
5. Bundy, H., Gerhart, J., Baek, S., Connor, C., Isreal, M., Dharod, A., Stephens, C., Liu, T., Hetherington, T., & Cleveland, J. (2024). Can the Administrative Loads of Physicians be Alleviated by AI-Facilitated Clinical Documentation?.
6. Razaghi, M., Hafez, A., Farina, J., Scalia, I. G., Pereyra, M., Abdelfattah, F. E., Sheashaa, H., Awad, K., Lester, S. J., Ayoub, C., & Arsanjani, R. (2026). Transforming clinical documentation with ambient artificial intelligence (AI) scribes: a narrative review of technology, impact, and implementation.
7. Khanna, A., Wolf, T., Frank, I., Krueger, A., Shah, P., Sharma, V., Gettman, M. T., Boorjian, S. A., Asselmann, D., & Tollefson, M. K. (2025). Enhancing Accuracy of Operative Reports with Automated Artificial Intelligence Analysis of Surgical Video.
8. Dayan, D., Ortenzi, M., & Nizri, E. (2026). AI-based computer vision as an adjunct tool for CVS documentation in laparoscopic cholecystectomy.
9. Kewalramani, D., Jawa, R., Cuschieri, J., Kaafarani, H., Buchman, T. G., & Narayan, M. (2026). Navigating the artificial intelligence landscape in trauma, critical care and emergency general surgery: insights from the American Association for the Surgery of Trauma (AAST) 2025 Annual Meeting Panel Discussion.
10. Rimoli, E., Vashist, Y., & Marano, L. (2026). The Swiss Army Knife Illusion: Preserving the Human Dimension in the Age of AI.
11. Limon, D., Satish, V., Raghavan, N. R. S., Nguyen, P., & Rajesh, A. (2025). Artificial Intelligence in Surgery Revisited: A 2025 Guide to Understanding and Applying AI Models in Clinical Practice.
12. McLean, A. L. (2023). Artificial Intelligence in Surgical Documentation: A Critical Review of the Role of Large Language Models.
13. Zavaleta-Monestel, E., Arguedas-Chacón, S., & Cordero-Bermúdez, K. (2025). Artificial intelligence for surgical data management and decision support: lessons from wound care.
14. Decker, H., Trang, K., Ramirez, J. L., Colley, A., Pierce, L., Coleman, M. H., Bongiovanni, T., Melton, G. B., & Wick, E. C. (2023). Large Language Model–Based Chatbot vs Surgeon-Generated Informed Consent Documentation for Common Procedures.
15. Holland, A. M., Lorenz, W., Cavanagh, J. C., Smart, N., Ayuso, S. A., Scarola, G. T., Kercher, K. W., Jørgensen, L. N., Janis, J. E., Fischer, J. P., & Heniford, B. T. (2024). Comparison of Medical Research Abstracts Written by Surgical Trainees and Senior Surgeons or Generated by Large Language Models.
16. Sahhar, M., Nassar, J. E., Rouffiac, A.-E. J., Crow, K., Singh, M., Farias, M., Diebo, B. G., & Daniels, A. H. (2025). Enhanced Surgical Instructions Using Large Language Models.
17. Rosenberg, G. S., Magnéli, M., Barle, N., Kontakis, M. G., Müller, A. M., Wittauer, M., Gordon, M., & Brodén, C. (2024). ChatGPT-4 generates orthopedic discharge documents faster than humans maintaining comparable quality: a pilot study of 6 cases.
18. Vaziri, R., Nazari, I., Dorri, F., & Dorri, F. (2025). Generative Artificial Intelligence for Trauma and Fracture Surgery: Assessing the Efficacy of AI-Assisted Decision Support for Orthopedic Surgeons.
19. Bracken, A., Reilly, C., Feeley, A., Sheehan, E., Merghani, K., & Feeley, I. (2025). Artificial Intelligence (AI) – Powered Documentation Systems in Healthcare: A Systematic Review.
20. Stephen, P., Liang, A. S., Shah, S., Smith, M., Jeong, Y., Devon-Sand, A., Crowell, T., Delahaie, C., Hsia, C., Lin, S., Shanafelt, T. D., Pfeffer, M. A., Sharp, C., & García, P. (2024). Ambient artificial intelligence scribes: utilization and impact on documentation time.
21. Ong, C. S., Obey, N. T., Zheng, Y., Cohan, A., & Schneider, E. B. (2024). Reducing Surgeons' Routine Task Burden, Enhancing Surgical Decision Making, and Ensuring Patient Safety Using SurgeryLLM.
22. Padovano, M., Longo, F., Padovano, A., Frascio, M., & Nardo, B. (2025). Simulating AI-Mediated Clinical Notes for Surgeon–Specialist Communication: A Randomized Vignette Study.
23. Unadkat, K., Dhupati, P., & Noland, S. S. (2025). Ambient Artificial Intelligence in Hand Surgery: Insights, Pearls, and Pitfalls From Real-World Use.
24. Le, K. D. R., Tay, S., Choy, K. T., Verjans, J., Sasanelli, N., & Kong, J. C. (2024). Applications of natural language processing tools in the surgical journey.

25. Heslin, S. M. (2026). Artificial Intelligence documentation in trauma resuscitation: efficiency requires guardrails.
26. Adegbesan, A., Akingbola, A., Aremu, O., Adewole, O., Amamdikwa, J. C., & Shagaya, U. (2024). From Scalpels to Algorithms: The Risk of Dependence on Artificial Intelligence in Surgery.
27. Simone, B. D., Deeken, G., & Catena, F. (2025). Balancing Ethics and Innovation: Can Artificial Intelligence Safely Transform Emergency Surgery? A Narrative Perspective.
28. Morris, M. X., Fiocco, D., Caneva, T., Yiapanis, P., & Orgill, D. P. (2024). Current and future applications of artificial intelligence in surgery: implications for clinical practice and research.
29. Talwar, A. A., Talwar, A., Broach, R. B., Ungar, L., Hashimoto, D. A., & Fischer, J. P. (2025). Artificial intelligence unlocks the healthcare data lake.
30. Tinajero, C. A. C. (2025). The Pediatric Surgeon's AI Toolbox: How Large Language Models Like ChatGPT Are Simplifying Practice and Expanding Global Access.
31. Saadat, S., Darounkolaei, M. K., Qorbani, M., Hemmat, A., & Hariri, S. (2025). Enhancing Clinical Documentation with AI: Reducing Errors, Improving Interoperability, and Supporting Real-Time Note-Taking.
32. Loftus, T. J., Tighe, P., Filiberto, A. C., Efron, P. A., Brakenridge, S. C., Mohr, A. M., Rashidi, P., Upchurch, G. R., & Bihorac, A. (2019). Artificial Intelligence and Surgical Decision-making.