surgeries

MDPI

Review

Artificial Intelligence for Clinic Notes and Letters in Surgical Outpatient Practice: Evolution, Outcomes, Assessment Methods, and Future Directions

Ofir Ron [1,2,3], Warren Matthew Rozen [1,2,3,*], Peter Evans [1,2,3] and Manvir Singh [1,2,3]

[1] Peninsula Clinical School, Monash University, Melbourne, VIC 3199, Australia; ronofir3@gmail.com (O.R.); manvirsvm@gmail.com (M.S.)
[2] Department of Plastic and Reconstructive Surgery, Bayside Health, Frankston, VIC 3199, Australia
[3] Department of Surgery, Bayside Health, Melbourne, VIC 3199, Australia
* Correspondence: warrenrozen@hotmail.com

Abstract

Background/Objectives: Documentation burden is a major source of dissatisfaction among outpatient surgeons, and artificial intelligence (AI) tools—from general-purpose language models to purpose-built ambient scribes—have emerged as potential solutions. This narrative review examines their evolution, evaluates current evidence on effectiveness, and discusses assessment methods and future directions. **Methods:** Searches of MEDLINE, Scopus, Web of Science, and PubMed from inception to June 2026 identified 1843 records; following deduplication and PRISMA-guided screening, 20 studies met the eligibility criteria, supplemented by seven additional references identified via citation tracking, for 27 included sources. Findings were synthesised narratively across five domains: documentation efficiency, clinician wellbeing, patient experience, assessment methods, and implementation challenges. **Results:** Documentation efficiency showed the most consistent benefit, with reduced note-writing time reported across general, orthopaedic, and neurosurgical settings, though the most rigorous studies were derived from multispecialty and primary care cohorts rather than surgical practice. Wellbeing benefits were also frequently reported, but largely from uncontrolled designs and non-surgical populations. Documentation quality was more equivocal, varying by tool, specialty, and metric, with persistent concerns around factual inaccuracy, poor capture of informed consent, and unresolved questions of privacy, governance, and medicolegal liability. **Conclusions:** AI-assisted documentation shows a consistent efficiency benefit and probable wellbeing benefit, but surgery-specific evidence remains limited, documentation quality is tool- and specialty-dependent, and safety, governance, and equity concerns remain substantially unresolved. Future research should prioritise multicentre surgical trials with patient safety as a co-primary endpoint, specialty-tailored models, structured governance frameworks, and evaluation across linguistically diverse settings.

Keywords: ambient AI scribe; clinical documentation; large language models; surgical outpatient practice; clinician burnout

Check for updates

Academic Editor: Salomone Di Saverio

Received: 23 July 2026
Revised: 18 September 2026
Accepted: 22 September 2026
Published: 25 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Clinical documentation forms an integral component of surgical outpatient practice. Clinical notes and correspondence not only facilitate communication between surgeons, other medical professionals, and patients, but also fulfil critical administrative, medico-legal,

and billing requirements. However, documentation burden has increased substantially over recent decades following widespread adoption of electronic health records (EHRs). In particular, outpatient surgeons have been identified as one of the clinician groups most dissatisfied with EHR documentation requirements [1,2].

Furthermore, surgeons often write succinct notes that attempt to prioritise the most critical information, and prior research suggests this brevity comes at a cost. Previous qualitative analyses of postoperative clinical notes from 18 general surgeons found that 28% of notes failed to document any symptomatic change across their patients, and two thirds failed to document any functional changes [3].

Hence, these findings suggest opportunities for technologies that improve documentation quality and completeness while reducing burden. Recent advances in artificial intelligence (AI)—particularly large language models (LLMs), automatic speech recognition, and ambient AI scribes—have generated considerable interest as potential solutions to the aforementioned issues.

This narrative review asks whether the current evidence supports routine adoption of AI-assisted documentation in surgical outpatient practice, and if not, what would need to be established first. It traces how these tools have evolved, and whether general-purpose models and purpose-built ambient scribes differ enough in capability and failure mode to be treated separately. It then examines what the evidence shows on documentation efficiency, clinician wellbeing, patient experience, and quality; how well current assessment methods capture the failure modes that matter in surgical correspondence; and which questions of safety, governance, and equity remain unresolved.

2. Methods

This review was conducted as a structured narrative review with a systematic literature search, reported in accordance with an adapted PRISMA 2020 framework. An adapted PRISMA-style flow diagram is provided to summarise the literature-identification and selection process (Figure 1). Included studies are summarised by outcome domain in Tables 1–4.

2.1. Search Strategy

MEDLINE, Scopus, Web of Science, and PubMed were searched from inception to June 2026. The search strategy was based on an exploration of AI-assisted clinical documentation, ambient AI scribes, speech recognition systems, clinic letters, and surgical outpatient documentation. Search terms included combinations of the following: “ambient AI scribe”, “ambient artificial intelligence documentation”, “large language model clinical notes”, “AI-assisted clinical documentation”, “surgical outpatient documentation”, “clinic letters artificial intelligence”, “ChatGPT OR Copilot OR Bard OR Gemini”, “documentation burden burnout”, and “physician documentation quality”. The Boolean operators (AND, OR) were used to combine terms across concept domains. Search results were subsequently imported into Covidence for deduplication, followed by title and abstract screening, full-text review, and study selection. Additional studies were identified through citation tracking and manual reference screening of included studies and relevant reviews.

2.2. Eligibility Criteria

Studies were included if they (1) evaluated AI-assisted or ambient AI documentation tools in a clinical setting; (2) reported outcomes related to documentation efficiency, quality, clinician experience, or patient experience; (3) were published in English; and (4) were available as full-text publications. Both primary research (including randomised controlled trials, cohort studies, quality-improvement projects, and pilot studies) and review articles

were included where relevant to the evolution, effectiveness, assessment, or future direction of AI-assisted documentation in clinical practice. Priority was given to studies examining surgical or outpatient settings.

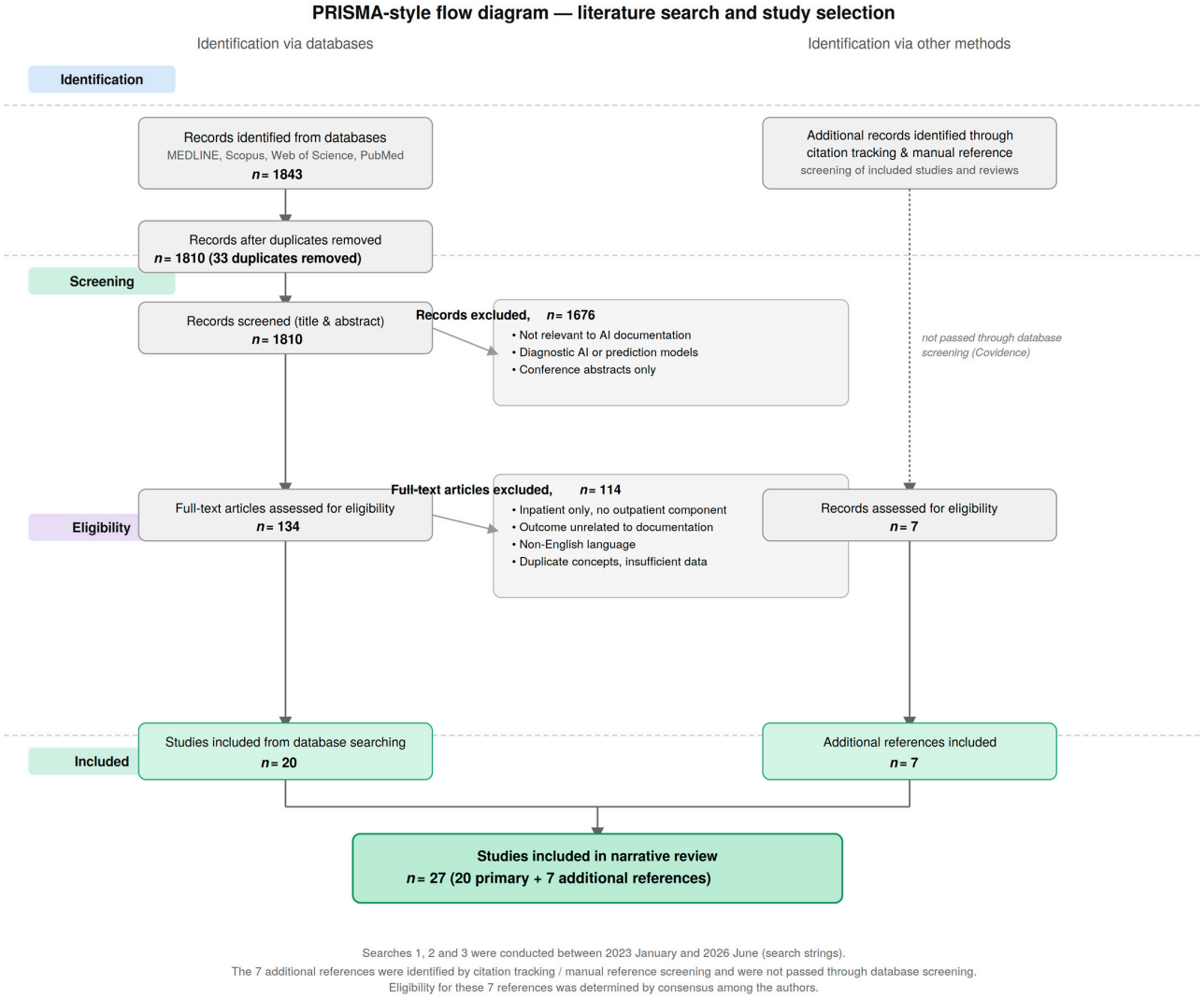


Figure 1. PRISMA-style flow diagram for literature search and study selection. Records identified via database searching (n = 1843) were screened and assessed for eligibility following the PRISMA 2020 process; a further 7 references identified through citation tracking and manual reference screening were assessed for eligibility by author consensus without passing through database screening. Twenty studies from database searching and 7 additional references were included (n = 27).

Studies were excluded if they focused solely on diagnostic AI applications without documentation relevance, examined inpatient-only settings with no outpatient component, or were available only as conference abstracts.

2.3. Study Selection

Two reviewers independently screened titles and abstracts using Covidence. Full-text screening was performed independently, with discrepancies resolved through discussion and, where necessary, adjudication by a senior author.

2.4. Data Extraction

A standardised extraction form was developed a priori and piloted before full extraction. Two reviewers extracted data independently, with discrepancies resolved by

consensus or senior adjudication. Extracted variables included study design, cohort characteristics, clinical setting, AI tool evaluated, primary outcomes, and key findings.

2.5. Data Synthesis

Due to the anticipated heterogeneity across individual study designs, AI tools, clinical settings, and outcome measures, the findings were synthesised narratively, organised by the various outcome domains: documentation efficiency, clinician wellbeing and burnout, patient experience, assessment methods, and implementation challenges. Outcomes were summarised using study-level findings and narrative interpretation. The decision to conduct a narrative rather than formal systematic synthesis reflects the evolving and heterogeneous nature of this field, where pooling of quantitative data across studies remains premature. Consistent with a narrative rather than systematic approach, formal risk-of-bias assessment using a standardised appraisal tool was not performed; the anticipated heterogeneity in study design, setting, and outcome measure rendered pooled quality scoring uninformative as a basis for evidence weighting. The PRISMA-style flow diagram and summary tables were used to guide and structure the narrative synthesis rather than to support formal meta-analytic inference.

Generative artificial intelligence (Claude Sonnet 4.6, Anthropic) was used to assist in generating the PRISMA-style flow diagram (Figure 1) and the summary tables of included studies (Tables 1–4); both outputs were subsequently edited, reviewed, and verified by the authors. GenAI was not used for study screening, data extraction, or narrative synthesis.

3. How Has AI-Assisted Documentation in Surgery Evolved over Time?

It is useful to distinguish the two broad categories of AI documentation tools, since the evidence base spans both and conflating them obscures important differences in capability and failure mode.

The first comprises general-purpose large language models (LLMs)—most prominently, GPT-3.5 and GPT-4 (OpenAI). While not being purpose-built for documentation, they can be prompted by a clinician to generate or restructure a note from supplied case information. Notably, studies of ChatGPT-assisted clinic letters have found meaningful intergenerational performance changes for this LLM: GPT-3.5 sometimes worsened readability metrics relative to clinician-written letters, whereas GPT-4 typically showed no significant difference, suggesting iterative model improvement carries real clinical consequences [4].

The second category comprises purpose-built ambient AI platforms—including DAX Copilot (Microsoft/Nuance), Lyrebird AI, Heidi AI, and Knowtex, among others—which are specifically designed to listen passively throughout a consultation and generate structured documentation without prompting, dictation, or manual data entry, integrating directly with EHR platforms via clinician- or specialty-specific templates. These represent a qualitatively different workflow integration to general-purpose chatbots retrofitted onto documentation tasks.

This progression is illustrated concretely within a single surgical specialty by two studies published two years apart. In 2024, Dubinski and colleagues compared a general-purpose large language model, ChatGPT, against an institution's existing dictation-based speech-recognition software for generating neurosurgical discharge summaries and operative reports [5]. In this earlier mode of use, ChatGPT did not listen to the clinical encounter directly—clinicians instead fed it case information to generate a draft, replacing a dictate-and-type workflow with a prompt-and-generate one. By 2026, Elmolla and colleagues evaluated a purpose-built ambient AI scribe, Lyrebird AI, across paediatric neurosurgery and multidisciplinary craniofacial outpatient clinics—software that passively

listens throughout the consultation and drafts a structured letter without prompting, dictation, or manual data entry [6].

Read together, these two neurosurgery-specific studies trace the shift from general-purpose chatbots retrofitted onto existing dictation workflows toward purpose-built, passively listening ambient scribes embedded directly in the clinical encounter. Ambient AI scribes in surgical outpatient practice, as in other outpatient specialties, therefore represent a natural progression toward workflow augmentation and administrative automation.

4. Evidence for Clinical and Operational Outcomes

4.1. Documentation Efficiency

Across eight studies (Table 1), documentation efficiency emerges as the most consistent benefit attributed to AI-assisted documentation, with reductions in clinician-facing documentation time reported in every cohort in which it was measured directly. The magnitude of benefit, however, varies by an order of magnitude—from a modest 20% reduction in a multispecialty quality-improvement cohort to near-elimination of documentation time in a research-built pipeline—reflecting differences in comparator, workflow, and what each study counted as documentation time.

Table 1. Studies reporting documentation efficiency outcomes.

Study	Population/Setting	Key Outcome	Main Finding
Duggan et al., 2025 [7]	46 clinicians; multispecialty outpatient	Time per note	20.4% reduction in time per note (10.3 to 8.2 min), with improved same-day encounter closure and reduced after-hours documentation.
Guo et al., 2025 [8]	Multispecialty academic health system	Note-writing time and note length	Note-writing time fell despite note length increasing, alongside reported reductions in documentation burden.
Alboksmaty et al., 2025 [9]	Systematic review of 9 studies; primary care and outpatient	Documentation burden and workflow	Consistent reductions in documentation burden and improved workflow efficiency; clinicians reported greater capacity to focus on patients.
Harvey et al., 2025 [10]	3 surgeons (general, orthopaedic, neurosurgical); outpatient clinics	EHR audit metrics	Effects on objective EHR metrics varied in direction by surgeon: after-hours time fell for the neurosurgeon but rose for the general and orthopaedic surgeons.
Carmichael et al., 2026 [11]	119 consecutive general and colorectal surgical consultations; regional Victoria, Australia	Clinician-facing documentation time	Median documentation time 0.71 vs. 2.28 min with AI ($p < 0.001$); regression-adjusted reduction 1.35 min (95% CI −1.75 to −0.95). Consultation duration unchanged; editing time longer with AI (0.71 vs. 0.39 min, $p < 0.001$).
Etagiuri et al., 2026 [12]	279 patients; single-surgeon arthroplasty practice (RCT)	Note-completion and staff processing time	Surgeon note-completion time 11.76 vs. 13.90 min ($p < 0.001$); office staff processing time 3.46 vs. 4.71 min ($p < 0.001$).
Rodenhouse et al., 2026 [13]	19 orthopaedic providers; outpatient clinics	EHR efficiency metrics	Time on notes per appointment fell 41%, per day 31%, and after-hours EHR time 61%; gains scaled with utilisation rate.
Gasque et al., 2025 [14]	20 hepatobiliary/pancreatic outpatient consultations; Spanish language	Documentation time	Median documentation time fell from 15.98 to 1.08 min using a Whisper and ChatGPT pipeline.

Two interpretive caveats temper this apparent consistency. First, the largest and most methodologically robust efficiency studies are derived from multispecialty and primary care settings rather than surgical practice; of the surgical studies, only the arthroplasty randomised trial and the regional general surgery pilot include a comparison group and a sample size permitting meaningful inference. Findings from general outpatient populations should therefore be read as indicating a plausible direction of effect rather than as established surgical evidence, since surgical consultations differ in structure, documentation content, and baseline templating practices.

Second, efficiency gains appear to redistribute clinician effort rather than simply remove it. The regional general surgery pilot found that although total documentation time fell substantially, editing time nearly doubled, reflecting a shift from composing text during dictation to reviewing a pre-generated draft [11]. The surgical outpatient pilot reinforces this heterogeneity: after-hours documentation time moved in opposite directions for different surgeons within the same institution using the same tool [10]. Taken together, these findings suggest that baseline documentation habits—including existing reliance on templates or brief structured notes—may determine benefit at least as much as specialty does.

4.2. Clinician Wellbeing and Cognitive Workload

The direction of effect across these studies (Table 2) is consistent: reductions in documentation burden are accompanied by improvements in perceived workload, satisfaction, and burnout scores. The randomised crossover trial provides the strongest design in this domain and adds an important finding—that the magnitude of benefit differs meaningfully between commercial platforms, with only one of two tools producing a significant reduction in time-in-note despite both improving burnout scores [15]. This suggests that "ambient AI" should not be treated as a single intervention when interpreting wellbeing outcomes.

Table 2. Studies reporting clinician wellbeing and cognitive workload outcomes.

Study	Population/Setting	Key Outcome	Main Finding
Harvey et al., 2025 [10]	3 surgeons; surgical outpatient clinics	Self-reported burnout symptoms	Self-reported burnout symptoms declined across participating surgeons, with non-significant reductions in mental demand and perceived rush.
Chowdhury et al., 2026 [15]	Multispecialty outpatient physicians; randomised crossover trial comparing DAX v2.0 and Nabla v1.5	Burnout (Mini-Z) and time-in-note	Nabla reduced time-in-note 9.5% vs. control ($p = 0.02$); DAX showed no significant change. Both platforms improved Mini-Z burnout scores and reduced task load and work exhaustion ($p < 0.001$).
Stults et al., 2025 [16]	Multispecialty clinicians; Sutter Health	Satisfaction and cognitive load	Improved clinician satisfaction at work and decreased cognitive load, alongside reduced time on notes per appointment.
Furrukh et al., 2025 [17]	6 plastic surgery residents (before–after pilot)	Perceived workload (NASA-TLX) and burnout (Mini-Z)	NASA-TLX fell from 74.64 to 51.94 ($p = 0.04$), and after-hours documentation from 3.6 to 0.75 h/week ($p = 0.02$); Mini-Z burnout change was non-significant.

These findings must nonetheless be interpreted cautiously. Most studies in this domain are quality-improvement evaluations or uncontrolled before–after designs, which are subject to selection bias—clinicians who adopt AI scribes voluntarily may differ systemati-

cally from those who do not—and the absence of control groups limits causal inference. The possibility that observed improvements reflect novelty effects, Hawthorne bias, or concurrent workflow changes cannot be excluded on the basis of current evidence.

The surgery-specific evidence is thinner still. The only dedicated surgical wellbeing study identified is a six-participant resident pilot in which the burnout instrument itself did not reach significance, and the supporting surgical data came from a three-surgeon outpatient pilot [10,17]. Any inference that ambient AI reduces burnout among practising surgeons therefore rests substantially on extrapolation from non-surgical populations rather than on direct evidence.

4.3. Patient Experience

Patient experience is the least studied outcome domain in the literature, and only two included studies report it. A retrospective pilot across outpatient clinics at an academic medical centre—including orthopaedic surgery and otolaryngology—found increases in Press-Ganey patient experience scores associated with provider use of an ambient scribe [18]. In the only surgical outpatient study to report patient-facing outcomes directly, surgeon-reported ability to give patients undivided attention improved significantly, but patient-reported satisfaction showed only small, non-significant changes that varied in direction by specialty [10].

The consistent signal across the literature is therefore clinician-perceived rather than patient-reported: clinicians report greater capacity to attend to patients when documentation is delegated [9,16], while patient satisfaction generally remains stable rather than improving measurably. This supports a conservative interpretation—that documentation automation does not appear to come at the expense of the consultation—rather than a positive claim that it improves patient experience. No study identified evaluated patient experience in surgical practice as a primary endpoint, and none assessed whether patients' awareness of being recorded influences what they disclose.

5. How Are These Outcomes Assessed?

Assessment of AI-assisted documentation systems has evolved considerably and now encompasses four main domains: workflow efficiency, documentation quality, clinician experience, and implementation outcomes. Workflow measures remain the most commonly reported and include documentation time, time spent on notes, after-hours charting ("pajama time"), encounter closure rates, same-day note completion, and total EHR interaction time. These are summarised in Table 1 above. The sections below address the assessment of documentation quality and the emerging methodological literature on how such assessment should be conducted.

5.1. Documentation Quality

Documentation quality has largely been assessed using validated instruments (Table 3)—including the Sheffield Assessment Instrument for Letters (SAIL) and the Physician Documentation Quality Instrument (PDQI-9)—alongside readability metrics and clinician preference ratings. Readability is measured using formulaic indices—including the Flesch–Kincaid Grade Level, Flesch Reading Ease, SMOG Index, and Gunning Fog Index—which estimate the educational level required to understand a given text. Blinded clinician preference ratings provide a further subjective measure of overall letter quality.

Table 3. Studies reporting documentation quality outcomes.

Study	Population/Setting	Assessment Instrument	Main Finding
Stoneham et al., 2024 [4]	Hand surgery outpatient clinic letters	Flesch Reading Ease; Flesch–Kincaid Grade Level	GPT-3.5 significantly worsened readability relative to clinician-written letters; GPT-4.0 showed no significant difference. Both increased letter length.
Elmolla et al., 2026 [6]	20 paediatric neurosurgery and craniofacial outpatient consultations	Readability indices; blinded rater preference	Improved readability on two of four indices; blinded rater preferred AI letters in 75% of cases; no unsafe or incorrect content identified, but qualifying diagnostic nuance was sometimes lost.
Etagiuri et al., 2026 [12]	Arthroplasty outpatient practice (RCT)	Blinded six-dimension quality rating	AI-generated letters rated significantly higher across all six quality dimensions assessed.
Patel et al., 2025 [19]	Urology outpatient clinic letters; major London hospital	SAIL	Significantly higher global SAIL quality scores (7.5 vs. 6.0), with improved readability and content validity.
Palm et al., 2025 [20]	5 specialties including orthopaedics; AI-generated vs. physician-authored notes	PDQI-9	Physician-authored notes scored higher overall and on accuracy, succinctness and internal consistency; AI notes scored higher on thoroughness and organisation.
Reddy et al., 2026 [21]	Veterans' Health Administration; 11 ambient AI vendors, simulated encounters	Modified PDQI-9	Human-generated notes scored higher across all ten domains assessed, with the largest gaps in cases requiring complex clinical reasoning.

Documentation quality is the domain in which findings diverge most sharply, and this divergence is systematic rather than random. Results depend on which tool is evaluated, which version of that tool, which instrument is applied, and how complex the clinical material is. The two PDQI-9 comparisons illustrate this directly: one found AI notes outperforming physician-authored notes on thoroughness and organisation while underperforming on accuracy, whereas a larger multi-vendor evaluation found human notes superior across every domain assessed, with the gap widening in cases requiring complex reasoning [20,21]. Any claim that AI matches or exceeds physician documentation quality must therefore be qualified by domain, vendor, and case complexity rather than stated generally.

The same pattern holds within surgery. Readability improved in paediatric neurosurgery and letter quality improved in arthroplasty, yet in hand surgery, an earlier model generation actively worsened readability [4,6,12]. Model version alone accounts for part of this spread, which cautions against generalising from studies using superseded models.

A distinction that recurs in the surgical literature is between outright inaccuracy and subtler loss of clinical nuance. The paediatric neurosurgery evaluation identified no unsafe or incorrect content, but found that clinician-dictated letters more reliably retained qualifying language—describing a finding as an incidental one, for example, rather than stating it as unqualified [6]. In surgical documentation, such qualifiers carry diagnostic and prognostic weight, and their loss would not be detected by instruments designed to identify factual error. This suggests current quality instruments may be insufficiently sensitive to the failure modes that matter most in surgical correspondence.

Most existing studies evaluate a single tool in a single subspecialty with small sample sizes. Generalisability across surgical subspecialties therefore cannot be assumed, and

future validation work should treat surgical documentation as heterogeneous rather than as a single category.

5.2. Assessment Methods and Evaluation Frameworks

Assessment methodology is itself becoming a subject of study (Table 4), and three distinct approaches are represented here. Transcription accuracy can be quantified independently of downstream note quality using established speech-recognition metrics—an important separation, since errors introduced at the transcription stage propagate into any note generated from that transcript [14].

Table 4. Studies evaluating assessment methodology for AI-generated documentation.

Study	Population/Setting	Assessment Approach	Main Finding
Gasque et al., 2025 [14]	20 hepatobiliary outpatient consultations; Spanish language	ASR metrics (WER, MER, word/character information measures)	WER 0.07 and CER 0.04—slightly above published Spanish-language benchmarks but within acceptable range; 85% of generated notes rated satisfactory.
Guo et al., 2026 [22]	Clinician edits to ambient AI drafts across 33 specialties	Qualitative content analysis of edits	Clinical content edits predominated—orders, tests and procedures (39.9%), symptom refinements (30.3%) and medication changes—with comparatively few grammar or style edits.
Wang et al., 2025 [23]	40 real clinical recordings; multi-vendor	SCRIBE multi-method framework (human review, automated metrics, LLM-based evaluation, adversarial simulation)	Standard automated metrics correlated poorly with human clinical judgement ($r < 0.2$); domain-adapted metrics performed substantially better (coherence $r = 0.51$; clarity $r = 0.69$). Simulation testing found failure to flag implausible laboratory values in 60% of cases.

Clinician editing behaviour offers a second, indirect measure of reliability. The finding that edits concentrate on clinical content rather than grammar or style suggests that AI systems reliably produce well-structured documentation but require substantive clinical oversight [22]. Because omission or inclusion of specific symptoms and orders can plausibly alter downstream management, this oversight is more than a formality.

The most significant methodological development is the recognition that commonly used automated metrics correlate poorly with human clinical judgement, with correlation coefficients below 0.2 across several quality dimensions [23]. This finding has direct implications for how the studies in Table 3 should be read, and it motivated the SCRIBE framework—a structured multi-method approach integrating human expert review, automated metrics, LLM-based evaluation, and adversarial simulation testing, on the basis that no single method adequately captures all dimensions of performance. Adversarial testing within that framework identified failure modes of particular relevance to surgery, including failure to recognise newly introduced medications and retention of implausible values without flagging. SCRIBE represents the most systematic evaluation approach proposed to date, and signals that the field is beginning to formalise assessment—a necessary precondition for meaningful standardisation across surgical implementations.

6. Discussion

This review suggests that the evaluation of AI-assisted documentation is undergoing a conceptual shift—from a narrow focus on efficiency toward multidimensional assessment encompassing documentation quality, clinician editing behaviour, patient experience, implementation outcomes, and safety. Taken together, the evidence reviewed here points

toward a consistent direction of efficiency benefit, a probable but less well-evidenced well-being benefit, a more nuanced and tool-dependent picture on quality, and a set of safety, governance, and equity concerns that remain substantially unresolved.

6.1. Composition of the Evidence Base

An important interpretive caveat applies across all outcome domains considered here. Although this review is framed around surgical outpatient practice, a substantial proportion of the available evidence is derived from multispecialty, primary care, or mixed outpatient cohorts. Where surgery-specific studies do exist, they are predominantly small single-centre pilots, simulation studies, or single-surgeon series; only two—a randomised trial in arthroplasty [12] and a quasi-experimental pilot in regional general surgery [11]—include a comparison group and a sample size permitting meaningful statistical inference. Several outcome domains, notably patient experience and clinician wellbeing, rest almost entirely on non-surgical evidence.

This matters because surgical outpatient documentation differs materially from the settings in which most of this evidence was generated. Surgical clinic letters carry procedural and anatomical detail, operative planning information, and consent discussions that have no direct analogue in general medical consultations. Findings from general outpatient populations should therefore be read as indicative of a plausible direction of effect rather than as established surgical evidence, and the conclusions drawn in this review are qualified accordingly.

6.2. Accuracy, Hallucination, and Surgical Risk

The most significant safety concern involves factual inaccuracies and hallucinations. While modern LLM-based scribes report lower error rates (1–3%) than dictation-based ASR systems (7–11%), they introduce distinct failure modes—fabrication, omission, speaker misattribution, and contextual misinterpretation [24]. Adoption appears to be outpacing rigorous safety validation, with most existing evidence focused on efficiency and satisfaction rather than systematic safety assessment [24].

Three findings from the studies reviewed above converge on a pattern of particular relevance to surgery. First, accuracy degrades as case complexity increases: in the neurosurgical evaluation, factual correctness was substantially lower for craniotomy and tumour resection operative reports than for more routine cases, which the authors attributed to the comparative rarity of such cases in the underlying training data [5]. This is precisely the direction surgical documentation moves in as procedures become more complex or atypical. Second, informed consent and risk–benefit discussions were captured poorly in surgical outpatient practice, producing short, generic summaries of conversations that carry direct medicolegal weight [10]. This is a distinct failure mode from hallucination, and one specific to the consent-heavy nature of surgical consultation. Third, the only surgical randomised trial recorded more than double the medical errors in the AI arm compared with conventional documentation, a difference that did not reach statistical significance but is clinically difficult to dismiss given the sample size [12].

Read together, these findings indicate that the surgical contexts in which AI documentation is most likely to fail are those in which documentation errors carry the greatest consequence. The recommendation for continued full clinician review of AI-generated documentation before distribution, made by the trial authors, should be extended broadly given the absence of larger confirmatory studies.

6.3. Privacy, Confidentiality, and Governance

Ambient AI systems record and process sensitive patient–clinician conversations, creating obligations under multiple regulatory frameworks. In Australia, collection and

processing must comply with the Privacy Act 1988 (Cth), the Australian Privacy Principles, and, in Victoria, the Health Records Act 2001. In the European Union, the EU AI Act (2024) classifies AI systems used in clinical decision support as high-risk, imposing conformity assessment and post-market monitoring requirements that may extend to ambient documentation tools. The Victorian Government's guidance on generative AI in the public sector goes further, explicitly prohibiting entry of patient health data into publicly available AI tools without authorisation—guidance that has already restricted surgical AI research in this state to private settings [11,25].

Questions of data residency, vendor access to identifiable patient data, and medicolegal responsibility in the event of a documentation error remain unresolved across all jurisdictions. The framework below (Table 5) translates these obligations into concrete institutional requirements, synthesised from existing regulatory instruments and offered as a minimum standard rather than a validated protocol.

Table 5. Institutional governance framework for ambient AI documentation tools.

Phase	Requirement	Why It Matters
Pre-implementation	Privacy impact assessment; vendor contract reviewed for data residency, sub-processor disclosure, retention periods, model-training use of clinical audio, and breach notification; determination of whether the tool falls within software-as-a-medical-device regulation	Offshore processing and clinical-suggestion features attract obligations that pure transcription tools do not, and these must be established before deployment rather than after
Patient notification and consent	Consent obtained and documented at the point of consultation, not via a waiting-room notice or admission-form clause; plain-language explanation of what is recorded, how long audio is retained, and whether any third party can access it; declining must not affect care, and the workflow must function without the tool rather than around it	Paediatric, interpreter-mediated, and impaired-capacity consultations each require a defined pathway rather than default application of the adult self-consenting one
Data lifecycle	Audio deleted once the note is finalised unless a specific documented reason to retain exists; retention periods for transcripts and drafts defined in advance and enforced technically; access to raw audio logged and restricted; de-identification standards specified for any retained recordings	Retention governed by policy alone is not enforceable; technical enforcement is what actually protects the data
Accountability	Signing clinician remains responsible, stated explicitly in local policy; defined standard for adequate review before a note enters the record; audit trail distinguishing AI-generated from clinician-authored content; reporting pathway for tool-attributable errors	Responsibility left to inference is the gap most likely to surface in medicolegal dispute
Post-implementation monitoring	Ongoing surveillance rather than a one-off procurement decision: periodic audit of note accuracy against source encounters, monitoring of consent-decline rates, and review of reported documentation errors	Mirrors the post-market monitoring logic already applied to high-risk AI systems under the EU AI Act

6.4. Strength of the Evidence

Current evidence remains dominated by observational and quality-improvement studies. Only two randomised trials were identified for this review—one surgical [12] and one multi-specialty [15]—and surgery-specific validation frameworks are essentially absent. Sample sizes in the surgical literature are generally small, several studies are single-surgeon or single-centre, and simulated rather than real-world settings account for a meaningful proportion of the evidence. These limitations constrain the strength of any causal claim about the effect of AI-assisted documentation on documentation quality, clinician wellbeing, or patient outcomes in surgical practice.

6.5. Patient Safety as an Evaluation Endpoint

Patient safety is largely absent from the literature as a measured outcome. Of the studies reviewed, only the arthroplasty randomised trial reported documentation-related errors as a discrete endpoint, and as described above, it did so as a secondary measure within a trial powered for documentation time [12]. This is the central methodological problem in the field: efficiency endpoints are inexpensive to measure and yield significant results in small samples, whereas safety endpoints are comparatively rare events requiring sample sizes an order of magnitude larger. The result is a body of literature that appears reassuring on safety largely because it has not looked.

Future evaluations should treat safety as a co-primary rather than a secondary endpoint, and should specify indicators appropriate to surgical documentation rather than adopting generic note-quality measures. Candidate indicators include discrepancies between medications, doses, or allergies recorded in the note and those actually prescribed; omission or misstatement of laterality, procedure, or anatomical site; failure to document specified elements of the informed consent discussion; loss of diagnostic qualifiers of the kind identified in paediatric neurosurgery [6]; incorrect or omitted follow-up intervals and investigation orders; and downstream events attributable to a documentation error, such as a missed result, an incorrect booking, or a communication failure with the referring clinician. Several of these are countable from routine records without additional data collection.

Two design implications follow. First, individual centres are unlikely to accrue sufficient events, so multicentre collaboration or a shared registry of documentation-related incidents will be necessary for meaningful safety inference. Second, error rates should be reported against a clinician-authored comparator rather than against zero, since conventional dictation and templating also generate errors; the relevant question is whether AI-assisted documentation changes the rate or the type of error, and current evidence answers neither. Until trials are designed to answer them, recommendations for routine adoption in surgical practice rest on efficiency data alone, and the urgency of generating this evidence increases as adoption accelerates ahead of it.

7. How Is AI-Assisted Documentation Continuing to Evolve?

Future development is likely to focus on more sophisticated platforms embedded directly in surgical workflows, including automated generation of referral responses, postoperative letters, multidisciplinary summaries, and patient education materials, with progressively less clinician editing required as systems mature.

Specialty-specific models represent a further pressing direction. A 2026 narrative review of ambient AI scribe technology found that satisfaction with AI documentation tools varied substantially across clinical disciplines, with primary care clinicians reporting the highest satisfaction (85%), compared with only 50% of surgical subspecialists and 36.4% of medical subspecialists—differences attributed primarily to poor specialty-specific customisation, particularly in physical examination sections requiring manual revision [26]. Documentation performance is therefore likely to improve when models are tailored to a domain's terminology and conventions, which could particularly benefit orthopaedics, plastic surgery, neurosurgery, and urology given their reliance on precise procedural and anatomical language.

However, several technical and ethical hurdles must be addressed before these developments can be realised in routine surgical practice. Data interoperability remains a significant barrier: many surgical centres rely on legacy EHR systems with proprietary architectures that do not natively support real-time API integration with ambient AI platforms, requiring costly middleware solutions or full system replacement [26]. Liability

frameworks for AI-driven documentation errors are also largely undefined—it is currently unclear whether responsibility for an AI-generated inaccuracy that influences clinical management rests with the clinician who signed the note, the institution that deployed the tool, or the AI vendor, and existing medical indemnity arrangements do not consistently address this [24,27].

7.1. Training and Competency for Surgical Trainees

Trainees bear a disproportionate share of the documentation burden and are likely to be among the earliest and heaviest users of these tools, yet no structured competency framework for AI-assisted documentation currently exists in surgical curricula. At minimum, such a framework should cover the distinctive failure modes of AI-generated notes (fabrication, omission, speaker misattribution, and loss of qualifying language, which are less conspicuous than familiar dictation errors [6,24]); a defined pre-signing verification workflow for the elements that matter most—medications, allergies, laterality, site, consent content, and follow-up; the medicolegal weight of attestation, given that responsibility rests with the signing clinician however the note was generated [27]; and how to obtain consent for recording and proceed when a patient declines.

Another concern specific to training also warrants study. Composing a note requires the trainee to synthesise the encounter, commit to an impression, and articulate a plan—part of how clinical reasoning develops and how supervisors assess it. Delegating that synthesis to an ambient system early in training may reduce burden at the cost of a formative activity, yet no study has examined the effect of ambient documentation on trainee learning. Until one does, supervised review of AI-generated notes is better treated as a workplace-based assessment activity than as an efficiency measure, with competency in unassisted documentation established first. Without competency frameworks of this kind, the risk of over-reliance and reduced documentation vigilance may offset efficiency gains.

7.2. Linguistic and Cultural Diversity

An important and underexplored consideration for future research is the performance of ambient AI systems across diverse linguistic backgrounds. To date, the vast majority of published studies have been conducted in English-dominant settings with homogeneous clinician and patient populations. A notable exception is the HPB SmartNotes study by Gasque and colleagues, which generated SOAP notes from Spanish-language surgical consultations using a Whisper and ChatGPT pipeline; the authors observed a slightly higher word error rate in Spanish than published benchmarks would predict, indicating scope for improvement [14]. That a purpose-built pipeline underperformed its English-language benchmark in the one non-English surgical evaluation available is not a reassuring baseline from which to generalise.

Evaluating this properly requires more than repeating existing study designs in other languages. Four methodological features would strengthen such work. First, performance should be reported disaggregated by language of consultation and by speaker background—including clinicians and patients using English as a second language—rather than pooled, with pre-specified equivalence margins against English-language benchmarks so that non-inferiority is tested rather than assumed. Second, interpreter-mediated consultations should be evaluated as a distinct category rather than excluded: three-party audio, overlapping speech, and the need to attribute content correctly between patient, interpreter, and clinician present failure modes that single-language evaluation cannot detect, and these consultations are common in surgical outpatient practice in multicultural health systems. Third, clinician editing burden should be reported as an equity outcome in its own right—editing minutes per note, stratified by language group—since a tool that achieves comparable final note

quality only by requiring substantially more correction for accented speech distributes its benefit unequally even where accuracy metrics appear acceptable. Fourth, adversarial and simulation testing of the kind proposed within the SCRIBE framework [23] should include accented and code-switched speech, which is more tractable than assembling large real-world cohorts and can be conducted before deployment rather than after.

Recruitment should reflect the linguistic composition of the catchment population rather than the convenience of the research setting. In Australian surgical outpatient practice, where a substantial proportion of patients speak a language other than English at home, a validation study drawing only on English-speaking participants does not establish that a tool is safe for the population it will be used on.

7.3. Research Priorities

Priorities for future research include multicentre prospective studies within surgical outpatient settings, surgery-validated assessment frameworks incorporating the patient safety endpoints described in Section 6.5, cost-effectiveness analyses, and longer-term evaluation of whether efficiency gains translate into sustained burnout reduction [26]. Looking further ahead, ambient scribes are likely to evolve beyond passive documentation into proactive clinical support—what has been described as the "AI-augmented clinician"—where the AI not only records the encounter but assists with ordering, referral workflows, and real-time contextual reasoning [26]. That trajectory makes the governance, safety, and training questions raised above more pressing rather than less, since each expansion of function widens the gap between what these systems do and what has been formally evaluated.

8. Conclusions

AI-assisted documentation has evolved rapidly from transcription support to sophisticated ambient platforms generating clinic notes and correspondence directly from patient–clinician interactions. The evidence reviewed reflects a field in transition, in which evaluation metrics are shifting from an initially narrow focus on efficiency toward multidimensional assessment encompassing documentation quality, clinician editing behaviour, patient experience, and safety. Evidence spanning urology, neurosurgery, orthopaedics, plastic surgery, general surgery, and hepatobiliary surgery indicates a consistent direction of efficiency benefit and a probable wellbeing benefit, with a more equivocal picture on documentation quality that depends heavily on specialty, tool, and metric. Much of this evidence, however, is derived from multispecialty and primary care cohorts, and surgery-specific studies remain small and largely uncontrolled.

The consent-documentation gap identified in surgical outpatient practice represents a distinct failure mode not captured by efficiency metrics alone, and the observation that AI accuracy degrades as case complexity increases is of particular concern in a specialty where documentation complexity and clinical consequence rise together. Patient safety has scarcely been measured as an endpoint in its own right, and challenges around accuracy, hallucination, governance, data privacy, trainee competency, and the absence of surgery-specific validation frameworks remain important barriers. Future work should prioritise multicentre surgical outpatient trials powered for safety as well as efficiency, specialty-tailored models, structured governance and consent frameworks, and evaluation across the diverse linguistic and clinical contexts in which surgeons practice—including regional and non-English-dominant settings where current evidence is almost entirely absent.

Author Contributions: Conceptualisation, M.S. and P.E.; methodology, M.S. and O.R.; investigation, M.S., O.R. and W.M.R.; data curation, M.S.; writing—original draft preparation, M.S.; writing—

review and editing, O.R., W.M.R. and P.E.; visualisation, M.S.; supervision, P.E. and W.M.R.; project administration, P.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This narrative review did not involve human participants, animal subjects, or the collection of primary data, and therefore did not require ethical approval.

Informed Consent Statement: Not applicable. This study did not involve human subjects.

Data Availability Statement: No new data were created or analysed in this study. Data sharing is not applicable to this article.

Acknowledgments: During the preparation of this manuscript, the authors used Claude Sonnet 4.6 (Anthropic) for the purposes of generating the PRISMA-style flow diagram (Figure 1) and the summary tables of the included studies (Tables 1–4). The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gaffney, A.; Woolhandler, S.; Cai, C.; Bor, D.; Himmelstein, J.; McCormick, D.; Himmelstein, D.U. Medical documentation burden among US office-based physicians in 2019: A national study. *JAMA Intern. Med.* **2022**, *182*, 564–566. [CrossRef] [PubMed]
2. Sinsky, C.; Colligan, L.; Li, L.; Prgomet, M.; Reynolds, S.; Goeders, L.; Westbrook, J.; Tutty, M.; Blike, G. Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties. *Ann. Intern. Med.* **2016**, *165*, 753–760. [CrossRef] [PubMed]
3. Miller, J.M.; Velanovich, V. The natural language of the surgeon's clinical note in outcomes assessment: A qualitative analysis of the medical record. *Am. J. Surg.* **2010**, *199*, 817–822. [CrossRef] [PubMed]
4. Stoneham, A.C.; Walker, L.C.; Newman, M.J.; Nicholls, A.; Avis, D. Can artificial intelligence make elective hand clinic letters easier for patients to understand? *J. Hand Surg. Eur. Vol.* **2024**, *49*, 1269–1270. [CrossRef] [PubMed]
5. Dubinski, D.; Won, S.Y.; Trnovec, S.; Behmanesh, B.; Baumgarten, P.; Dinc, N.; Konczalla, J.; Chan, A.; Bernstock, J.D.; Freiman, T.M.; et al. Leveraging artificial intelligence in neurosurgery—Unveiling ChatGPT for neurosurgical discharge summaries and operative reports. *Acta Neurochir.* **2024**, *166*, 38. [CrossRef] [PubMed]
6. Elmolla, M.; Khanom, A.A.; Ali, A.M.S.; Duncan, C.; Hennedige, A.; Vakharia, V.N. Integration of AI-generated clinic letters in complex paediatric neurosurgery outpatient settings. *Childs Nerv. Syst.* **2026**, *42*, 80. [CrossRef] [PubMed]
7. Duggan, M.J.; Gervase, J.; Schoenbaum, A.; Hanson, W.; Howell, J.T.; Sheinberg, M.; Johnson, K.B. Clinician experiences with ambient scribe technology to assist with documentation burden and efficiency. *JAMA Netw. Open* **2025**, *8*, e2460637. [CrossRef] [PubMed]
8. Guo, Y.; Wang, J.; Hu, D.; Tam, S.; Gilman, C.; Chow, E.; Perret, D.; Pandita, D.; Zheng, K. Evaluating ambient artificial intelligence documentation: Effects on work efficiency, documentation burden, and patient-centered care. *J. Am. Med. Inform. Assoc.* **2025**, *32*, 273–282. [CrossRef] [PubMed]
9. Alboksmaty, A.; Aldakhil, R.; Hayhoe, B.W.J.; Ashrafian, H.; Darzi, A.; Neves, A.L. The impact of using AI-powered voice-to-text technology for clinical documentation on quality of care in primary care and outpatient settings: A systematic review. *eBioMedicine* **2025**, *118*, 105861. [CrossRef] [PubMed]
10. Harvey, C.J.; Wong, V.; Huynh, W.; Lee, J.P.; Woo, R.K. Ambient AI-assisted clinical documentation in surgical outpatient care: A preliminary study of usability, workflow, and patient experience. *World J. Pediatr. Surg.* **2025**, *8*, e001073. [CrossRef] [PubMed]
11. Carmichael, G.J.; Di Tano, V.; Wright, C.; Al-Sharea, A.; Kovoor, J.G.; Gonzalez-Chica, D.; Jacob, M.O. Artificial intelligence-enhanced dictation improves efficiency in a regional general surgery clinic: A quasi-experimental pilot study. *ANZ J. Surg.* 2026; *online ahead of print*. [CrossRef] [PubMed]
12. Etagiuri, A.; Chaudhry, D.; van der Merwe, J.; Nickol, M.; van der Merwe, J.M. Evaluating the impact of artificial intelligence scribes on clinical workflow and documentation quality: A randomized controlled trial. *J. Arthroplast.* **2026**. [CrossRef] [PubMed]
13. Rodenhouse, A.J.; Davis, E.; Burger, P.; Nicandri, G.; DiGiovanni, B. Utilization of an artificial intelligence-based documentation system improves provider efficiency in outpatient orthopaedic clinics: Reducing the afterhours burden of the electronic health record—A pilot study. *J. Am. Acad. Orthop. Surg.* **2026**, *34*, e1093–e1101. [CrossRef] [PubMed]
14. Gasque, R.A.; Zaietta, N.; Mollard, L.; Chahdi Beltrame, M.; Lenz Virreira, M.E.; Quiñonez, E.G.; Mattera, F.J. HPB SmartNotes: The impact of artificial intelligence on surgeon workload in the outpatient office. *EngMedicine* **2025**, *2*, 100101. [CrossRef]

15. Chowdhury, A.; Casey, M.; Wilson, J.; Pollak, K.I.; Goldstein, B.A.; Bedoya, A.; Poon, E.G. Comparing ambient scribes: A randomized crossover clinical trial addressing ambient scribe technologies' impact on physician burnout. *J. Am. Med. Inform. Assoc.* **2026**, *33*, 990–999. [CrossRef] [PubMed]
16. Stults, C.D.; Deng, S.; Martinez, M.C.; Wilcox, J.; Szwerinski, N.; Chen, K.H.; Driscoll, S.; Washburn, J.; Jones, V.G. Evaluation of an ambient artificial intelligence documentation platform for clinicians. *JAMA Netw. Open* **2025**, *8*, e258614. [CrossRef] [PubMed]
17. Furrukh, A.J.; Lister, R.C.; Lalikos, J.F. Implementation of Dragon Ambient Experience by plastic surgery residents improves wellness and decreases burnout: A pilot study. *Plast. Reconstr. Surg. Glob. Open* **2025**, *13*, e7287. [CrossRef] [PubMed]
18. Davis, E.; Davis, S.; Haralambides, K.; Gleber, C.; Nicandri, G. Ambient AI documentation and patient satisfaction in outpatient care: Retrospective pilot study. *JMIR AI* **2026**, *5*, e78830. [CrossRef] [PubMed]
19. Patel, T.; Khan, N.; Huntley, J.; Raison, N. From pen to portal: Using artificial intelligence to evaluate the impact of electronic patient information and communication (EPIC) on the quality of urology clinic letters at a major London hospital. *Cureus* **2025**, *17*, e97486. [CrossRef] [PubMed]
20. Palm, E.; Manikantan, A.; Mahal, H.; Belwadi, S.S.; Pepin, M.E. Assessing the quality of AI-generated clinical notes: Validated evaluation of a large language model ambient scribe. *Front. Artif. Intell.* **2025**, *8*, 1691499. [CrossRef] [PubMed]
21. Reddy, A.; Gunnink, E.; Wheat, C.L.; Pawlikowski, S.; Payne, C.M.; Wiltz, S.; Hubert, T.L.; Kirsh, S.; Carey, E.; Hill, D.; et al. Rapid evaluation of artificial intelligence technology used for ambient dictation in primary care: Comparing the quality of documentation of artificial intelligence–generated and human-produced clinical notes. *Ann. Intern. Med.* **2026**, *179*, 765–772. [CrossRef] [PubMed]
22. Guo, Y.; Hu, D.; Yang, Z.; Kim, S.; Tran, B.; Lee, J.; Vallabhaneni, S.; Zehrung, R.; Sutari, S.; Tam, S.; et al. What do clinicians edit in ambient AI-drafted clinical documentation? A qualitative content analysis. *J. Am. Med. Inform. Assoc.* **2026**, *33*, 1457–1465. [CrossRef] [PubMed]
23. Wang, H.; Yang, R.; Alwakeel, M.; Kayastha, A.; Chowdhury, A.; Biro, J.M.; Sorrentino, A.D.; Handley, J.L.; Hantzmon, S.; Bessias, S.; et al. An evaluation framework for ambient digital scribing tools in clinical applications. *npj Digit. Med.* **2025**, *8*, 358. [CrossRef] [PubMed]
24. Topaz, M.; Peltonen, L.M.; Zhang, Z. Beyond human ears: Navigating the uncharted risks of AI scribes in clinical practice. *npj Digit. Med.* **2025**, *8*, 569. [CrossRef] [PubMed]
25. Victorian Government, Department of Government Services. *Guidance for the Safe and Responsible Use of Generative Artificial Intelligence in the Victorian Public Sector*; Victorian Government: Melbourne, Australia, 2024; updated March 2025. Available online: https://www.vic.gov.au/guidance-safe-responsible-use-gen-ai-vps (accessed on 21 July 2026).
26. Razaghi, M.; Hafez, A.; Farina, J.M.; Scalia, I.G.; Pereyra, M.; Abdelfattah, F.E.; Sheashaa, H.; Awad, K.; Lester, S.J.; Ayoub, C.; et al. Transforming clinical documentation with ambient artificial intelligence (AI) scribes: A narrative review of technology, impact, and implementation. *Cardiovasc. Diagn. Ther.* **2026**, *16*, 11. [CrossRef] [PubMed]
27. Anderson, T.N.; Mohan, V.; Gold, J.A. Ethical considerations for clinical adoption of ambient digital scribe technology. *J. Am. Med. Inform. Assoc.* **2026**, *33*, 770–775. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.