

DOCUMENTATION BURDEN

Ambient documentation and the clinician's day: a review of the evidence

ABSTRACT

Ambient artificial intelligence documentation is deployed at scale, but the claims made for it outrun the measurements. This paper appraises the randomized, controlled observational and deployment literature and weights each finding by its design. The randomized evidence shows a real effect that is small and specific to the product: one tool cut time in the note by 9.5% while a second tool, randomized in the same trial, changed nothing. The frequently repeated figure of about an hour saved per day is supported by none of the studies reviewed.

Type Systematic appraisal of the literature **References** 28 **Reading time** 15 min

Last reviewed September 2026

1 The question, and the size of the honest answer

Ambient documentation systems listen to a clinical encounter, transcribe it, and return a drafted note for a clinician to edit and sign. The question is practical: what have they been measured to do, and how good is the measurement? This paper appraises the published evidence, weights it by study design, and separates the findings that survive a control group from those that do not.

The short answer is that the effect is real, modest, and a property of the individual product rather than of the category. In a three-group pragmatic randomized trial of 238 outpatient physicians across 14 specialties, one ambient product reduced time in the note by 9.5% (95% CI -17.2% to -1.8% ; $P=0.02$) while a second, randomized in the same trial over the same months, changed it not at all (-1.7% ; $P=0.66$).¹ A 24-week stepped-wedge trial measured 0.36 fewer hours per day on notes.² A randomized crossover trial of 160 clinicians found no meaningful difference in pajama time between two products.³ A randomized study of AI-drafted inbox replies found read time rose 21.8%.⁴

Four claims follow. First, the measured benefit sits almost entirely in note composition, a minority of a clinician's time in the record. Second, the largest reported figures come from uncontrolled pre-post designs run while burnout was already falling for unrelated reasons. Third, the claim that ambient documentation saves about an hour a day is traceable to no study appraised here. Fourth, two of this literature's most cited papers are routinely cited for the opposite of what they report.

2 The baseline: where the day went before ambient AI

Ambient documentation is aimed at a target measured well before the intervention existed, and that baseline is less uniform than the round numbers in circulation suggest.

What the audit logs and the stopwatches found

The founding measurement is a time-and-motion study of 57 US physicians observed for 430 hours across four specialties in four states. During the office day, 27.0% of total time went to direct clinical face time and 49.2% to EHR and desk work. The 21 physicians who kept after-hours diaries reported one to two hours of additional work nightly.⁵ This is the origin of the ratio quoted ever since, nearly two hours of EHR and desk work for every hour of face time, and it is only that: the sample is 57 physicians, the authors warn the data came from self-selected, high-performing practices, and the observation predates the surge in portal messaging.

Event-log studies extended the estimate. Three years of Epic logs from 142 family medicine physicians put EHR time at 355 minutes of an 11.4-hour workday per clinical full-time equivalent, 86 of them after clinic hours, with clerical and administrative work taking 157 minutes (44.2%) and inbox management 85 minutes (23.7%).⁶ More than 31 million transactions from 471 primary care physicians split the day almost evenly, 3.08 hours on office visits against 3.17 hours on desktop medicine.⁷ The largest dataset, roughly 100 million encounters from 155,000 physicians across 417 health systems, reports 16 minutes 14 seconds per encounter: chart review 33%, documentation 24%, ordering 17%.⁸ That last figure is often quoted as an industry constant, but every encounter in it came from Cerner Millennium systems, and audit-log definitions are not interchangeable between vendors.

A cross-sectional study of 307 primary care physicians across 31 clinics gives both level and spread: median 36.2 minutes of EHR time per visit, of which 6.2 minutes was pajama time and 7.8 minutes inbox time, with clinic-level medians from 23.5 to 47.9 minutes. Above-median teamwork on order entry was associated with 3.81 fewer minutes per visit (95% CI 0.49–7.13) and a clinic pharmacy technician with 7.87 fewer (95% CI 2.03–13.72).⁹

Two facts here constrain what any documentation tool can achieve. Note authorship is a minority of time in the record, sharing the day with chart review, orders and the inbox. And the variation between clinics running the same software exceeds almost any measured ambient effect, while staffing changes in that same dataset moved more minutes per visit than most ambient products have.

What a clinical note is made of

The documentation these tools replace is largely not original composition. Across a decade at one academic medical center, median outpatient note length rose 60.1%, from 401 to 642 words, and notes written in 2018 contained a mean of 29.4% directly typed text (99% CI 28.2%–30.7%).¹⁰ A character-level analysis of 23,630 notes by 460 clinicians found 18% manually entered, 46% copied and 36%

imported.¹¹ That 18% circulates as a fact about clinical notes generally, but it describes inpatient progress notes on one general medicine service over eight months of 2016, in a research letter. The outpatient figure of 29.4% typed text is the better one to quote.

The outcome the tools are aimed at

Burnout is the endpoint most deployment studies report, and its national baseline moved sharply over the adoption period. In a repeated cross-sectional survey of 7,643 participants, 45.2% of physicians reported at least one symptom of burnout in 2023 against 62.8% in 2021 ($P<.001$) and 43.9% in 2017 ($P=.16$), and physicians remained at higher risk than other US workers (OR 1.82; 95% CI 1.63–2.05).¹² Burnout fell by about 28%, or 17.6 percentage points, between 2021 and 2023, back to its 2017 level, without help from ambient AI, and every uncontrolled before-and-after measurement taken since rides on that trend.

3 The deployment literature and the tier it occupies

Organizational reports

The most visible publications are deployment reports from The Permanente Medical Group, which gave ambient scribes to 10,000 physicians in October 2023 and recorded use in more than 303,000 encounters within 10 weeks, describing use “linked with reduced time spent in documentation and in the EHR.”¹³ A one-year follow-up covers more than 2.5 million uses; it deposits no abstract, so nothing beyond that scale figure is quoted here.¹⁴

These papers are strong evidence of feasibility at scale and none of effect size: uncontrolled, without a comparison group, authored by the organization that made the deployment decision, in a venue that publishes organizational case studies, not trials. A figure of roughly 15,791 hours saved is attributed to this deployment across trade press, and it could not be confirmed to appear in either peer-reviewed publication. It is organizational reporting, not a measured outcome.

Single-site pre-post studies

At Stanford, a survey of 48 physicians reported task load falling 24.42 points, burnout 1.94 points and usability rising 10.9 points (all $p<.001$), on a paired analysis of 38 respondents.¹⁵ Its companion audit-log study of 45 physicians found the scribe used in 9,629 of 17,428 encounters (55.25%) and median reductions of 0.57 minutes per note, 6.89 minutes per day of documentation and 19.95 minutes per day of total EHR time.¹⁶ A two-site pre-post survey enrolling 1,430 clinicians reported burnout at one site falling from 50.6% to 29.4% at 42 days ($\chi^2=42.4$; $P<.001$).¹⁷

The counterweight is a six-month evaluation of 97 ambulatory clinicians across two commercial vendors at a third academic health system: against the prior three months, reductions of 0.35 minutes per note and 2.07 minutes per day.¹⁸ Same design, similar specialty mix, a per-day estimate an order of

magnitude smaller.

LIMITATIONS OF THE PRE-POST SURVEY LITERATURE

These findings are association evidence and cannot bear a causal reading. Participants were volunteers who self-selected into pilots, which selects for enthusiasm toward the tool being evaluated. Response rates decay steeply: the two-site study reported 30.4% at 42 days, 22.0% at 84 days and 11.1% at the second site, so its burnout headline rests on the fraction of an enthusiast sample still willing to answer a fourth survey. None of these studies has a comparison group, and the national burnout rate was falling independently over the same interval.¹² A fall from 50.6% to 29.4% is a legitimate observation, not an expected deployment outcome.

Controlled observational studies

Three observational studies used comparison groups, and they are less flattering. A peer-matched cohort of 99 providers across 12 specialties against 76 matched controls, at median utilization of 47%, found favorable trends in provider engagement, a statistically significant *worsening* of after-hours EHR time, and “no significant benefits to patient experience, documentation, or measures of provider productivity.”¹⁹ Secondary summaries frequently list this paper as positive evidence for the product it evaluated. On the efficiency endpoints it is negative, and any review citing it for after-hours time savings has inverted its result.

The most carefully confounded study covers 10,344 encounters and 100 attending physicians in a tertiary academic emergency department. Ambient use was associated with 72.6 fewer seconds of on-shift documentation per encounter ($P < .001$), roughly 24 minutes across a 20-encounter shift, while after-shift documentation time *increased* by 9.1 seconds ($P = .004$). It included negative-control outcomes and a placebo permutation test.²⁰ A cross-sectional study of 198,178 emergency encounters, of which 4.3% used ambient AI and 8.0% a human scribe, found ambient AI associated with a 1.6-minute reduction in adjusted median documentation time against 3.3 minutes for human scribes, and total wRVUs per shift hour no different among groups.²¹ The confidence interval published for that ambient estimate appears to be a transcription error and is not reproduced here.

The financial question is least answered: the one cohort evaluating physician revenue, patient volumes and claim denials among adopters and nonadopters is a research letter with no full abstract, so no number from it is quoted here.²²

4 The randomized evidence

Four randomized comparisons carry the causal claims in this literature (Table 1). They are recent, and more informative about what does not change than about what does.

The three-group pragmatic trial matters most, because it randomized two commercial products against usual care within one physician population. The first was used in 33.5% of 24,696 visits and the second in 29.5% of 23,653; only the second reduced time in the note significantly. Clinically significant inaccuracies were reported as occurring occasionally, at 2.7 and 2.8 on five-point scales.¹ One trial therefore holds both a positive and a null efficiency result for the same category. Any sentence of the form “ambient AI reduces documentation time” is underspecified until it names the product.

The stepped-wedge trial covered 71,487 notes, 27,092 of them (38%) written with ambient AI. Time on notes fell 0.36 hours per day. The reduction in work outside work of 0.50 hours per day was explicitly reported as not robust, losing significance once the top 3% of daily observations were removed.² That disclosure is the trial’s most consequential sentence, because work outside work is the outcome clinicians care about most.

The crossover trial separated the product from the category: both tools reduced personal and work burnout scores, but the differences between them were not meaningful and none appeared in pajama time.³ Pajama time is the burden the category is marketed against, and where the controlled evidence is weakest.

The fourth trial addresses the inbox rather than the note. Fifty-two physicians were randomized to immediate or delayed access to AI-generated draft replies, against 70 contemporary controls. Read time rose 21.8% (95% CI 5.2% to 41.0%; $P=.008$), reply time did not change significantly (-5.9% ; $P=.33$), and replies grew 17.9% longer ($P<.001$).⁴ The study is frequently cited as evidence that generative AI reduces inbox burden. On the time metrics it shows the opposite: physicians read for longer, replied no faster, and sent longer messages. They reported liking the feature, which is a real finding and a different one.

TABLE 1

The four randomized comparisons, and what each did and did not move.

Trial	Design	Participants	Documentation-time result	Well-being or experience result
Lukac 2025 ¹	Three-group pragmatic RCT	238 physicians, 14 specialties	Nabla −9.5% time in note (P=0.02); DAX Copilot −1.7% (P=0.66), not significant	Mini-Z: DAX +2.83, Nabla +2.69; task load: DAX −39.9, Nabla −31.7
Afshar 2025 ²	Stepped-wedge randomized RCT	66 practitioners; 71,487 notes, 38% ambient	Note time −0.36 h/day; work outside work −0.50 h/day, not robust	Exhaustion and disengagement −0.44 (95% CI −0.62 to −0.25; P<0.001); fulfillment +0.14, reported nonsignificant
Chowdhury 2026 ³	Open-label randomized crossover	160 randomized; 136 analyzed	B minus A = −3.19 min in notes/day (95% CI −4.87 to −1.50); no meaningful difference in pajama time	Satisfaction 2.51 vs. 1.91 (difference 0.60; 95% CI 0.32–0.90); burnout differences not meaningful
Tai-Seale 2024 ⁴	Randomized waiting-list QI study	52 randomized; 70 contemporary controls	Read time +21.8% (P=.008); reply time −5.9% (P=.33), not significant	Reply length +17.9% (P<.001)

5 Effect sizes, and the “about an hour a day” claim

Across designs the effects are consistent in one respect: they are small, and they shrink as the design gets stronger. A scoping review of 12 included studies puts time saved per note between 0.2 and 2.1 minutes and after-hours reductions between 1.6 and 15.2 minutes.²³ The lower bound of that per-note range is twelve seconds.

−9.5% largest significant time-in-note reduction in a randomized trial, for one of two products tested	0.2–2.1 min time saved per note across 12 studies in a scoping review	−0.36 h/day note time in the stepped-wedge randomized trial	+21.8% inbox read time with AI-generated draft replies
--	---	---	--

TABLE 2

Reported effects on documentation time, ordered by design strength. None approaches an hour per day.

Study	Design	Measured change
Lukac 2025 ¹	Three-group RCT	–9.5% time in note for one product; –1.7% and not significant for the other
Afshar 2025 ²	Stepped-wedge RCT	–0.36 h/day on notes; work outside work not robust
Chowdhury 2026 ³	Randomized crossover	–3.19 min/day between products; pajama time no different
Preiksaitis 2026 ²⁰	Retrospective cohort	–72.6 s/encounter on shift; +9.1 s after shift
Dutta 2026 ²¹	Retrospective cross-sectional	–1.6 min adjusted median; human scribes –3.3 min; wRVUs unchanged
Haberle 2024 ¹⁹	Peer-matched cohort	After-hours EHR time significantly worse; no productivity benefit
Ma 2025 ¹⁶	Uncontrolled pre-post	–0.57 min/note; –6.89 min/day documentation; –19.95 min/day total EHR
Lawrence 2026 ¹⁸	Uncontrolled pre-post	–0.35 min/note; –2.07 min/day
Nallapaneni 2026 ²³	Scoping review	0.2–2.1 min saved per note; 1.6–15.2 min after hours

Against Table 2, the claim that ambient documentation saves clinicians about an hour a day cannot be sourced. The largest randomized per-day estimate is 0.36 hours, roughly 22 minutes, from the trial whose adjacent work-outside-work result did not survive a sensitivity analysis.² The largest uncontrolled estimate is 19.95 minutes of total EHR time per day, and a second study of the same design found 2.07 minutes.^{16,18} The scoping review's range tops out at 2.1 minutes per note,²³ and the three-group trial found no significant change at all for one of its two products.¹ An hour a day is roughly three times the largest randomized figure and appears in none of these studies. It is a marketing number, and a reader who meets it should ask which study it came from, because none appraised here is a candidate.

The pattern beneath the numbers matters more than the headline. Time moves in the note and moves less, or the wrong way, everywhere else: after-hours time worsened in one matched cohort¹⁹ and rose slightly after shift in the best-controlled emergency study,²⁰ pajama time did not differ between products,³ wRVUs per shift hour did not differ across scribe conditions,²¹ and inbox read time rose when drafts were introduced.⁴ One emergency medicine scoping review puts it in a phrase: these systems “shift rather than eliminate documentation effort.”²⁸

6 Note quality

Efficiency is half the question; the other half is whether the note is good. The strongest vendor-neutral evidence is a blinded comparison in the Veterans Health Administration, in which 11 AI scribe tools and 18 human note takers documented five standardized primary care cases and 30 blinded raters scored the notes on a modified PDQI-9 instrument. Human notes scored higher on all five cases, most widely in acute low back pain, 43.8 against 20.3, a difference of –23.5 (95% CI –29.2 to –17.9). AI notes scored

lower on all 10 domains, worst on thoroughness (−1.23), organization (−1.06) and usefulness (−1.03).²⁴ A narrative review of 18 studies reports frequent documentation omissions and occasional clinically significant hallucinations alongside consistent reductions in documentation burden.²⁵

WHAT THE NOTE-QUALITY COMPARISON DOES AND DOES NOT SHOW

The comparison used simulated cases with standardized patients, and the human notes were not written under real-world time pressure or interruption load. It measures the ceiling of human documentation against AI output, not the notes clinicians produce on a full clinic day. Read against the finding that the average 2018 outpatient note was 70.6% templated or copied text,¹⁰ the practical comparison is unresolved: AI notes are worse than careful human notes on every measured domain, and how they compare with routine ones is unknown.

7 The state of the evidence base

Three 2026 reviews converge. A PRISMA-ScR scoping review screened 4,772 studies and included 29: accuracy concerns appeared in nine of them, reported impacts included improved clinician well-being (n=12), reduced documentation burden (n=18) and improved patient-clinician interaction (n=10), and only three examined organizational outcomes such as cost and productivity.²⁶ A second applied the Technology Readiness Level framework and found most digital scribe systems still at TRL 3–4, with heterogeneous validation methods, mostly simulated or retrospective data and limited real-world testing, so cross-system comparison is not currently possible.²⁷ A third, in emergency medicine, included 27 sources and documented significant racial and dialect-based disparities in the underlying speech recognition systems, with higher word error rates for Black than White speakers.²⁸ It appears in a rapid-publication venue and is cited for framing and that equity finding, not for effect estimates.

One consequence deserves stating plainly, because secondary summaries get it wrong. No meta-analysis of ambient documentation with pooled quantitative effect estimates exists; everything published through 2026 in this appraisal is a scoping or narrative review.^{23,25,26,27,28} Pooling is not merely absent, it is currently invalid: the studies use different products, different denominators, different baselines, utilization rates from 29.5% to 55.25% of eligible encounters,^{1,16} and mostly simulated or retrospective validation data.²⁷ A weighted average across that set would have no referent, and any claim beginning “meta-analyses show” will not survive a check.

8 What would settle it

The gaps are specific, and naming them is more useful than another effect estimate. Multi-site randomized trials with organizational and financial endpoints are missing: only three of 29 studies in the largest scoping review examined cost or productivity,²⁶ and the cohort that addresses them publishes no abstract from which a direction of effect can be read.²² Patient experience remains a hypothesis, and the matched cohort that measured it found no significant benefit.¹⁹ Note quality has been assessed rigorously only against standardized patients, so a blinded comparison on real encounters would answer what that study cannot.²⁴ Head-to-head product randomization should be the default, since the between-product difference within one trial exceeded the category effect.¹ And reporting needs a common denominator and a common baseline before a pooled estimate can mean anything.²⁷

Until that work exists, the defensible summary is narrow. Ambient documentation reduces the time spent composing the note by a measurable amount, usually under two minutes, and the amount depends on which product is deployed. It has not been shown to reduce after-hours work durably, to improve productivity, to improve patient experience, or to produce notes matching human documentation on quality. The questions worth putting to any claim are which product, which endpoint and which comparison group; a claim of an hour a day should be treated as unsourced until a study is named.

REFERENCES

Entries 1–4 are the randomized trials and carry the causal claims made above; entries 5–12 are the pre-ambient baseline literature; entries 13–22 are deployment reports and observational cohorts, cited as association evidence only; entries 23 and 25–28 are scoping and narrative reviews, not meta-analyses. Entries 14 and 22 publish no abstract, so they are cited for existence and scope only and no number is taken from them.

1. Lukac PJ, Turner W, Vangala S, et al. *Ambient AI Scribes in Clinical Practice: A Randomized Trial*. NEJM AI. 2025;2(12). doi:10.1056/Aloa2501000 RCT
2. Afshar M, Baumann MR, Resnik F, et al. *A Pragmatic Randomized Controlled Trial of Ambient Artificial Intelligence to Improve Health Practitioner Well-Being*. NEJM AI. 2025;2(12). doi:10.1056/Aloa2500945 RCT
3. Chowdhury A, Casey M, Wilson J, et al. *Comparing ambient scribes: a randomized crossover clinical trial addressing ambient scribe technologies' impact on physician burnout*. Journal of the American Medical Informatics Association. 2026;33(5):990–999. doi:10.1093/jamia/ocag018 RCT

4. Tai-Seale M, Baxter SL, Vaida F, et al. *AI-Generated Draft Replies Integrated Into Health Records and Physicians' Electronic Communication*. JAMA Network Open. 2024;7(4):e246565. doi:10.1001/jamanetworkopen.2024.6565 RCT
5. Sinsky C, Colligan L, Li L, et al. *Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties*. Annals of Internal Medicine. 2016;165(11):753–760. doi:10.7326/M16-0961 TIME-MOTION STUDY
6. Arndt BG, Beasley JW, Watkinson MD, et al. *Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations*. Annals of Family Medicine. 2017;15(5):419–426. doi:10.1370/afm.2121 COHORT
7. Tai-Seale M, Olson CW, Li J, et al. *Electronic Health Record Logs Indicate That Physicians Split Time Evenly Between Seeing Patients And Desktop Medicine*. Health Affairs (Millwood). 2017;36(4):655–662. doi:10.1377/hlthaff.2016.0811 AUDIT-LOG STUDY
8. Overhage JM, McCallie D Jr. *Physician Time Spent Using the Electronic Health Record During Outpatient Encounters: A Descriptive Study*. Annals of Internal Medicine. 2020;172(3):169–174. doi:10.7326/M18-3684 AUDIT-LOG STUDY
9. Rotenstein LS, Holmgren AJ, Horn DM, et al. *System-Level Factors and Time Spent on Electronic Health Records by Primary Care Physicians*. JAMA Network Open. 2023;6(11):e2344713. doi:10.1001/jamanetworkopen.2023.44713 CROSS-SECTIONAL
10. Rule A, Bedrick S, Chiang MF, et al. *Length and Redundancy of Outpatient Progress Notes Across a Decade at an Academic Medical Center*. JAMA Network Open. 2021;4(7):e2115334. doi:10.1001/jamanetworkopen.2021.15334 RETROSPECTIVE ANALYSIS
11. Wang MD, Khanna R, Najafi N. *Characterizing the Source of Text in Electronic Health Record Progress Notes*. JAMA Internal Medicine. 2017;177(8):1212–1213. doi:10.1001/jamainternmed.2017.1548 RESEARCH LETTER
12. Shanafelt TD, West CP, Sinsky C, et al. *Changes in Burnout and Satisfaction With Work-Life Integration in Physicians and the General US Working Population Between 2011 and 2023*. Mayo Clinic Proceedings. 2025;100(7):1142–1158. doi:10.1016/j.mayocp.2024.11.031 NATIONAL SURVEY
13. Tierney AA, Gayre G, Hoberman B, et al. *Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation*. NEJM Catalyst Innovations in Care Delivery. 2024;5(3). doi:10.1056/CAT.23.0404 DEPLOYMENT REPORT
14. Tierney AA, Gayre G, Hoberman B, et al. *Ambient Artificial Intelligence Scribes: Learnings after 1 Year and over 2.5 Million Uses*. NEJM Catalyst Innovations in Care Delivery. 2025;6(5). doi:10.1056/CAT.25.0040 DEPLOYMENT REPORT

15. Shah SJ, Devon-Sand A, Ma SP, et al. *Ambient artificial intelligence scribes: physician burnout and perspectives on usability and documentation burden*. Journal of the American Medical Informatics Association. 2025;32(2):375–380. doi:10.1093/jamia/ocae295 PRE-POST SURVEY
16. Ma SP, Liang AS, Shah SJ, et al. *Ambient artificial intelligence scribes: utilization and impact on documentation time*. Journal of the American Medical Informatics Association. 2025;32(2):381–385. doi:10.1093/jamia/ocae304 PRE-POST AUDIT-LOG
17. You JG, Dbouk RH, Landman A, et al. *Ambient Documentation Technology in Clinician Experience of Documentation Burden and Burnout*. JAMA Network Open. 2025;8(8):e2528056. doi:10.1001/jamanetworkopen.2025.28056 PRE-POST SURVEY
18. Lawrence K, Polet C, Malhotra K, et al. *Implementing Artificial Intelligence-Enabled Ambient Documentation Technology for Ambulatory Clinicians: An Innovation Evaluation*. Journal of General Internal Medicine. 2026; published online ahead of print. doi:10.1007/s11606-026-10448-w PRE-POST EVALUATION
19. Haberle T, Cleveland C, Snow GL, et al. *The impact of nuance DAX ambient listening AI documentation: a cohort study*. Journal of the American Medical Informatics Association. 2024;31(4):975–979. doi:10.1093/jamia/ocae022 MATCHED COHORT
20. Preiksaitis C, Alvarez A, Winkel M, et al. *Ambient AI Scribes and Emergency Department Documentation Burden: Retrospective Cohort Study*. JMIR AI. 2026;5:e92193. doi:10.2196/92193 RETROSPECTIVE COHORT
21. Dutta S, You JG, Dunham L, et al. *Medical Scribe and Ambient Artificial Intelligence Impact on Emergency Physician Documentation Burden and Clinical Productivity*. Annals of Emergency Medicine. 2026; published online ahead of print, S0196-0644(26)00266-0. doi:10.1016/j.annemergmed.2026.04.022 CROSS-SECTIONAL
22. Holmgren AJ, Fenton CL, Thombly R, et al. *Ambient Artificial Intelligence Scribes and Physician Financial Productivity*. JAMA Network Open. 2026;9(1):e2553233. doi:10.1001/jamanetworkopen.2025.53233 RESEARCH LETTER
23. Nallapaneni S, Crafton CL, Cabrera CI, et al. *Ambient Artificial Intelligence Scribes: A Scoping Review With Implications for Otolaryngology*. The Laryngoscope. 2026; published online ahead of print. doi:10.1002/lary.70625 SCOPING REVIEW
24. Reddy A, Gunnink E, Wheat CL, 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*. Annals of Internal Medicine. 2026;179(6):765–772. doi:10.7326/ANNALS-25-02772 COMPARATIVE EVALUATION

25. Razaghi M, Hafez A, Farina JM, et al. *Transforming clinical documentation with ambient artificial intelligence (AI) scribes: a narrative review of technology, impact, and implementation*. Cardiovascular Diagnosis and Therapy. 2026;16(1):11. [doi:10.21037/cdt-2025-454](https://doi.org/10.21037/cdt-2025-454)
NARRATIVE REVIEW
26. Yu SHT, Li J, Carland JE, et al. *Adoption and utility of digital scribes in clinical practice — A scoping review*. International Journal of Medical Informatics. 2026;220:106630. [doi:10.1016/j.ijmedinf.2026.106630](https://doi.org/10.1016/j.ijmedinf.2026.106630) SCOPING REVIEW
27. Kerimoğlu E, Notermans FV, Silkens MEWM, et al. *Validating Digital Scribes: A Scoping Review of Evaluation Practices and Clinical Use*. Journal of Medical Systems. 2026;50(1):62. [doi:10.1007/s10916-026-02392-3](https://doi.org/10.1007/s10916-026-02392-3) SCOPING REVIEW
28. Wolfe J, Welsh SM. *Ambient Documentation Systems in Emergency Medicine: A Scoping Review of Clinical Precision, Patient Experience, Throughput, and Quality*. Cureus. 2026;18(4):e106643. [doi:10.7759/cureus.106643](https://doi.org/10.7759/cureus.106643) SCOPING REVIEW

RELATED PAPERS

Evaluating clinical language models

How benchmark scores are produced, and why so few of them describe clinical work.

Human oversight of clinical AI

What happens to reviewers who are handed a confident wrong answer.

A write-path architecture for an AI-native EMR

Where a model may write, and where a clinician must.