Ambient AI scribes: what the evidence actually shows.

The honest 2026 read for clinicians and health-system leaders deciding on ambient documentation: what the largest deployment proved, what the randomized trials found, why time-savings distributions matter more than vendor averages, where notes go wrong, and the assumptions that make or break the business case. Every claim carries a numbered primary source. As of August 2026.

2,576,627

patient encounters in the largest published deployment[5,6]

8,581

clinicians in the largest controlled cohort of scribe time effects[14]

26

numbered primary sources, each verified against the original

Contents

Executive summary

Ambient artificial intelligence (AI) scribes — tools that listen to a clinical encounter and draft the note for the clinician to review and sign — are among the most widely deployed generative-AI technologies in clinical care, and the evidence has finally begun to catch up with the adoption.

The largest published deployment, at Kaiser Permanente's Permanente Medical Group, reached 7,260 physicians and 2,576,627 patient encounters in 63 weeks, with an estimated 15,791 hours of documentation time saved[5,6]. The first randomized trials, in 2025, complicated the marketing story: one scribe cut time-in-note 9.5% versus control while the other showed no significant change[9]. The largest controlled cohort — 8,581 clinicians across five academic health systems — measured savings of about 13 minutes of electronic health record (EHR) time and 16 minutes of documentation time per 8 scheduled patient hours, no significant change in after-hours EHR time, and only about 32% of adopters using the scribe in at least half their visits[14]. Accuracy studies show omissions outnumber fabrications by roughly two- to threefold[17,19], and human-written notes still outscore AI notes in blinded quality evaluations[18,20].

The honest verdict: adoption at scale is proven, burnout relief is the most consistent finding, time savings are real but modest and unevenly distributed, the financial return rests on assumptions you should test locally, and the clinician-review step remains load-bearing. This paper walks through each claim with primary sources, then gives the measurement plan for a defensible pilot.

A note on scope

This paper appraises published evidence. It is not clinical, legal, or purchasing advice; confirm regulatory status, contract terms, and compliance obligations for any specific tool with your own governance, IT, and counsel.

1

The burden the scribe is aimed at

Before weighing the intervention, weigh the problem, because it is measured and large — and it predates AI.

A time-and-motion study across four specialties found that during the office day, physicians spent **27.0% of their time on direct clinical face time with patients and 49.2% on EHR and desk work** — as the authors put it, "for every hour physicians provide direct clinical face time to patients, nearly 2 additional hours is spent on EHR and desk work within the clinic day"[1]. An EHR event-log study of 142 family physicians measured **5.9 hours of an 11.4-hour workday in the EHR**, including **1.4 hours after clinic** — the after-hours charting clinicians call "pajama time"[2]. The multisite scribe study discussed below opens from the same literature: an average of **2.3 hours of documentation per 8 hours of patient care**[14].

The burden is not just time; it tracks with distress through validated instruments. In a national survey, physicians rated their EHR's usability at **45.9 of 100 on the System Usability Scale (SUS)** — in the bottom 9% of products ever tested with that instrument, a grade of F — and after adjustment, **each 1-point more favorable usability score was associated with 3% lower odds of burnout** (odds ratio 0.97; 95% confidence interval [CI], 0.97–0.98)[3].

5.9 hours

of an 11.4-hour workday spent in the electronic health record by 142 family physicians, including 1.4 hours after clinic.

Source: Arndt BG, et al. Tethered to the EHR. Ann Fam Med. 2017;15(5):419–426[2].

That is the opening for an ambient scribe: take the drafting of the note off the clinician's hands, and perhaps the burden with it. The rest of this paper asks how much of that promise the evidence supports.

2

Scale is settled: the Kaiser Permanente deployment

The number everyone quotes — "2.5 million uses" — comes from two papers in NEJM Catalyst Innovations in Care Delivery by The Permanente Medical Group (TPMG), the physician group of Kaiser Permanente's Northern California region[4,5]. They are the closest thing the field has to a national-scale case study, and they reward a careful read.

The launch was fast. In October 2023, TPMG began offering the tool to **10,000 physicians at 21 locations in Northern California**; within the first 10 weeks it had been "adopted by 3,442 physicians who used it in 303,266 patient encounters," including **968 physicians who used the tool more than 100 times**[4,7]. The one-year follow-up reports the totals now in wide circulation: **7,260 Permanente physicians used the technology across 2,576,627 patient encounters** during a **63-week evaluation from October 2023 through December 2024**, with an estimated **15,791 hours of documentation time saved**[5,6]. By the end of that first year, **more than 3,400 physicians had used the technology during at least 100 patient visits**[6]. Physician surveys were favorable — **84% reported a positive effect on communication and 82% said overall work satisfaction improved** — and among surveyed patients, **56% reported a positive impact on visit quality while none reported a negative one**[6].

15,791

hours of documentation time saved, estimated across 2,576,627 patient encounters in a 63-week evaluation — a comparison of users against nonusers, not a randomized measurement.

Source: The Permanente Medical Group. NEJM Catalyst Innov Care Deliv. 2025[5]; American Medical Association[6].

Quality assurance was operational rather than assumed: a standing program scored note quality with a modified Physician Documentation Quality Instrument (PDQI), and **1,252 questionnaire**

responses averaged 4.35 of 5 points[8]. The technology has since been deployed across Kaiser Permanente's **8 regions, 600 medical offices, and 40 hospitals, supporting more than 4 million encounters** as of March 2025[8].

Now the caveats, which the deployment's own authors would recognize. This is an observational rollout: no randomization, no patient outcomes, and no note-level error rate across the millions of encounters. The hours-saved figure is an estimated comparison of users against nonusers, not a randomized measurement. The setting — a large, integrated group with a single EHR and enterprise-grade support — is the ceiling of what a rollout can look like, not the median. And one detail matters for vendor claims: TPMG changed scribe vendors between the initial pilot and the wider rollout, and the public record of the deployment names neither vendor[6] — so no single product can claim this evidence as its own. The Kaiser record settles one question definitively: **ambient scribes can be adopted at scale, quickly, and clinicians keep using them.** Whether they cause the improvements clinicians feel is a question for controlled designs — which now exist.

3

What the randomized and controlled evidence shows

Through 2024, the published base was thin. A rapid review of real-world evidence, published in JMIR AI in 2025, screened the literature and reported: "Of the 1450 studies identified, 6 met the inclusion criteria" — and those six were an observational study, a case report, a peer-matched cohort, and surveys. Its verdict: "the studies fell short when compared to standardized frameworks," and "the currently available evidence is sparse"[13]. The tools reached millions of encounters before the first randomized trial reported. In 2025 and 2026, that changed.

The first randomized controlled trial (RCT) made the effect tool-specific

At UCLA Health, 238 outpatient physicians across 14 specialties were randomized to one of two widely used scribes — Microsoft Dragon Ambient eXperience (DAX) Copilot or Nabla — or usual care, across roughly 72,000 encounters[9,10]. On the primary outcome, log-measured writing time-in-note, "Nabla users experienced a 9.5% (95% confidence interval [CI], -17.2% to -1.8%; P=0.02) decrease in time-in-note versus the control group, whereas DAX users exhibited no significant change versus the control group (-1.7%; 95% CI, -9.4% to +5.9%; P=0.66)"[9]. In absolute terms, the winning arm's per-note writing time fell by an estimated 41 seconds (4 minutes 30 seconds to 3 minutes 49 seconds), against an 18-second drop in control[10].

Both arms showed potential improvements on burnout and task-load instruments, but the authors were explicit that "these secondary end point findings need confirmation in larger, multicenter trials"[9]. Two more results matter: the scribes were used in only about a third of eligible visits (DAX in 33.5% of 24,696 visits; Nabla in 29.5% of 23,653)[9], and physicians reported clinically significant inaccuracies "occasionally" — about 2.7–2.8 on a 5-point scale — most commonly omissions and pronoun errors[9,10].

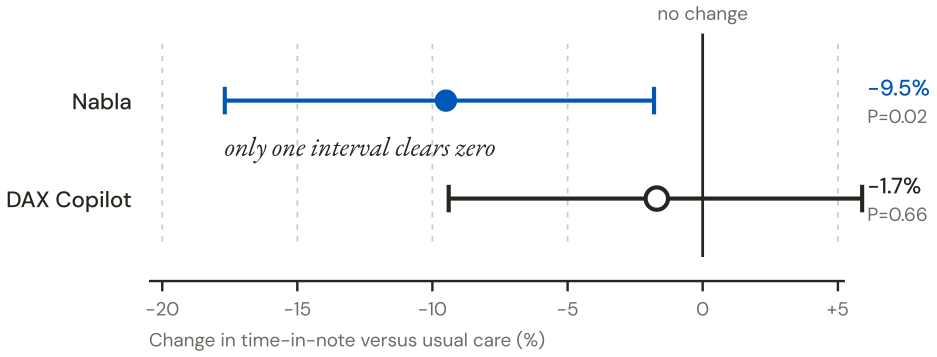

Figure 1 — Two ambient scribes, one randomized trial, two different answers. Point estimates with 95% confidence intervals for the change in log-measured time-in-note against usual care, among 238 outpatient physicians at UCLA Health. Source: Lukac PJ, et al. NEJM AI. 2025[9].

A pragmatic trial found relief without transformation

A 24-week, stepped-wedge, individually randomized trial of 66 practitioners across ambulatory clinics in two states found ambient AI produced "a significant reduction in work exhaustion/interpersonal disengagement (-0.44 points)" but "a nonsignificant increase in professional fulfillment" on its coprimary outcomes; time on notes fell 0.36 hours per day, an after-hours-work reduction was sensitive to

outlier exclusion, and documentation quality and billing compliance held steady. Notably for governance, the team monitored the tool in production and reported "No drift in software performance was detected"[11].

A crossover trial found the tools differ — again

At Duke, 160 outpatient clinicians used two ambient scribes in randomized sequence. Both reduced burnout scores on the Copenhagen Burnout Inventory, but one product outperformed the other on satisfaction and showed "greater reductions in average minutes-in-notes per day" (a difference of 3.19 minutes per day), while "efficiency metrics like pajama time were largely unaffected"[12].

A six-system pre-post study produced the headline burnout number

Across six academic and community health systems, among 263 analyzed clinicians using one ambient platform for 30 days — 186 of whom completed the burnout measure — "the proportion of participants experiencing burnout decreased significantly from 51.9% to 38.8% (odds ratio, 0.26; 95% CI, 0.13–0.54)," with improvements in cognitive task load and after-hours documentation[15]. Two honesty notes: this is a voluntary, survey-based, 30-day pre-post design — the study type most flattered by novelty — and the author list includes a coauthor affiliated with Abridge AI Inc., the maker of the platform studied[15]. The direction of the finding is consistent with the randomized trials; the size should be read with those design limits in view.

The pattern across every controlled design is remarkably stable. Burnout and exhaustion measures move reliably in the right direction. Clock-measured time moves modestly, and it moves differently for different tools — twice now, in randomized comparisons, one scribe delivered measurable time savings and another did not[9,12].

Any sentence that begins "AI scribes save…" without naming the tool, the setting, and the denominator is a marketing sentence, not an evidence sentence.

4

Time savings: read the distribution, not the average

The largest controlled study of scribe time effects is a 2026 multisite cohort in JAMA: 8,581 ambulatory clinicians — 1,809 of them scribe adopters — across five US academic health systems (Mass General Brigham, Emory, UCSF, Yale New Haven, UC Davis), using Ambience, Nuance DAX Copilot, or Abridge on Epic, analyzed in a difference-in-differences framework[14]. The headline: adoption was associated with "13.4 (95% CI, 9.1–17.7) fewer minutes of EHR time" and "16.0 (95% CI, 13.7–18.3) fewer minutes of documentation time" per 8 scheduled patient hours, plus "0.49 (95% CI, 0.17–0.81) additional weekly visits delivered" — relative decreases of 3.0% in total EHR time and 10.0% in documentation time. "Electronic health record time outside work hours did not change significantly"[14].

Those averages are honest, and they are also the least interesting thing in the paper. What matters for a buyer is the distribution underneath:

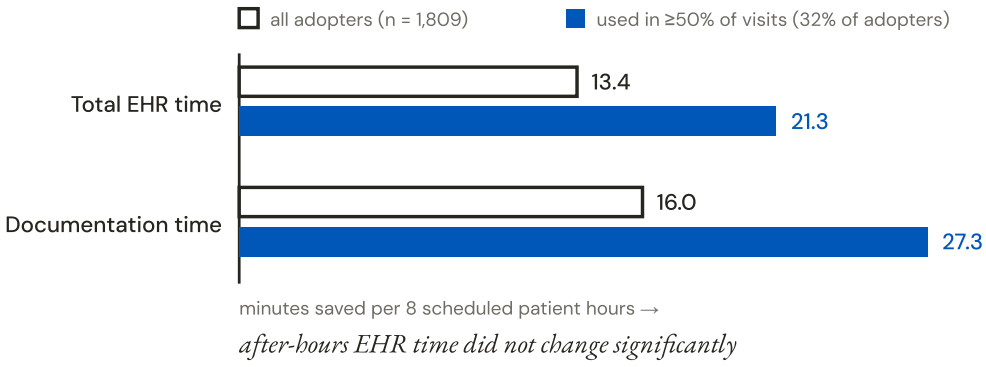

Figure 2 — The average is a blend, not a person. Minutes saved per 8 scheduled patient hours among all scribe adopters and among the minority who used the tool in at least half their visits, across five academic health systems. Source: Rotenstein LS, et al. JAMA. 2026[14].

— **Utilization is the hidden variable.** Clinicians who used the scribe in 50% or more of visits saved 21.3 minutes of EHR time and 27.3 minutes of documentation time — roughly two and three times the average effect, respectively — and delivered a full additional visit per week. But "only about 32% of adopters used their AI

scribe that frequently"[14]. The average blends heavy users who gained a lot with light users who gained little.

- **Benefits concentrate in predictable groups.** Primary care clinicians saved 25.0 EHR minutes and 26.9 documentation minutes; female clinicians, advanced practice clinicians, and residents also saw greater-than-average changes[14].
- **The saved time does not simply vanish from the workday.** Documentation time fell more than total EHR time, and after-hours EHR time did not significantly change — a signature consistent with clinicians reallocating minutes to inbox, chart review, and other EHR work rather than leaving earlier[14]. The Duke trial's flat "pajama time" result points the same way[12].

32%

of scribe adopters used the tool in half or more of their visits. That minority saved 21.3 EHR minutes and 27.3 documentation minutes per 8 scheduled patient hours.

Source: Rotenstein LS, et al. JAMA. 2026;335(16):1408–1417[14].

Now set that against how the category is marketed. Microsoft's Dragon Copilot announcement cites "5 minutes of time-savings per encounter" — sourced, in its own footnote, to a July 2024 Microsoft survey of 879 clinicians using DAX Copilot[26]. The independent randomized trial of its DAX Copilot lineage measured no significant change in time-in-note versus control, and the winning tool in that trial saved about 41 seconds per note[9,10]. Both statements can be simultaneously true — a vendor's survey of its own users is not a randomized comparison — but only one belongs in a base-case business model.

When a vendor average and a controlled distribution disagree, plan on the distribution.

5

Accuracy: how notes go wrong, and how systems catch it

Most effectiveness studies never check whether the note is true. The studies that do have converged on a consistent error anatomy, and the single most useful skill for a buyer is checking the **unit** behind any quoted rate — per sentence, per note, or per case — because the honest numbers range from about 1% to over 30% depending on that unit.

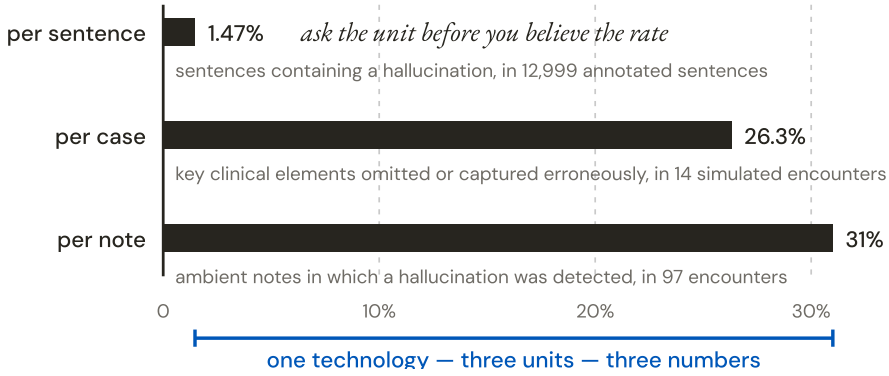

Figure 3 — Three published error rates from three studies, each measured against a different unit. The numbers are not comparable with one another; the unit is the whole difference. Sources: Asgari E, et al. npj Digit Med. 2025[17]; Anderson TN, et al. Mayo Clin Proc Digit Health. 2025[19]; Palm E, et al. Front Artif Intell. 2025[18].

Per sentence, errors are rare — and asymmetric

A clinician-in-the-loop framework — built, for disclosure, by clinician-researchers at the vendor Tortus AI — annotated 12,999 sentences of large language model (LLM)-generated clinical documentation and "observed a 1.47% hallucination rate and a 3.45% omission rate" — omissions (real information left out) more than double hallucinations (content invented)[17]. The same group showed the rates are not fixed properties of the technology: "By refining prompts and workflows, we successfully reduced major errors below previously reported human note-taking rates"[17]. Rates move with the model, the prompt, and the workflow — which means a vendor's number from last year does not describe this year's product, in either direction.

Per note, fabrication is common enough to demand review

A validated evaluation across 97 encounters and five specialties — conducted, for disclosure, by authors affiliated with the vendor Suki AI — compared LLM-drafted "ambient" notes with physician-drafted "gold" notes: "Hallucinations were detected in 20% of gold notes and 31% of ambient notes" — a statistically significant gap. The same study found the two note types close on overall quality (gold 4.25 vs ambient 4.20 of 5), with gold notes better on accuracy, succinctness, and internal consistency, ambient notes better on thoroughness and organization — and reviewers, blinded, actually preferred the ambient notes 47% to 39%[18]. Fuller, better-organized, more often subtly wrong: that is the trade a reviewer needs to know they are managing.

Per case, stress tests map the failure surface

An Oregon Health & Science University (OHSU) team ran 14 simulated ambulatory encounters through five ambient scribe platforms and graded every element. Mean note error rate was "26.3% (95% CI, 17.0%–31.0%)" of key clinical elements omitted or captured erroneously; "errors of omission were most frequent across all platforms, comprising a mean of 76.3%… of all errors"; and "an average of 3.0 (95% CI, 0–4, range 0–21) errors per case had potential for moderate-to-severe harm" on the Agency for Healthcare Research and Quality (AHRQ) harm scale[19]. Upstream speech recognition mattered: across the four platforms that produce transcripts, transcripts averaged 13.9 errors per case, and 19.5% of transcript errors were transmitted into the final note[19]. Only 35.8% of clinical elements were captured correctly by all five platforms — the tools do not even fail in the same places[19]. The starkest example: in a simulated pneumonia-with-sepsis case where the clinician articulated immediate admission and a specific antibiotic dose, one platform's note "stated that admission would be considered and antibiotics would only be initiated if infection confirmed"[19]. Simulation overstates routine-visit rates by design; use it to understand modes, not to forecast your baseline.

76.3%

mean share of all note errors that were errors of omission — the dominant and quietest failure mode, and the reason clinician review is load-bearing.

Source: Anderson TN, et al. Mayo Clin Proc Digit Health. 2025;3(4):100292[19].

Head to head with humans, humans still win on quality

A Veterans Health Administration evaluation put 11 AI scribe tools and 18 human note-takers through five standardized primary-care cases, with 30 blinded raters scoring notes on the modified PDQI-9. "Across all 5 clinical cases, human-generated notes received higher overall modified PDQI-9 scores than AI-generated notes" — most dramatically in an acute low back pain case (human 43.8 vs AI 20.3 of 50). The authors' conclusion is the field's to-do list: "independent, vendor-neutral evaluations of note quality are essential before large-scale clinical deployment"[20].

How real systems manage the risk

Three published control layers. First, **universal clinician review**: every published deployment keeps the clinician reading, editing, and signing each note — the omission-dominant error profile is why, since a missing allergy is visible only to a reviewer who remembers the encounter. Second, **standing quality assurance (QA)**: TPMG's program combined in-workflow star ratings of draft notes, structured PDQI scoring, and staff review of more than 3,600 clinician free-text comments — surveillance as a routine, not a one-time validation[8]. Third, **production monitoring**: the Wisconsin trial checked for performance drift over its 24 weeks and reported finding none — the right habit, since the absence of drift is something you verify, not assume[11]. One warning belongs beside all three: reviewer vigilance is imperfect — in simulation work the OHSU team cites, clinicians editing AI-generated drafts frequently failed to recognize clinically relevant errors — so audit signed notes, not just drafts[19].

6

The vendor landscape: documented versus marketed

The market is crowded and capital-rich — one 2025 evaluation counted "over 90 platforms currently available for clinical use," and "as of early 2024, nearly a third of medical groups… reported adoption"[19]. The useful discipline is separating three kinds of statement: what a vendor discloses about itself (fact, but self-reported), what it claims about outcomes (marketing until independently measured), and what independent studies have tested (the smallest pile). The figures below are each company's own dated announcements — as of August 2026, with the standing caution that funding measures investor conviction, never clinical performance.

Vendor	Own disclosures (dated)	Independent evidence identified
Abridge	$300M Series E led by Andreessen Horowitz, June 2025; "partnering with more than 150 enterprise health systems"[22]	Platform in the six-system burnout study, which included an Abridge-affiliated coauthor[15]; among the tools in the JAMA multisite cohort[14]
Microsoft Dragon Copilot	Announced March 2025; carries the Nuance DAX ambient lineage; "captures multiparty, multilingual patient-clinician conversations" into specialty-specific draft notes[26]	DAX Copilot arm of the UCLA RCT: no significant time-in-note change versus control[9]; among tools in the JAMA multisite cohort[14]
Nabla	$70M Series C, June 2025, total $120M; "used by 85,000 clinicians across more than 130 healthcare organizations"[24]	Nabla arm of the UCLA RCT: -9.5% time-in-note versus control (P=0.02)[9]
Ambience Healthcare	$243M Series C, July 2025; "supports more than 100 ambulatory subspecialties, EDs, and inpatient specialties"; Epic, Oracle Cerner, athenahealth[23]	Among tools in the JAMA multisite cohort[14]; no vendor-specific independent trial identified
Suki	$168M raised to date after a January 2025 Zoom Ventures investment; assistant spans documentation, dictation, coding, chart Q&A; "used by 350 health systems and clinics"[25]	Vendor-affiliated authors published the PDQI-9 note-quality evaluation[18]; no vendor-independent trial identified

Reconfirm every cell with the vendor at contract time; these figures age quickly.

Three structural observations. First, **the evidence is lopsided**: a handful of tools carry nearly all the vendor-specific independent evidence — and the largest deployment record of all, Kaiser's, names no vendor at all and spans a vendor change[6], so no product can claim it. An empty cell is a prompt to demand a study, not proof of poor performance. Second, **no published study ranks named tools on accuracy under one rubric** — the VA evaluation tested 11 tools but reported them against human notes, not against each other by name[20] — so any "most accurate scribe" claim is, as of August 2026, unbacked by a neutral head-to-head. Third, **integration claims are checkable; outcome claims usually are not.**

"Deployed inside Epic" is a verifiable binary. "Saves five minutes per encounter" is a customer anecdote until someone measures it against a comparison group — and when someone did, for one major tool, the measured answer was null.[9,26]

7

An honest ROI frame

The return-on-investment (ROI) conversation is where scribe procurement most often goes wrong, because the arithmetic is simple and the assumptions are not. Here is the frame the evidence supports, assumption by assumption.

Start from measured inputs, not brochure inputs

The defensible per-adopter numbers are those from controlled studies: roughly 13 minutes of EHR time and 16 minutes of documentation time saved per 8 scheduled patient hours, about half an additional visit per week[14], and — from a separate single-system cohort at UCSF — a productivity signal of "1.81 (95% CI, 0.86–2.75) greater RVUs per week relative to nonadopters" among 698 adopters, which "translates to $3044 annually per physician, using the 2025 Medicare Physician Fee Schedule"[16]. (A relative value unit, or RVU, is the standard US unit of billed physician work; RVU gains can reflect

more visits or richer coding.) The JAMA multisite cohort's parallel revenue estimate was even more modest: about $167 per adopting clinician per month in marginal evaluation-and-management (E/M) revenue[14]. Set those figures against your negotiated per-clinician subscription price and implementation cost, and the margin is thin. That is the honest starting point.

The assumption that breaks most models is utilization

Every per-clinician saving above scales with actual use, and in the best multisite data only about 32% of adopters used the scribe in half or more of their visits[14]; in the UCLA trial, the tools were used in roughly a third of eligible visits[9]. A model that licenses 100 clinicians and assumes 100 heavy users will overstate the return several-fold. Model three tiers — heavy, light, and lapsed — from your own pilot telemetry, not from hope.

Count the costs the vendor calculator omits

Clinician review time is already netted out of log-measured savings (the minutes above are what remained after editing), so do not add it back — but do count implementation, training, consent workflows, and the standing QA program that a responsible deployment runs[8]. Ramp time is real, and light users can lose time on unfamiliar drafts before they gain it.

Be careful what you wish for on coding

Revenue gains that come from richer E/M coding are only a benefit if the documentation truthfully supports the codes. The UCSF authors flagged the open question themselves: further research should "ascertain whether increased RVUs reflect more clinical services or accurate coding rather than upcoding"[16]. Encouragingly, they found no change in claim denials[16], and the Wisconsin trial found billing compliance intact under professional coder review[11] — but an ROI case built primarily on code-level drift is a compliance exposure, not a return.

Put burnout on its own line — and take it seriously there

The most consistent return in the whole literature is felt relief: burnout prevalence falling from 51.9% to 38.8% in 30 days across six systems[15], exhaustion improving in randomized designs[9,11,12]. That is a real asset for retention and recruitment in a workforce-constrained decade, and it resists honest conversion to dollars. The Peterson Health

Technology Institute's 2025 assessment — drawn from the systems actually running these tools — is the fairest one-line summary of the financial state of play: "ambient scribes are potentially effective at reducing clinician documentation time and cognitive load," while "health system leaders indicate that there are gaps in evidence regarding the impact of ambient scribes on productivity and financial performance"[21].

38.8%

burnout prevalence among 186 clinicians after 30 days on one ambient platform, down from 51.9% (odds ratio, 0.26; 95% CI, 0.13–0.54) — in a voluntary, survey-based pre-post design.

Source: Olson KD, et al. JAMA Netw Open. 2025;8(10):e2534976[15].

The bottom line: with measured inputs, a scribe program pencils out as a modest financial case plus a meaningful workforce case — and the spread between a thin positive and a clear negative is determined almost entirely by utilization depth and specialty mix. Which is why the last section matters most.

8

What to measure in a pilot

The published evidence cannot tell you what a scribe will do in your clinics; it tells you what to measure so your pilot can. A defensible design: at least 6 weeks of baseline and 12 weeks of intervention, a concurrent comparison group of similar non-users (the multisite cohort shows how far difference-in-differences can go when randomization is impractical[14]), and results reported as distributions, not averages.

What to measure in a pilot

Print this page and pin it beside the project plan.

Time, from logs — never from memory

- Total EHR time, documentation time, and after-hours EHR time per 8 scheduled patient hours, from EHR audit logs (Epic Signal or equivalent), matching the definitions in the multisite study so you can benchmark directly[14].
- Report the per-clinician distribution: your median user's change, your top-quartile change, and the share of clinicians who saved nothing.

Utilization, per clinician

- Share of eligible visits with the scribe used, per clinician per week — the single strongest predictor of benefit[14]. Set an explicit target (the published heavy-user threshold is ≥50% of visits) and track the share of adopters reaching it.

Accuracy, with the unit named

- A structured audit of a random sample of *signed* notes against recordings or clinician recall: hallucination rate and omission rate per note, harm-graded (the AHRQ scale gives you a published rubric[19]).
- Count omissions separately — they are the dominant and quietest failure mode[17,19].
- Repeat the audit after any vendor model update; rates are version-specific[17].

Experience, with validated instruments

- A burnout measure (Mini-Z, Professional Fulfillment Index, or Copenhagen Burnout Inventory) before and after, so your result is comparable to the trials[9,11,12]. Expect the survey to move more than the clock; that gap is the published norm, not a failure.

Money, honestly

- Visits per week, RVUs per week, E/M level mix, and claim-denial rate for users versus comparison[14,16]. If E/M levels shift upward, trigger a documentation-integrity review rather than booking the gain[16].

Governance, from day one

- Patient consent workflow appropriate to your jurisdictions; a standing QA channel for clinician-reported errors[8]; a monitoring plan that checks for performance drift on a schedule[11]; and pre-agreed kill criteria — the utilization floor, accuracy floor, and time-savings floor below which you exit the contract.

Design: at least 6 weeks of baseline and 12 weeks of intervention, with a concurrent comparison group of similar non-users. Report distributions, not averages.

A pilot that measures these eight things will end with the only ROI number that matters: yours.

About this paper and further reading

This white paper extends the ambient-documentation coverage maintained on aimocs.com:

The Kaiser deployment analysis

aimocs.com/guides/kaiser-2-5m-encounter-deployment-analyzed

The evidence appraisal

aimocs.com/guides/ai-scribes-what-the-evidence-actually-shows

The documentation-time and burnout literature

aimocs.com/guides/documentation-time-and-burnout-evidence

Hallucination and omission rates, keyed by unit

aimocs.com/guides/hallucination-and-omission-rates-in-scribe-notes

The vendor landscape

aimocs.com/guides/ai-scribe-vendor-landscape-2026

The transparent ROI model

aimocs.com/guides/ambient-ai-roi-calculator

The implementation checklist

aimocs.com/guides/implementation-checklist-for-medical-groups

Method

Every factual claim carries a numbered source, and every load-bearing figure was verified against the primary source — the journal abstract or open-access full text (retrieved via the publisher, PubMed, or PubMed Central), or the organization's own announcement; a claims register with verbatim source quotes accompanies this paper. Where a primary paper sits behind a paywall (the two NEJM Catalyst deployment papers), figures were verified against public reporting of those papers by the American Medical Association and Permanente Medicine and cited to both the paper and the public page. Vendor figures are the vendors' own disclosures, labeled as such. Nothing here is clinical advice. As of August 2026.

Sources

1 Sinsky C, et al. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. Ann Intern Med. 2016;165(11):753–760.
https://doi.org/10.7326/M16-0961
Direct-observation source for the 27.0% / 49.2% time split.

2 Arndt BG, et al. Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. Ann Fam Med. 2017;15(5):419–426.
https://doi.org/10.1370/afm.2121
EHR-log source for 5.9 hours/day and 1.4 after-hours.

3 Melnick ER, et al. The Association Between Perceived Electronic Health Record Usability and Professional Burnout Among US Physicians. Mayo Clin Proc. 2020;95(3):476–487.
https://doi.org/10.1016/j.mayocp.2019.09.024
SUS 45.9 (grade F) and 3% lower burnout odds per point.

4 Tierney AA, et al. Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation. NEJM Catalyst Innov Care Deliv. 2024;5(3).
https://doi.org/10.1056/CAT.23.0404
TPMG launch paper: 3,442 physicians, 303,266 encounters in 10 weeks.

5 The Permanente Medical Group. Ambient Artificial Intelligence Scribes: Learnings after 1 Year and over 2.5 Million Uses. NEJM Catalyst Innov Care Deliv. 2025.
https://doi.org/10.1056/CAT.25.0040
One-year deployment paper: 7,260 physicians, 2,576,627 encounters.

6 American Medical Association. AI scribes save 15,000 hours — and restore the human side of medicine. AMA (accessed August 2026).
https://www.ama-assn.org/practice-management/digital-health/ai-scribes-save-15000-hours-and-restore-human-side-medicine
Public verification of the year-1 figures: 15,791 hours, survey results, and the note that TPMG changed scribe vendors.

7 American Medical Association. AI scribe saves doctors an hour at the keyboard every day. AMA, March 2024 (accessed August 2026).
https://www.ama-assn.org/practice-management/digital-health/ai-scribe-saves-doctors-hour-keyboard-every-day
Public verification of the launch figures: 10,000 offered, 968 heavy users.

8 Permanente Medicine. Quality assurance informs large-scale use of ambient AI clinical documentation. March 26, 2025 (accessed August 2026).
https://permanente.org/quality-assurance-informs-large-scale-use-of-ambient-ai-clinical-documentation/
QA program design, PDQI 4.35/5 on 1,252 responses, more than 4M encounters across 8 regions.

9 Lukac PJ, et al. Ambient AI Scribes in Clinical Practice: A Randomized Trial. NEJM AI. 2025;2(12).
https://doi.org/10.1056/AIoa2501000
First RCT: Nabla −9.5% time-in-note, DAX no significant change; inaccuracies "occasionally."

10 UCLA Health. UCLA study finds AI scribes may reduce documentation time and improve physician well-being (accessed August 2026).
https://www.uclahealth.org/news/release/ucla-study-finds-ai-scribes-may-reduce-documentation-time
Institutional release: 72,000 encounters, 41-second per-note saving, funding sources.

11 Afshar M, et al. A Pragmatic Randomized Controlled Trial of Ambient Artificial Intelligence to Improve Health Practitioner Well-Being. NEJM AI. 2025;2(12).
https://doi.org/10.1056/AIoa2500945
Stepped-wedge trial: exhaustion improved, fulfillment did not; no drift detected.

12 Chowdhury A, et al. Comparing ambient scribes: a randomized crossover clinical trial addressing ambient scribe technologies' impact on physician burnout. J Am Med Inform Assoc. 2026;33(5):990–999.
https://doi.org/10.1093/jamia/ocag018
160-clinician crossover: both tools cut burnout; 3.19 min/day between-tool gap; pajama time unaffected.

13 Kanaparthy NS, et al. Real-World Evidence Synthesis of Digital Scribes Using Ambient Listening and Generative Artificial Intelligence for Clinician Documentation Workflows: Rapid Review. JMIR AI. 2025;4:e76743.
https://doi.org/10.2196/76743
6 of 1,450 studies met inclusion; evidence "sparse."

14 Rotenstein LS, et al. Changes in Clinician Time Expenditure and Visit Quantity With Adoption of Artificial Intelligence–Powered Scribes: A Multisite Study. JAMA. 2026;335(16):1408–1417.
https://doi.org/10.1001/jama.2026.2253 · open-access full text: https://pmc.ncbi.nlm.nih.gov/articles/PMC13044793/
8,581 clinicians; 13.4/16.0 fewer minutes; 0.49 extra visits; 32% heavy users; $167.37/month.

15 Olson KD, et al. Use of Ambient AI Scribes to Reduce Administrative Burden and Professional Burnout. JAMA Netw Open. 2025;8(10):e2534976.
https://doi.org/10.1001/jamanetworkopen.2025.34976
Six-system pre-post: burnout 51.9% → 38.8%; includes Abridge-affiliated coauthor.

16 Holmgren AJ, et al. Ambient Artificial Intelligence Scribes and Physician Financial Productivity. JAMA Netw Open. 2026;9(1):e2553233.
https://doi.org/10.1001/jamanetworkopen.2025.53233 · open-access full text: https://pmc.ncbi.nlm.nih.gov/articles/PMC12789954/
UCSF cohort: +1.81 RVU/week ≈ $3,044/year; no denial change; upcoding question flagged.

17 Asgari E, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. npj Digit Med. 2025;8:274.
https://doi.org/10.1038/s41746-025-01670-7
12,999 sentences: 1.47% hallucination, 3.45% omission; rates improvable via workflow.

18 Palm E, et al. Assessing the quality of AI-generated clinical notes: validated evaluation of a large language model ambient scribe. Front Artif Intell. 2025;8:1691499.
https://doi.org/10.3389/frai.2025.1691499
Hallucinations in 31% of ambient vs 20% of gold notes; Suki-affiliated authors.

19 Anderson TN, et al. Evaluating the Quality and Safety of Ambient Digital Scribe Platforms Using Simulated Ambulatory Encounters. Mayo Clin Proc Digit Health. 2025;3(4):100292.
https://doi.org/10.1016/j.mcpdig.2025.100292 · open-access full text: https://pmc.ncbi.nlm.nih.gov/articles/PMC12605248/
Five-platform simulation: 26.3% note error rate; 76.3% omissions; 3.0 harmful errors/case; more than 90 platforms.

20 Reddy A, 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(6):765–772.
https://doi.org/10.7326/ANNALS-25-02772
VA evaluation: 11 tools; human notes scored higher on all five cases.

21 Peterson Health Technology Institute. Leading Health Systems: AI-Powered Scribes Alleviate Clinician Burnout; Financial Impact Unclear. March 2025 (accessed August 2026).
https://phti.org/announcement/ai-scribes-reduce-clinician-burnout/
Assessment-body verdict on effectiveness vs financial evidence gaps.

22 Abridge. Abridge Secures $300M Series E Led by a16z. June 24, 2025 (accessed August 2026).
https://www.abridge.com/blog/series-e
Vendor disclosure: $300M round; 150+ enterprise health systems.

23 Ambience Healthcare. Ambience Healthcare Announces $243 Million Series C to Scale its AI Platform for Health Systems. July 29, 2025 (accessed August 2026).
https://www.ambiencehealthcare.com/blog/ambience-healthcare-announces-243-million-series-c-to-scale-its-ai-platform-for-health-systems
Vendor disclosure: round, scope, EHR integrations.

24 Nabla. $70M Series C: Just Getting Started. June 17, 2025 (accessed August 2026).
https://www.nabla.com/blog/70m-series-c
Vendor disclosure: $70M round, $120M total, 85,000 clinicians.

25 Suki. Suki Announces Strategic Investment from Zoom Ventures. January 30, 2025 (accessed August 2026).
https://www.suki.ai/press-releases/suki-announces-investment-from-zoom-ventures/
Vendor disclosure: $168M to date; product scope; 350 health systems and clinics.

26 Microsoft. Meet Microsoft Dragon Copilot: your new AI assistant for clinical workflow. Microsoft Cloud Blog, March 3, 2025 (accessed August 2026).
https://www.microsoft.com/en-us/microsoft-cloud/blog/healthcare/2025/03/03/meet-microsoft-dragon-copilot-your-new-ai-assistant-for-clinical-workflow/
Vendor announcement: capabilities and survey-based outcome claims (879-clinician Microsoft survey, July 2024).