



AIGW Policy Observatory
Global AI Governance & Workforce Transformation

Global AI Governance and Workforce Transformation Policy Observatory

AIGW HEALTH & WORKFORCE TRANSITION DEEP DIVE

Kaiser Permanente: When AI Enters Clinical Work, What Must Remain Human?

How ambient AI compressed documentation
while preserving physician review, patient
permission and clinical decision authority

3 October 2026

An abstract graphic in the bottom right corner consisting of several concentric, curved lines in shades of blue and purple, creating a sense of motion or a stylized 'C' shape.

CONTENTS

Contents

Executive Summary and Key Findings	03
------------------------------------	----

Why the Documentation Workflow Matters	07
--	----

From Proposed Notes to Governed Scale	09
---------------------------------------	----

Failure, Review and Human Authority	12
-------------------------------------	----

Adoption and Physician Work	16
-----------------------------	----

Quality Assurance and Measured Outcomes	20
---	----

Role Architecture and Governance	24
----------------------------------	----

Governance and Expanding Authority	28
------------------------------------	----

Evidence Boundaries and Research Priorities	34
---	----

Conclusion and Notes and References	39
-------------------------------------	----

EXECUTIVE SUMMARY

Executive Summary

Kaiser Permanente's deployment of ambient artificial intelligence offers an unusually useful example of AI entering a high-stakes professional workflow without being given the highest-stakes decision.

The technology listens to a patient-clinician conversation, converts the encounter into a transcript and proposed clinical note, and returns that proposed note for physician review and editing. In this use case, the AI does not diagnose the patient, recommend treatment or make the clinical decision. The physician remains responsible for reviewing and correcting the generated documentation before it becomes part of the medical record.[1][2] That boundary is what makes the case important.

Many discussions of "human-in-the-loop" AI treat human involvement as a single condition: either a person is present or not. Kaiser Permanente's deployment shows why that language is too imprecise for high-stakes work. Different kinds of authority can remain human:

- permission; • observation; • verification; • editing; • approval; • override; • final clinical judgment; • accountability.

In ambient documentation, the AI performs a meaningful part of the work — listening, transcribing, summarizing and first-pass note creation — while the physician retains responsibility for the clinical content that enters the record and for the care decision itself. The scale is substantial.

Between 16 October 2023 and 28 December 2024, 7,260 physicians in The Permanente Medical Group (TPMG) used ambient AI scribes in 2,576,627 patient encounters across specialties and locations. Of those physicians, 3,447 used the technology in at least 100 encounters.[1] The evaluation estimated aggregate savings of 15,791 hours of documentation time, equivalent to roughly 1,794 eight-hour working days, compared with nonusers over the year after implementation. [1] Those hours were estimated from differences in documentation time; they were not direct stopwatch measurements of 15,791 hours returned to patient care.

EXECUTIVE SUMMARY

The workload results were statistically significant but modest at the encounter level. Compared with nonusers, AI-scribe users experienced reductions in time spent on notes, time outside the 7 a.m.–7 p.m. workday and "pajama time." High-volume users saved more time per note than low-volume users, and the highest-volume third of users accounted for 89% of ambient AI activations.[1] The experience data were also favorable but should remain correctly scoped.

In a survey of 102 adult and family medicine physicians, 84% reported that AI scribes had a positive effect on visit interactions and 82% reported a positive effect on work satisfaction.[1] In a patient survey, 56% said the scribe had a positive effect on visit quality and the remainder reported no effect; 92% were comfortable with or neutral about AI use, while 8% reported some discomfort. [1] The deployment therefore provides strong evidence of:

- scaled adoption;
- documentation-time compression;
- reduced clerical burden;
- a redesigned physician workflow;
- human review of

- AI-generated documentation;
- patient permission;
- preserved physician clinical decision authority.

It does not establish that ambient AI improved diagnosis, treatment quality, mortality or other clinical outcomes. It also does not establish a workforce-reduction effect.

No reviewed source shows that Kaiser Permanente reduced physician staffing because of ambient AI scribes. The more defensible workforce interpretation is that the technology changes how physician time is allocated, not that it substitutes for the physician role.

This distinction is especially important because the organization itself framed the deployment as a workforce-sustainability intervention.

Documentation burden is a contributor to clinician burnout and turnover. Kaiser Permanente's leaders positioned ambient AI as a way to reduce administrative load, increase focus on patients and support physician well-being and retention.[1][3] The implementation also shows that preserving human authority does not mean simply adding a final manual click.

EXECUTIVE SUMMARY

Kaiser Permanente created a quality-assurance system around the technology. In a separate 10-week quality-assurance pilot supporting broader deployment, more than 1,000 physicians participated across roughly 63,000 patient encounters. Clinicians rated proposed notes, submitted free-text feedback and completed a modified Physician Documentation Quality Instrument. The organization used those data to modify training, provide feedback to the vendor, stage deployment and establish post-deployment monitoring.[2] The QA process identified both value and failure.

proposed notes were rated positively overall, but users also reported:

- omitted information; • difficulty distinguishing multiple speakers; • erroneous assumptions; • the need for physician correction.[2]

That is a crucial governance signal. The safety architecture does not depend on the model being perfectly accurate. It depends on: bounded AI authority + qualified human review + feedback + monitoring + escalation to improvement. This

Deep Dive therefore argues that the most useful lesson from Kaiser Permanente is not "AI scribes save physicians time." It is:

In high-stakes work, automation can be scaled when the organization is explicit about what the AI may do, what the human must still do, and how errors are detected before they become authoritative. AIGW calls this the Human Authority Stack. For any high-stakes workflow, leaders should separately identify: Permission → Review → Verification → Approval → Override → Final Authority → Accountability Not every workflow requires every layer. But "human in the loop" should never substitute for specifying them.

1. The AI performs documentation work, not clinical decision-making — Kaiser Permanente describes the ambient tool as capturing the encounter and producing a proposed note for physician editing. Its leaders explicitly state that the tool does not make care decisions or recommendations and that clinicians remain the medical decision-makers.[2] This is the central governance boundary of the case.

EXECUTIVE SUMMARY

2. Documentation was compressed at enterprise scale — Between October 2023 and December 2024, 7,260 TPMG physicians used ambient AI across 2,576,627 encounters. Estimated aggregate documentation-time savings were 15,791 hours, or about 1,794 eight-hour workdays.[1] The figure is an estimated workload effect, not evidence of equivalent staffing reduction.

3. Human review is operational, not symbolic — The AI-generated note is a proposed note. Physicians review and edit it before adding it to the medical record.[2][3] Quality-assurance feedback also documented cases in which AI output omitted information or made erroneous assumptions that physicians corrected.[2]

4. Patient agency is built into the workflow — Clinician use is voluntary, and physicians seek patient permission before recording the visit. Audio is not retained, according to Kaiser Permanente's implementation descriptions.[2] This adds a distinct form of human authority: the patient can refuse the recording context.

5. Benefits were concentrated among users who adopted the tool heavily — The highest-volume third of users generated 89% of ambient AI activations. High users saved about 0.7 minutes

per note relative to nonusers, compared with about 0.15 minutes for low users.[1] Adoption intensity therefore matters when interpreting average benefit.

6. Physician and patient experience was generally positive but not universal — Among 102 surveyed physicians, 84% reported improved visit interactions and 82% improved work satisfaction. Among 116 patients who recalled scribe use, 56% reported improved visit quality; 8% reported some discomfort with AI use.[1] Positive experience should not be translated into clinical-outcome evidence. 7. Safe scale required organizational infrastructure — Kaiser Permanente used:

- training; • peer support; • patient-facing materials; • quality assurance; • physician feedback; • vendor feedback loops; • staged deployment; • ongoing monitoring.[2][4]

The technology did not scale through model performance alone. 8. The strongest workforce effect is time reallocation, not job substitution — The evidence supports task compression and workflow redesign. It does not support physician layoffs, reduced physician staffing or substitution of the physician's clinical role.

CONTEXT

Why the Documentation Workflow Matters

1. Why This Case Matters — Health care exposes a weakness in generic AI workforce language. In many industries, people say: AI handles the routine work; humans handle the complex work. That statement is directionally useful but insufficient in medicine. A clinical workflow contains different forms of responsibility:

- collecting information; • recording information; • interpreting information; • forming a diagnosis; • recommending treatment; • discussing options; • obtaining consent; • making decisions; • documenting the reasoning; • signing the record.

These activities may occur in the same encounter but do not carry the same risk. Ambient AI can automate a substantial part of documentation without being granted authority to diagnose or prescribe. That makes the case a clean example of selective automation. The organization did not ask: Can the model participate in the encounter? It asked, more precisely: Which part of the encounter can the model perform while the physician remains responsible for clinical judgment and the medical record?

This is a more mature governance question. The lesson extends well beyond health care. In any

high-stakes profession — banking, law, audit, public administration, safety-critical operations — AI governance should begin with: Which authority is being delegated? not merely: Is a human still involved?

2. The Starting Point: Documentation as a Workforce Burden — The case begins with a problem that is administrative in form but strategic in workforce consequence. Clinical documentation is necessary. A medical record supports:

- continuity of care; • communication across teams; • billing and compliance; • legal accountability; • quality measurement; • future clinical reasoning.

But the burden of producing that record has become substantial.

Physicians often document during the patient encounter, after the patient leaves, during evenings or on days that were not scheduled for clinical work. Kaiser Permanente's research team explicitly tracked this burden through electronic-health-record measures including:

CONTEXT

- time in notes per appointment; • time in orders; • time outside 7 a.m.–7 p.m.; • "pajama time"; • work on unscheduled days.[1]

The problem is therefore not that documentation has no value. It is that the method of producing it can consume physician attention and time that might otherwise be available for:

- patient interaction; • clinical reasoning; • rest; • non-clerical professional work.

2. The work before ambient AI — A simplified traditional workflow is:

Patient speaks → physician listens → physician simultaneously types / takes notes → physician reconstructs the encounter afterward → physician completes the clinical note → physician signs the record The physician is both:

- the clinical decision-maker; • the primary documentation producer.

Those two roles compete for attention inside the same encounter. Ambient AI changes that allocation. It does not remove documentation. It changes who — or what — produces the first-pass note.

3. The Intervention: AI Produces a proposed note, Not an Authoritative Record — The core workflow is relatively simple. With patient permission, the ambient system listens to the clinical encounter. It then: 1. captures the conversation; 2. generates a transcript; 3. synthesizes the encounter into a proposed clinical note; 4. returns the proposed note to the clinician; 5. leaves the physician to review and edit before the note enters the medical record.[2]

Kaiser Permanente states that audio recordings are not retained in the workflow it deployed.[2] This design creates an important distinction between generation and authority. The AI generates content. The physician retains authority over whether that content is acceptable as part of the medical record.

THE WORKFLOW

The intervention: AI generates a proposed note, the physician decides

FIGURE 1.1

workflow evidence

Before → AI intervention → after

Workflow element	Before	AI intervention	After / human role
Listen to patient	Physician	Ambient tool also listens	Physician remains responsible for clinical interpretation ^[2]
Capture details	Physician types / recalls	AI transcribes and summarizes	Physician checks what was captured ^[2]
proposed note	Physician creates note	AI creates first-pass note	Physician edits / corrects ^[2]
Clinical reasoning	Physician	No clinical decision authority in this use case	Physician retains reasoning ^[2]
Diagnosis / treatment	Physician	Tool does not make care decisions or recommendations	Physician retains final clinical authority ^[2]
Record entry	Physician	AI prepares proposed note	Physician determines final note content ^[2]
Patient permission	Physician/patient interaction	Tool can be activated only with permission in deployment model	Patient retains agency over recording context ^[2]

Source: [2] Permanente Medicine / Kaiser Permanente [2]

Note: This documented use case concerns clinical documentation, not diagnosis or treatment recommendation.

DEPLOYMENT

3. AIGW interpretation — This is not "AI replacing the physician in documentation." It is better described as: AI taking first-pass production responsibility while the physician retains epistemic and professional responsibility. That distinction becomes essential when errors are possible.

4. Stage One: Rapid Adoption in Northern California — TPMG first enabled ambient AI technology in October 2023 for roughly 10,000 physicians and staff across diverse settings and specialties in Northern California.[4][5] The early deployment was unusually rapid. Within the first 10 weeks:

- 3,442 physicians had used the technology; • it had been activated in 303,266 patient encounters;
- 968 physicians had used it in at least 100 encounters; • one physician had used it in more than 1,200 encounters.[4]

The organization supported adoption through:

- multiple training formats; • at-the-elbow peer support; • patient-facing materials; • rapid-cycle

vendor upgrades; • ongoing monitoring.[4]

This matters because the adoption story is not only a technology story. A model can be technically capable and still fail operationally if clinicians do not trust it, understand it or fit it into their work. 4. Voluntary use as an adoption mechanism — Kaiser Permanente did not make clinician use compulsory. Physicians could decide whether the tool was useful for their practice and patients.[2] [14] That design choice had two consequences.

First, it allowed adoption to emerge from perceived utility rather than mandate. Second, it means evaluation data are influenced by self-selection. Physicians with the highest documentation burden may have had the strongest incentive to adopt. That is exactly what the later first-year analysis found.[1]

4. Evidence boundary — High adoption among physicians who perceive high value does not prove that the same benefit would apply equally to every specialty, workflow or clinician. The deployment is evidence of scalable use. It is not evidence of universal usefulness.

DEPLOYMENT

5. Stage Two: From Regional Pilot to Enterprise Quality Assurance — As Kaiser Permanente expanded ambient documentation more broadly, the organization created a formal quality-assurance process rather than treating early physician enthusiasm as sufficient evidence for scale.[2] A 10-week QA pilot supporting the broader deployment recruited more than 1,000 physicians and included roughly 63,000 patient encounters.[2] The QA team included:

- physicians; • administrators; • informaticists; • safety officers; • quality and risk leaders; • evaluators.[2]

This multidisciplinary structure matters because an AI documentation system can fail along several dimensions at once. A note can be:

- linguistically fluent but clinically incomplete; • accurate in one domain but wrong in another; • technically functional but difficult to use; • useful to clinicians but uncomfortable to patients; • efficient but biased; • safe in common encounters but weak in edge cases.

No single metric can capture all of those problems.

5. The feedback architecture — Kaiser Permanente

used several channels. Clinicians could:

- rate notes on a five-star scale; • provide free-text feedback; • participate in forums; • complete a modified Physician Documentation Quality Instrument measuring dimensions including accuracy, thoroughness and perceived bias.[2]

The organization analyzed more than 3,600 free-text comments and received 1,252 responses to the modified PDQI questionnaire.[2] The average PDQI score was 4.35 out of 5. Of rated encounters, 47% received five stars, 31% four stars, while 7% received one or two stars.[2] The point is not that the system was perfect. It was not. The point is that feedback was incorporated into:

- clinician training; • deployment sequencing; • vendor improvement; • post-deployment monitoring.[2]

5. AIGW interpretation — The quality system turns the clinician from passive end user into part of the AI governance infrastructure. The physician does not only consume the model's output. The physician helps identify where the model fails and supplies the feedback needed to improve the operating system around it.

HUMAN AUTHORITY

Failure, Review and Human Authority

6. What Failure Looked Like — The most useful governance evidence in the case is not the positive rating. It is the documented failure modes. Kaiser Permanente reported that some clinicians observed problems including:

- difficulty tracking multiple speakers;
- omitted information;
- erroneous assumptions;
- proposed notes requiring correction.[2]

These are not trivial defects. A clinical note can influence future care. An omission or incorrect statement can persist if it is not detected before the record becomes authoritative. That is why physician review matters. 6. Hallucination is not the only risk — Generative-AI safety discussions often focus on hallucination. Ambient documentation has a broader error surface. Possible failures include: 6. Omission — The patient said something important, but the generated note leaves it out.

6. Attribution error — The system assigns a statement to the wrong speaker. 6. Unsupported inference — The system transforms ambiguous conversation into a stronger clinical statement than was actually made. 6. Over-documentation — The system includes excessive or low-value material, creating note bloat. 6. Under-documentation — The note lacks important context or detail. 6. Mis-prioritization — The note captures information but obscures what is clinically important. 6. Bias or inequity — Performance may vary across accents, speech patterns, disabilities, languages or demographic groups.

Kaiser Permanente's first-year evaluation explicitly noted that it lacked sufficient linked demographic data to assess differential experiences and that questions remained about performance for patients with disabilities affecting speech.[1] 6. AIGW takeaway — The human control layer must be designed against the actual failure modes of the system.

HUMAN AUTHORITY

A generic instruction to "review the output" is weaker than a review process that tells clinicians what types of errors they should actively look for. 7. The Human Authority Stack — The Kaiser case suggests that high-stakes AI governance should be decomposed into distinct authority layers. AIGW proposes the following stack. 7. Layer 1 — Permission — Can the AI system participate in the interaction at all? In the Kaiser deployment, patients are asked for permission before the visit is recorded for ambient documentation.[2]

This is patient agency, not physician oversight. 7. Layer 2 — Observation — Can the human see what the AI produced? The physician receives the AI-generated proposed note. Observation is necessary but weak by itself. 7. Layer 3 — Verification — Is the human expected to assess whether the output is accurate and complete? Yes. Physician review and correction are directly evidenced.[2] 7. Layer 4 — Approval — Can the AI output become authoritative without human acceptance?

In the described workflow, the generated content remains a proposed note until physician

review/edit before entry into the medical record.[2] This is stronger than merely having a human nearby. 7. Layer 5 — Override — Can the human change or reject the AI output? Editing and correction are directly evidenced. The public sources do not provide a detailed enterprise-wide override taxonomy for every field or automated action. 7. Layer 6 — Final Authority — Who makes the consequential professional decision?

Kaiser Permanente explicitly states that clinicians and care teams remain the medical decision-makers and that the ambient documentation tool does not make care decisions or recommendations.[2] 7. Layer 7 — Accountability — Who remains professionally responsible for the outcome? The physician remains responsible for the clinical record and care decision. The public sources do not reduce professional accountability to the AI vendor. 7. Why the stack matters — A workflow can contain a human and still fail to preserve meaningful authority.

For example:

HUMAN AUTHORITY

- a human may review after an action has already occurred;
- a human may see an output but have no practical ability to challenge it;
- a human may approve hundreds of outputs too quickly for meaningful scrutiny;
- a human may formally sign while lacking the information needed to verify.

Therefore: human presence $\neq$ human authority. The authority stack gives organizations a more precise vocabulary. 8. Human Review Is Work — and It Has a Cost — Automation narratives often treat human review as if it were free. It is not. A physician reviewing an AI-generated note must still spend cognitive effort to determine:

- whether the patient's story was captured correctly;
- whether diagnoses are represented accurately;
- whether medications, plans and follow-up are correct;
- whether the note contains unsupported statements;
- whether the record reflects what actually occurred.

The AI can reduce first-pass note creation effort while creating a new verification workload. This is

why net time savings matter more than gross automated content.

8. The Kaiser evidence — The first-year evaluation found that high users saved about 0.7 minutes per note relative to nonusers, while low users saved about 0.15 minutes per note.[1] Across the full user population, the evaluation estimated 15,791 hours of time in notes saved.[1] Those results indicate that review burden did not eliminate the benefit overall. But they also show that the benefit was not a dramatic elimination of documentation work per encounter.

8. External evidence reinforces the point — A separate 2025 JAMA Network Open study at the University of Pennsylvania found ambient scribe use associated with 20.4% less time in notes per appointment and 30% less after-hours work, but clinicians reported mixed experiences with note quality and some found that editing requirements offset part of the time saved.[9]

HUMAN AUTHORITY

Human Authority and Physician Work

A 2026 randomized clinical trial across 238 outpatient physicians found different time effects across two ambient scribe products, demonstrating that benefits vary by tool and implementation.[10] 8. AIGW concept: verification economics — Every AI workflow has two productivity terms: automation benefit minus verification burden The relevant question is not: How much work did the AI generate? It is: How much reliable work remained after the human performed the required verification? For high-stakes AI, verification economics should be treated as a first-class performance metric.

9. Adoption Was Highly Uneven — and That Is Part of the Evidence — The first-year TPMG analysis shows that ambient AI adoption was concentrated. The top third of users by volume accounted for 89% of all activations.[1] Physicians with higher documentation burden before deployment were more likely to become heavy users. Adoption was especially high in:

- mental health; • primary care; • emergency medicine; • allergy; • cardiology.[1]

It was lower in specialties including:

- urology; • obstetrics/gynecology; • infectious disease.[1]

The exact reasons differ by specialty. Some workflows rely on:

- more templated documentation; • inpatient rather than outpatient care; • procedures with different note structures; • different patterns of conversation and examination.

9. Why this matters for workforce analysis — A system can be highly successful for some workers and almost irrelevant for others. Therefore "organization-wide deployment" should not be confused with "uniform work redesign." The workforce unit of analysis remains the specific workflow. For one physician, ambient AI may substantially reduce after-hours documentation. For another, editing the generated proposed note may feel slower than typing the note directly.[1] 9. AIGW takeaway — Averages can hide workflow fit.

Every AI productivity claim should eventually be segmented by:

WORKFLOW

- specialty; • task type; • encounter type; • user intensity; • baseline burden; • patient context.

The organizations most likely to gain are not always the organizations with the most sophisticated technology. They may be the ones that match the technology most carefully to the work. 10. The Physician Role After Ambient AI — The ambient scribe does not remove the physician role. It changes the composition of the physician's work. 10. Less direct production of documentation — The physician no longer has to create every sentence from a blank screen.

10. More verification of machine-produced documentation — The physician becomes responsible for checking a proposed note created by the system. 10. More potential attention available during the encounter — The tool is intended to reduce the need to divide attention between patient and keyboard. In the TPMG patient survey:

- 39% said their doctor spent more time speaking directly with them; • 56% reported no difference; •

5% reported less time; • 47% said the physician spent less time looking at the computer screen.[1]

These are patient perceptions, not direct time-motion measures. 10. Clinical judgment remains physician work — The ambient documentation system is not the clinical decision-maker in this deployment.[2] That preserves the core professional function. 10. Role architecture before and after — Before Physician = clinical listener + decision-maker + documentation documentation producer + reviewer + signer After Physician = clinical listener + decision-maker + AI-output verifier/editor + signer AI = conversation capture + transcript + first-pass synthesis + proposed note

10. AIGW interpretation — This is a case of administrative task compression inside a professional role. The most important role change is not that the physician becomes less skilled. It is that a greater share of physician value can be concentrated in:

- interaction; • judgment; • verification; • accountable decision-making.

WORKFLOW

That is augmentation in a strict sense. 11. Does Time Saved Become More Patient Care? – It is tempting to assume that any saved documentation time becomes additional clinical capacity. The evidence does not support that conclusion automatically. TPMG measured:

- reduced time in notes; • reduced time outside normal hours; • reduced pajama time.[1]

Those are real workload outcomes. But they do not tell us exactly how every saved minute was reallocated. Time could become:

- more patient interaction; • reduced after-hours work; • cognitive recovery; • other administrative work; • more visits; • non-work time.

The first-year study did not establish a direct causal chain from AI scribe use to higher patient throughput or improved clinical outcomes.[1] External research is similarly mixed.

A 2026 emergency-medicine study comparing ambient AI, human scribes and no scribe found reduced documentation time but no difference in clinical productivity measured as work RVUs per shift hour.[11] This is important. 11. AIGW rule: time saved is an intermediate outcome – The correct chain is:

AI use → documentation time changes → capacity becomes available → organization / physician decides where capacity goes → possible patient, workforce or business outcome The middle decision cannot be skipped. For workforce strategy, reduced after-hours work may itself be the desired outcome. Not every productivity gain should be converted into more output. In a burnout-sensitive profession, some productivity gain can rationally be converted into workload sustainability. 12.

The Kaiser case is important partly because it breaks a common AI productivity assumption.

WORKFLOW

Workforce Sustainability Is Different from Workforce Reduction

A technology can improve labor productivity without reducing the number of professionals. The organization repeatedly frames ambient documentation as a response to:

The organization cites clerical burden, burnout, work dissatisfaction and physician retention risk. [1][3][14]

In 2026, Permanente leaders also described the availability of advanced ambient AI as a recruiting differentiator for some physicians considering work at Kaiser Permanente.[12]

That is a different workforce mechanism from substitution. Substitution logic is the same output with fewer workers; sustainability logic is the same or better work with lower burden and stronger retention or attraction. The two can coexist in some settings, but the public evidence reviewed here supports the second much more strongly.

The evidence does not show physician layoffs caused by ambient AI, reduced physician headcount, fewer physician hires because of documentation automation, or redeployment away from physician roles.

The evidence does show physician time reallocation, reduced documentation burden, improved reported work satisfaction, and a potential retention and recruitment value proposition.

AIGW takeaway: AI workforce analysis should distinguish labor substitution from labor sustainability. A tool that reduces burnout risk may have workforce value even if headcount never falls.

Ambient AI is unusual because the technology enters the clinical encounter as a listener. That makes the patient's role in governance visible. Kaiser Permanente says physicians seek patient approval before using the ambient documentation tool.[2]

WORKFLOW

Patient Consent and Experience

The request for approval creates a consent checkpoint before the AI participates in the workflow.

Patient permission is more than privacy compliance. Patients may care about who is listening, where data goes, whether audio is stored, how the generated note is used, and whether refusing the AI affects care.

Kaiser Permanente says the audio is not retained in the deployed workflow and that patient data remain protected.[2]

Patient reactions were broadly favorable, but not unanimous. Among patients who recalled the tool being used:

56% reported a positive effect on visit quality; the remainder reported no effect.[1]

92% were comfortable or neutral; 8% reported some discomfort.[1]

The first-year authors explicitly noted the need to understand why some patients were uncomfortable and to continue educating patients and obtaining permission.[1]

AIGW interpretation: In AI governance, human authority is not always only employee authority. The affected external party may also have a meaningful control right. In this case, patient permission precedes physician supervision. That suggests a broader framework for high-stakes AI:

subject consent or notice, professional review, and organizational monitoring. Different humans may govern different parts of the system.

QUALITY & MEASUREMENT

Quality Assurance and Measured Outcomes

14. Quality Assurance as a Learning System — The Kaiser deployment shows that responsible AI implementation can be treated as a continuous organizational learning process. The QA system did not merely ask: Is the tool safe enough to launch? It also asked: How do we keep learning after launch? The process connected: user feedback → quality analysis → training changes → vendor feedback → product improvement → post-deployment monitoring.[2] This matters because generative AI systems can change through:

- vendor updates;
- model changes;
- new integrations;
- new specialties;
- new user behaviors.

A one-time validation cannot guarantee permanent performance.

14. Governance at scale — Kaiser Permanente later described a responsible-AI framework focused on patient safety and clinical impact that is used to evaluate AI tools across clinical settings.[13] Its current vendor requirements also state that AI systems used by Kaiser Permanente should be safe, reliable, equitable and trustworthy.[15] Those later governance mechanisms are broader than the original TPMG scribe case. They are useful as evidence that the organization institutionalized lessons from early deployments.

14. AIGW concept: governance learning loop — A high-stakes AI deployment should include: pre-deployment evaluation → limited rollout → structured feedback → error taxonomy → training / product correction → scaled rollout → continuous monitoring This is governance as an operating process, not a policy document.

QUALITY & MEASUREMENT

15. What the Evidence Says About Accuracy — and What It Does Not — Kaiser Permanente's QA reporting describes the generated documentation as largely accurate and well received, with an average modified PDQI score of 4.35 out of 5 in the QA sample.[2] The score is useful. It does not mean:

- every note was correct; • every specialty performed equally; • every clinically important fact was captured; • the technology is safe without physician review.

The documented low ratings and physician corrections make that clear.[2] 15. Accuracy has multiple dimensions — For clinical documentation, accuracy can mean:

- factual correctness; • correct attribution; • completeness; • correct medical terminology; • absence of unsupported inference; • appropriate emphasis; • consistency with the physician's actual assessment.

A single aggregate quality score can mask different failure modes. 15. The equity gap — The first-year TPMG study also acknowledged limited ability to assess differential experiences by sociodemographic group.[1] It specifically raised uncertainty about how well ambient AI captures encounters involving patients with disabilities that affect speech.[1] This is a crucial boundary. An enterprise tool can perform well on average and still create systematic problems for specific groups. 15. AIGW takeaway — Quality assurance should be disaggregated by:

- specialty; • language; • accent; • disability; • encounter type; • demographic context; • complexity.

A high average score is not the end of responsible deployment. 16. Measured Outcomes: Keep the Layers Separate — The case contains several kinds of evidence. They answer different questions.

MEASURED OUTCOMES

Measured outcomes: reach and workload

FIGURE 1.2

reported result

Reach, adoption and workload

Measure / evidence layer	Reported result	Boundary
Physicians using ambient AI — Adoption	7,260	TPMG, Oct 2023–Dec 2024 ^[1]
Patient encounters with use — Adoption / scale	2,576,627	Not unique patients; not clinical outcomes ^[1]
Physicians with ≥100 uses — Adoption intensity	3,447	Does not establish benefit by itself ^[1]
Share of activations from top-third users — Adoption concentration	89%	Highlights skewed use ^[1]
High-user note-time saving — Workflow / time	~0.7 min/note vs nonusers	Difference-in-differences estimate ^[1]
Low-user note-time saving — Workflow / time	~0.15 min/note vs nonusers	Smaller effect ^[1]
Aggregate estimated time in notes saved — Workforce / workload	15,791 hours	Estimated, not direct stopwatch measurement ^[1]
Equivalent 8-hour working days — Illustrative workload conversion	1,794	Not staff positions eliminated ^[1]

Source: [1] NEJM Catalyst Innovations in Care Delivery [1]

Note: Documentation-time effects are workload estimates and are not staffing reductions.

MEASURED OUTCOMES

Measured outcomes: experience and quality

FIGURE 1.3

reported result

Physician, patient and QA results

Measure / evidence layer	Reported result	Boundary
Physician survey positive visit interaction — Experience	84% of 102 respondents	Adult/family medicine survey; not all 7,260 users ^[1]
Physician survey positive work satisfaction — Experience	82% of 102	Self-reported ^[1]
Patient positive visit quality — Experience	56%	Small survey; remainder no effect ^[1]
Patient comfortable/neutral with AI — Acceptance	92%	8% expressed discomfort ^[1]
QA modified PDQI average — Documentation quality	4.35/5	QA pilot sample; not every note ^[2]
QA free-text comments reviewed — Governance / monitoring	>3,600	Feedback volume, not outcome ^[2]
QA patient encounters — QA scale	~63,000	Separate QA deployment context ^[2]

Source: [1] NEJM Catalyst Innovations in Care Delivery; [2] Permanente Medicine / Kaiser Permanente [1] [2]

Note: Survey and quality-assurance samples are distinct from the full first-year user population.

ROLE ARCHITECTURE

Measured Outcomes and Role Architecture

16. The most important measurement boundary — Do not write: Ambient AI saved 15,791 hours, therefore Kaiser Permanente could remove the equivalent of 1,794 physician workdays from staffing. The study does not establish that. The hours were estimated reductions in documentation time distributed across thousands of physicians. The evidence is about burden and time allocation. It is not a full-time-equivalent reduction analysis.

17. From Tool Adoption to Role Architecture — The Kaiser case can be mapped through the AIGW Transition Stack. 17. Layer 1 — Task — AI takes on:

- transcription; • summarization; • first-pass note first-pass note creation.

17. Layer 2 — Workflow — The sequence changes from: physician captures + writes to: AI generates a

proposed note → physician verifies / edits → physician finalizes. 17. Layer 3 — Role — The physician spends less time as a clerical producer and more of the documentation workflow as:

- reviewer; • verifier; • accountable editor.

The clinical role remains human. 17. Layer 4 — Capability — The technology increases the importance of:

- verification; • error recognition; • clinical judgment; • selective trust; • patient communication; • understanding when not to use the tool.

17. Layer 5 — Workforce — The strongest evidenced effect is:

- workload relief; • potential workforce sustainability.

ROLE ARCHITECTURE

No direct staffing response is established. 17. AIGW interpretation — This makes Kaiser an important counterexample to a simplistic automation narrative. AI can transform a role substantially without creating an immediate worker-flow event. The physician is neither replaced nor unchanged. The job becomes AI-mediated professional work. 18. What Must Remain Human? — The case gives a more precise answer than "judgment." Several human functions remain distinct. 18. 1. Clinical interpretation — The physician determines what the patient's symptoms, history and findings mean.

18. 2. Clinical decision-making — The physician decides diagnosis, treatment and care plan. The ambient tool is not intended to make those decisions in this use case.[2] 18. 3. Verification of the generated record — The physician checks whether the proposed note accurately represents the encounter. 18. 4. Correction — The physician changes inaccurate or incomplete output. 18. 5. Professional accountability — The record and care remain clinician responsibilities.

18. 6. Relationship and communication — The physician remains the professional in direct

relationship with the patient. The technology may change attention allocation, but it does not become the care relationship. 18. 7. Ethical judgment about use — The clinician decides whether the tool is appropriate for a particular encounter and seeks patient permission. 18. AIGW takeaway — "Human work" after automation should not be defined only by tasks the machine cannot perform.

Some work remains human because society assigns authority, accountability and relationship to a qualified person even when technology can technically assist. That is especially important in regulated professions. 19. The Difference Between Human Review and Human Authority — Review is often presented as a sufficient safety control. It may not be. A review process can become weak when:

- volume is too high;
- reviewers become over-reliant on AI;
- AI output appears fluent and credible;
- important omissions are difficult to notice;
- review expectations are unclear;
- the human lacks time to reconstruct the source interaction.

ROLE ARCHITECTURE

Role Architecture and Governance

Ambient documentation has one important advantage over some other AI workflows: the physician was physically present for the encounter. That means the reviewer possesses first-hand source context. The physician can compare the proposed note against:

- memory of the conversation; • clinical reasoning;
- examination; • existing patient information.

This is stronger than a reviewer who sees only the AI output after the fact. 19. AIGW concept: source-proximate review — Human review is stronger when the reviewer is close to the original source of truth. Strong review

- reviewer witnessed or produced the underlying evidence; • reviewer understands the domain; • reviewer has authority to reject or change output.

Weak review

- reviewer sees only the generated answer; • reviewer lacks context; • reviewer is under time pressure; • approval is procedural rather than substantive.

The Kaiser workflow is relatively strong on source proximity because the physician participated in the encounter. That does not eliminate automation bias. It improves the conditions for meaningful verification. 20. Adoption Design: Voluntary Use, Champions and Workflow Fit — The first-year evaluation suggests that adoption strategy was itself part of the intervention. Kaiser Permanente used:

- voluntary physician adoption; • physician champions; • multiple training channels; • peer support; • vendor iteration; • EHR integration improvements.[1][4]

ROLE ARCHITECTURE

The organization also switched vendors during the evaluation period as it expanded deployment. The second vendor offered direct EHR integration, and prior users tended to continue using the new product.[1] This creates an important distinction between: model capability and implementation discipline. A useful model with poor workflow integration can fail. A slightly better model may still fail if users do not understand when it adds value. 20. 1. Let use reveal where pain is highest — The most burdened specialties adopted heavily.[1]

20. 2. Reduce friction — Direct EHR integration made the workflow easier.[1] 20. 3. Preserve user choice — Voluntary use allowed physicians to decide when the tool fit a particular encounter.[14] 20. Boundary — Voluntary use also creates selection bias in evaluation. High satisfaction among adopters should not be interpreted as satisfaction among every eligible clinician. 21. A Broader Responsible-AI Architecture Emerged Around the Case — By 2025–2026, Kaiser Permanente was describing a broader governance architecture for clinical AI.

Its Federation Care Delivery Technology and Quality teams developed a responsible-AI

framework focused on patient safety and clinical impact, with ambient documentation among the technologies used to develop and apply the framework.[13] Kaiser Permanente's current supplier requirements state that AI tools should be:

- safe; • reliable; • equitable; • trustworthy.[15]

Permanente leaders have also emphasized governance, transparency and accountability as prerequisites for trust in clinical AI.[16] These later mechanisms should not be retroactively treated as the exact governance framework used at every point in the original 2023 deployment. They are better interpreted as institutionalization after scale. 21. AIGW interpretation — A mature AI organization may evolve through three governance stages: Stage 1 — Use-case governance

- Can this tool be deployed safely in this workflow?

Stage 2 — Portfolio governance

- Can the organization apply common standards across many AI systems?

Stage 3 — Supplier / ecosystem governance

GOVERNANCE

- Can expectations be embedded into procurement, monitoring and vendor relationships?

Kaiser Permanente's ambient AI experience appears to have contributed to movement from Stage 1 toward Stages 2 and 3. 22. The Follow-On Frontier: When Documentation Starts Doing More — Ambient AI began as a documentation tool. The boundary may not stay there.

By 2026, Permanente leaders were discussing the potential for ambient systems to capture social determinants of health — such as food insecurity, housing instability or transportation barriers — more systematically when those issues arise during visits.[17] This is a logical extension of ambient listening. It is also a governance transition. 22. first-pass note creation is lower authority — The system summarizes what was said. 22. Identifying social needs is more interpretive — The system may decide that a statement is relevant to a defined risk category.

22. Triggering action is higher authority — If the system automatically routes patients, creates alerts or initiates interventions, the decision-right architecture changes again.

22. AIGW principle: authority must be re-reviewed when function expands — A system that is safe as a documentation assistant should not automatically inherit approval to become:

- a risk classifier; • a recommender; • a triage system; • an autonomous agent.

Every increase in function can increase authority. Governance should follow the capability boundary, not the product brand.

23. External Benchmark: Ambient AI Evidence Is Growing — and Still Uneven — Kaiser Permanente's experience is one of the largest documented deployments, but it sits inside a broader evidence base. 23. University of Pennsylvania — A 2025 JAMA Network Open quality-improvement study of 46 clinicians found ambient scribe use associated with:

GOVERNANCE

- 20.4% less time in notes per appointment; • 30% less after-hours work per workday; • higher same-day note closure; • lower perceived documentation burden.[9]

Clinicians also reported mixed note-quality experiences and sometimes substantial editing requirements.[9]

23. Multi-system burnout study — A 2025 JAMA Network Open study of 263 clinicians across ambulatory settings reported that after 30 days of ambient AI use, the proportion reporting burnout decreased from 51.9% to 38.8%, alongside improvements in cognitive task load and after-hours documentation.[18] This is not Kaiser Permanente evidence and should not be used as proof of Kaiser-specific burnout reduction.

23. Randomized clinical trial — A randomized trial involving 238 outpatient physicians found that one ambient scribe product significantly reduced time

in notes relative to control, while another did not show the same significant reduction, highlighting product and implementation heterogeneity.[10]

23. JAMA 2026 caution — TPMG researchers themselves wrote in JAMA in 2026 that ambient scribes may contribute to multiple dimensions of the health-care quintuple aim, but that evidence remains uneven across population outcomes, cost, patient experience, clinician well-being and equity.[19] 23. AIGW takeaway — The external literature strengthens three conclusions: 1. documentation-time and burden benefits are increasingly credible; 2. benefit size varies by implementation, product and user; 3. evidence on clinical outcomes, cost and equity is less mature.

This is exactly why case-study claims should remain narrow. 24. AIGW Framework: The Human Authority Stack for High-Stakes AI — The Kaiser case allows a general framework that can be reused across sectors.

AUTHORITY FRAMEWORK

AIGW framework: the human authority stack

FIGURE 1.4 framework mapped to the case
Seven layers of authority in the Kaiser workflow

Layer	Core question	Kaiser ambient AI example
Permission	May the system participate?	Patient permission before recording ^[2]
Observation	Can the human see the output?	Physician receives proposed note ^[2]
Verification	Must the human check accuracy?	Physician reviews generated documentation ^[2]
Approval	Can output become authoritative without human acceptance?	proposed note is edited before entry to record ^[2]
Override	Can the human reject/change output?	Physician corrections directly evidenced ^[2]
Final Authority	Who makes the consequential decision?	Clinician remains medical decision-maker ^[2]
Accountability	Who remains responsible?	Professional responsibility remains human ^[2]

Source: [2] Permanente Medicine / Kaiser Permanente [2]

Note: The stack is an AIGW analytical framework; the examples reflect the public description of this use case.

DECISION IMPLICATIONS

Governance and Leadership Implications

24. Banking — AI may proposed note a credit memo. Question: Who makes the lending decision?

24. Audit — AI may produce analysis. Question: Who signs off the audit judgment? 24.

Manufacturing — AI may flag a defect. Question: Who can override the classification or disposition?

24. Customer service — AI may resolve a request. Question: Which actions can occur without human escalation? 24. Public administration — AI may rank or recommend.

Question: Who holds legal authority over the final determination? 24. The critical rule — Do not ask only: Is there a human in the loop? Ask: Which layer of authority remains human, and is that authority meaningful in practice?

25. 1. Automate administrative production before professional judgment — High-value early use cases may be those that remove burden without transferring consequential decision authority. 25. 2. Treat generated content as a proposed note when error matters — The system should make clear when output is:

- suggestion; • proposed note; • recommendation; • decision.

Those categories should not blur. 25. 3. Budget for verification — Review is work. Include it in:

- productivity models; • staffing; • training; • workflow design.

25. 4. Measure error types, not only average quality — Track:

DECISION IMPLICATIONS

• omissions; • incorrect assumptions; • attribution errors; • note bloat; • specialty-specific failure; • equity-related performance.

25. 5. Preserve patient agency — Explain:

• what is being recorded; • what is retained; • what the AI does; • what the physician still decides.

25. 6. Build feedback into operations — Every correction can become evidence for:

• training; • model improvement; • vendor management; • workflow redesign.

25. 7. Reassess governance when functionality expands — A documentation system that later begins:

• recommending; • triaging; • ordering; • routing;

requires a new authority review. 26. Decision Implications for CHROs and Workforce Leaders — The case has direct workforce implications even without headcount reduction. 26. Measure work quality, not only labor cost — Useful outcomes include:

• after-hours work; • cognitive load; • burnout; • retention; • recruitment; • patient interaction.

26. Distinguish automation from deprofessionalization — A professional can use AI for lower-value administrative work while retaining — or increasing the relative importance of — professional judgment. 26. Train verification as a capability — The future professional may need to become exceptionally good at:

• challenging generated output; • detecting omission; • interpreting uncertainty; • knowing when not to trust automation.

DECISION IMPLICATIONS

26. Protect reflective practice — If note writing previously created a moment of reflection, leaders should ask whether automation removes a useful cognitive step. External research has raised this possibility.[9] The solution may include workflow practices that preserve clinical reflection without recreating unnecessary clerical work. 27. Define the authority boundary before procurement — A vendor evaluation should specify:

- what the model can generate; • what it cannot decide; • what requires human review; • what is logged; • what is retained; • what can be overridden.

27. Build QA before scale — Do not wait for broad deployment to define:

- error categories; • monitoring metrics; • feedback channels; • escalation paths.

27. Evaluate workflow fit by specialty — One adoption number can conceal large differences. 27. Maintain model-change governance — Vendor updates can change behavior. A governance programme should track:

- model version; • product changes; • new integrations; • performance drift.

27. Make users part of monitoring — Clinicians can provide high-value domain feedback, but the organization should not depend only on voluntary anecdotes. Structured monitoring is necessary. 28. Decision Implications for Policymakers and Regulators — The Kaiser case suggests several questions for high-stakes AI policy. 28. Regulate by function and authority, not only by technology label — A system that proposed notes documentation is different from one that recommends treatment. 28. Require clarity on human responsibility — Providers should be able to explain:

- who reviews; • who approves; • who can override; • who is accountable.

28. Monitor equity after deployment — Pre-deployment testing cannot capture every real-world population and workflow. 28. Preserve meaningful consent where appropriate — Patients should understand when ambient systems are listening and what happens to the data. 28. Encourage outcome-specific evidence — Efficiency evidence should not be presented as clinical-safety or health-outcome evidence. A mature policy environment should demand that claims remain tied to the outcome actually measured.

EVIDENCE BOUNDARIES

Evidence Boundaries and Research Priorities

29. What the Evidence Establishes — The reviewed evidence establishes that:

- The Permanente Medical Group enabled ambient AI documentation for large numbers of physicians beginning in October 2023. • In the first 10 weeks, 3,442 physicians used the technology across 303,266 patient encounters. • From October 2023 through December 2024, 7,260 TPMG physicians used ambient AI across 2,576,627 encounters. • 3,447 physicians used the system in at least 100 encounters. • The top third of users accounted for 89% of activations. • High users experienced larger reductions in time in notes than low users.
- The first-year analysis estimated 15,791 hours of documentation time saved in aggregate. • AI-scribe users experienced statistically significant reductions in time in notes, time outside the normal workday and pajama time relative to nonusers. • 84% of 102 surveyed physicians reported a positive effect on visit interactions. • 82% reported a positive effect on work satisfaction.
- 56% of surveyed patients who recalled scribe use reported a positive impact on visit quality, with the rest reporting no effect.

- 8% of surveyed patients expressed some discomfort with AI use. • Physicians review and edit AI-generated proposed notes before they enter the medical record. • Patient permission before ambient recording is directly documented. • Kaiser Permanente says audio is not retained in the deployed workflow described in its materials. • Kaiser Permanente explicitly states that the tool does not make care decisions or recommendations in this use case. • Clinicians and care teams remain the medical decision-makers.

- A formal QA programme evaluated documentation quality and user feedback during enterprise expansion. • The QA process documented omission, speaker-tracking and erroneous-assumption failure modes. • Kaiser Permanente used clinician feedback to modify training, vendor feedback and deployment. • The organization subsequently developed broader responsible-AI evaluation and vendor-governance mechanisms.

30. What the Evidence Does Not Establish — The reviewed evidence does not establish:

EVIDENCE BOUNDARIES

- that ambient AI improved diagnosis accuracy; • that it improved treatment outcomes; • that it reduced mortality or readmissions; • that it reduced total cost of care; • that every specialty benefited equally; • that every eligible physician adopted the tool; • that the 15,791 hours were directly observed rather than estimated; • that saved documentation time became additional patient visits; • that saved time produced an equivalent reduction in physician staffing; • that Kaiser Permanente laid off physicians because of ambient AI; • that ambient AI reduced physician hiring; • that every generated note was accurate; • that human review catches every AI error; • that physician review eliminates automation bias; • that patient acceptance is universal; • that the tool performs equally across languages, accents, disabilities or demographic groups; • that the QA average score applies to all millions of encounters; • that the human-authority architecture described for documentation applies automatically to future AI recommendation or agentic tools; • that the broader 2025–2026 responsible-AI framework was identical to the governance system used at initial rollout.

31. Research Gaps to Watch — 1. What share of generated note text is ultimately edited by physicians at scale? 2. Which error types are most clinically consequential? 3. How often do physicians reject an AI-generated note entirely? 4. Does review burden decline as clinicians gain experience with the tool? 5. Does heavy use increase or decrease automation bias? 6. Do physicians become less likely to notice omissions over time?

7. How does performance vary by specialty? 8. How does performance vary by visit complexity? 9. How does performance vary across languages and accents? 10. How does the technology perform for patients with speech-related disabilities? 11. What happens in multi-speaker encounters? 12. What percentage of patients decline ambient recording? 13. Why do some patients remain uncomfortable with the technology? 14. Does use affect physician-patient trust over time? 15. Does reduced documentation burden reduce burnout longitudinally at TPMG?

EVIDENCE BOUNDARIES

16. Does it improve physician retention? 17. Does it affect recruitment acceptance rates? 18. Does it change visit volume? 19. Does it change patient outcomes? 20. What is the full economic return after license, integration, QA and review costs? 21. How much time does QA and monitoring require? 22. How are vendor model updates revalidated? 23. What threshold of error triggers product restriction or rollback? 24. Which functions can ambient AI add before the governance classification should change?

25. How are social-needs identification and downstream routing governed? 26. When should an ambient system be treated as clinical decision support rather than documentation support? 27. Can decision-right taxonomies be standardized across health systems? 28. How should regulators distinguish proposed note-generation AI from autonomous clinical AI? 32. Method and Source Quality — This Deep Dive separates five evidence layers.

32. A. TPMG first-year operational evaluation — The strongest quantitative source is the 2025 NEJM Catalyst commentary, Ambient Artificial Intelligence Scribes: Learnings after 1 Year and over 2.5 Million Uses.[1] It provides:

- adoption scale; • encounter volume; • user intensity; • difference-in-differences workload metrics; • estimated aggregate documentation-time savings; • physician survey evidence; • patient survey evidence; • implementation limitations.

The analysis is especially valuable because it includes electronic-health-record workload data rather than relying only on self-report. It remains an observational organizational evaluation. Adoption was voluntary, heavy users had higher baseline documentation burden, and the deployment changed vendors during the study period. The authors used within-physician difference-in-differences to account for secular change, but the study is not a randomized clinical trial. 32. B. Early TPMG implementation evidence — The 2024 NEJM Catalyst commentary documents the initial Northern California rollout. [4]

It provides:

- 10,000-person enablement scale; • 3,442 physician users; • 303,266 encounters; • training and support architecture; • rapid early adoption; • early physician and patient reactions.

It is useful for understanding implementation mechanics. It should not be confused with the later first-year outcome analysis.

EVIDENCE BOUNDARIES

32. C. Kaiser Permanente enterprise quality-assurance evidence — The 2025 NEJM AI case study and Kaiser Permanente's detailed summary describe the QA process used during wider deployment.[2][6] They provide evidence of:

- multidisciplinary evaluation; • structured note ratings; • modified PDQI use; • free-text feedback; • documented failure modes; • vendor feedback; • staged deployment; • post-deployment monitoring.

These sources are central to the governance analysis.

32. D. Later governance and institutionalization evidence — Permanente Medicine and Kaiser Permanente vendor materials from 2025–2026 provide evidence that the organization formalized a broader responsible-AI framework and supplier expectations.[13][15][16] These sources are used as follow-on governance evidence. They are not retroactively treated as the exact initial rollout

framework. 32. E. External ambient-AI research — Independent and external studies are used to benchmark:

- documentation-time effects; • editing burden; • burnout; • product heterogeneity; • clinical productivity.

These studies include JAMA Network Open and randomized-trial evidence.[9][10][11][18] They are not used to claim Kaiser-specific outcomes. 32. F. AIGW interpretation — The following concepts are AIGW analytical constructs, not labels used by Kaiser Permanente:

- Human Authority Stack; • source-proximate review; • verification economics; • governance learning loop; • selective automation as role architecture.

They are derived from the evidence but should remain identified as interpretation.

EVIDENCE BOUNDARIES

33. Safe claim — The Permanente Medical Group scaled ambient AI documentation to more than 7,000 physicians and 2.5 million encounters, reducing measured documentation burden while keeping AI output in proposed note form for physician review. In this use case, clinicians remain the medical decision-makers and patients are asked for permission before ambient recording. 33. Safe workforce interpretation — The strongest workforce effect is time reallocation and reduced administrative burden, not documented physician substitution.

33. Safe governance interpretation — The deployment demonstrates one practical model for bounded AI authority: the system generates a proposed note, the physician verifies and finalizes, and the physician retains clinical decision authority. 33. Avoid — Kaiser replaced medical scribes or physicians with AI. The reviewed case does not establish that workforce outcome. 33. Avoid — AI saved 1,794 physician jobs' worth of time. The 1,794 figure is a conversion of estimated documentation hours into eight-hour working days, not full-time positions.

33. Avoid — The AI improved clinical outcomes. The core studies measure workload, adoption, documentation quality and experience — not diagnosis or treatment outcomes. 33. Avoid — The

AI is safe because humans review it. Human review is an important control, but documented errors, automation bias and uneven performance remain relevant risks. Safe external formulation — A concise evidence-safe description of the case is:

Safe external formulation — The Permanente Medical Group scaled ambient AI documentation to 7,260 physicians and more than 2.5 million encounters, with an estimated 15,791 hours of documentation time saved over the evaluation period. The AI system produces proposed note clinical documentation that physicians review and edit before it enters the medical record; Kaiser Permanente states that the tool does not make care decisions or recommendations and that clinicians remain the medical decision-makers. The deployment also used patient permission, formal quality assurance, structured feedback and ongoing monitoring.

The evidence supports task compression and workflow redesign, not physician substitution or improved clinical outcomes.

Safe external formulation — The central AIGW interpretation is: Safe external formulation — High-stakes AI should be governed by explicit authority allocation, not by the vague label "human in the loop."

CONCLUSION

Conclusion

Kaiser Permanente's ambient-AI deployment is one of the clearest available examples of how high-stakes professional work can be automated selectively rather than indiscriminately. The system performs real work. It listens. It transcribes. It synthesizes. It proposed notes. At millions of encounters, those functions materially changed physician documentation work and reduced measured administrative burden. But the organization did not give the system the full authority of the professional role. The physician still:

- interprets the patient;
- makes the clinical decision;
- reviews the generated documentation;
- corrects errors;
- determines what enters the medical record.

The patient also retains a form of agency through permission before ambient recording. That structure matters more than the phrase "human in the loop." A human can be present in a workflow without holding meaningful authority. The Kaiser case is valuable because the human role can be described more precisely: patient permission → AI proposed note → physician verification/edit → physician finalization → physician clinical authority. The deployment also shows that human authority is not enough by itself.

Safe scale required an organizational system around the technology:

- training;
- workflow integration;
- clinician feedback;
- structured quality assurance;
- vendor

feedback; • patient communication; • ongoing monitoring.

The strongest lesson is therefore not that health care should keep AI away from high-stakes work. It is that high-stakes automation requires a deliberate allocation of authority. AI can remove clerical burden without inheriting the professional judgment attached to the role.

And where an AI system does begin to move from first-pass note creation toward recommendation, classification or autonomous action, the authority boundary must be reconsidered rather than assumed. For workforce strategy, the case adds an important archetype to the AIGW portfolio. Salesforce shows AI changing staffing requirements. DBS shows AI reshaping operating models. PwC shows AI compressing apprenticeship. LONGi and JD show domain experts moving into AI-training and system-operations work.

Kaiser Permanente shows something different: AI can transform professional work while leaving the professional role intact — if the organization deliberately separates task delegation from decision authority. That is likely to become one of the defining governance questions of the AI-enabled workplace. The future of human work will not be determined only by what AI is capable of doing. It will also be determined by what institutions decide AI is allowed to decide.

SOURCES

Notes and References

1. NEJM Catalyst Innovations in Care Delivery. *Ambient Artificial Intelligence Scribes: Learnings after 1 Year and over 2.5 Million Uses*. 2025. First-year organizational evaluation / commentary with EHR workload analysis and surveys.
Supports: 7,260 physicians; 2,576,627 encounters; 3,447 physicians with ≥100 uses; 89% activation concentration among high users; high/low user documentation-time effects; 15,791 estimated hours saved; 1,794 eight-hour days; physician and patient survey results; specialty adoption; limitations and equity gaps.
<https://catalyst.nejm.org/doi/full/10.1056/CAT.25.0040>
2. Permanente Medicine / Kaiser Permanente. *Quality assurance informs large-scale use of ambient AI clinical documentation*. 26 March 2025. First-party summary of NEJM AI quality-assurance case study.
Supports: voluntary clinician use; patient permission; transcript + proposed summary; physician editing; audio not retained; tool does not make care decisions/recommendations; clinicians remain decision-makers; 10-week QA pilot; >1,000 physicians; ~63,000 encounters; 3,600+ comments; 1,252 PDQI responses; 4.35/5 average; failure modes; vendor feedback.
<https://permanente.org/quality-assurance-informs-large-scale-use-of-ambient-ai-clinical-documentation/>
3. Permanente Medicine. *How AI is giving physicians more time for what matters most*. 2025. First-party workforce / physician-wellness interpretation.
Supports: physician documentation burden, workforce sustainability framing, voluntary adoption, patient focus, estimated working-day equivalence.
<https://permanente.org/how-ai-is-giving-physicians-more-time-for-what-matters-most/>
4. NEJM Catalyst / The Permanente Medical Group. *Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation*. 2024. Early TPMG implementation evaluation / commentary.
Supports: tool enabled for 10,000 physicians/staff; 3,442 physician users; 303,266 encounters in 10 weeks; 968 users with ≥100 encounters; training formats, peer support, patient materials, rapid-cycle upgrades and monitoring.
<https://divisionofresearch.kaiserpermanente.org/publications/ambient-artificial-intelligence-scribes-to-alleviate-the-burden-of-clinical-documentation/>

SOURCES

5. American Medical Association. *AI scribe saves doctors an hour at the keyboard every day*. 2024. Independent professional-association reporting based on TPMG rollout.
Supports: early implementation scale; vendor-selection considerations; patient consent; training; physician review responsibility.
<https://www.ama-assn.org/practice-management/digital-health/ai-scribe-saves-doctors-hour-keyboard-every-day>
6. NEJM AI. *Quality Assurance during the Rapid Implementation of an AI-Assisted Clinical Documentation Support Tool*. 2025. Quality-assurance case study.
Supports: risk-based QA, multidisciplinary evaluation, crowdsourced feedback, training modification, vendor feedback, deployment staging and transition to ongoing monitoring.
<https://doi.org/10.1056/AIcs2400977>
7. Kaiser Permanente Division of Research. *AI-assisted notetaking gains steady support from Kaiser Permanente physicians*. 1 April 2025. First-party research summary.
Supports: 7,260 physicians; 2,576,627 encounters; 3,447 heavy-use physicians; adoption patterns; reduced documentation and after-hours burden.
<https://divisionofresearch.kaiserpermanente.org/ai-assisted-notetaking-gains-steady-support-from-kaiser-permanente-physicians/>
8. American Medical Association. *AI scribes save 15,000 hours—and restore the human side of medicine*. 12 June 2025. Independent professional-association reporting.
Supports: 15,791 estimated hours; 1,794 working-day equivalence; 7,260 physicians; 2,576,627 encounters; distinction from clinical decision support; physician editing of notes.
<https://www.ama-assn.org/practice-management/digital-health/ai-scribes-save-15000-hours-and-restore-human-side-medicine>
9. JAMA Network Open. *Clinician Experiences With Ambient Scribe Technology to Assist With Documentation Burden and Efficiency*. 2025. External prospective pre-post quality-improvement study.
Supports: 20.4% reduction in time in notes; 30% reduction in after-hours work; mixed qualitative note-quality feedback; editing burden; possible loss of reflective note-writing; need for patient-outcome evaluation.
<https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2830383>

SOURCES

10. External randomized clinical study. *Ambient AI Scribes in Clinical Practice: A Randomized Trial*. 2025. External randomized evidence.
Supports: heterogeneous documentation-time effects across two commercial ambient AI products; importance of product and implementation differences.
<https://pubmed.ncbi.nlm.nih.gov/41497288/>

11. Annals of Emergency Medicine. *Medical Scribe and Ambient Artificial Intelligence Impact on Emergency Physician Documentation Burden and Clinical Productivity*. 2026. External observational comparison.
Supports: reduced documentation time with ambient AI and human scribes; no observed difference in clinical productivity measured as wRVUs per shift hour.
<https://pubmed.ncbi.nlm.nih.gov/42283666/>

12. Permanente Medicine. *Podcast: AI, innovation, and value-based care in medicine*. 2025. First-party executive discussion.
Supports: national ambient-AI scale; clinician review/edit before permanent EHR entry; physician-workforce recruitment value as described by leadership.
<https://permanente.org/podcast-ai-innovation-and-value-based-care-in-medicine/>

13. Permanente Medicine. *An eye on innovation to enhance care delivery*. 23 March 2026. First-party annual-report governance update.
Supports: Federation Care Delivery Technology + Quality responsible-AI framework; patient safety and clinical impact focus; ambient AI as an early evaluation use case.
Boundary: follow-on governance evidence, not the exact 2023 rollout framework.
<https://permanente.org/an-eye-on-innovation-to-enhance-care-delivery/>

14. Permanente Medicine. *Lessons learned from the Kaiser Permanente rollout of ambient AI scribes*. 20 March 2025. First-party implementation reporting.
Supports: patient consent, physician-workload goals, responsible implementation lessons and broad system rollout.
<https://permanente.org/lessons-learned-from-the-kaiser-permanente-rollout-of-ambient-ai-scribes/>

SOURCES

External ambient-scribe studies provide context; they do not validate Kaiser Permanente-specific outcomes. The supplier requirements and later innovation pages are governance context, not outcome evaluations of the 2023–2024 deployment.

15. Kaiser Permanente. *Artificial Intelligence Requirements for Vendors, Contractors and Suppliers*. n.d.. Formal organizational vendor-governance requirement.
Supports: organizational expectations that AI be safe, reliable, equitable and trustworthy; supplier governance.
Boundary: current governance framework, not evidence of the original ambient-scribe outcome.
<https://supplier.kp.org/requirements-guidelines/general/artificial-intelligence-requirements/>
16. Permanente Medicine. *Building trust in health care through better AI governance*. 28 July 2025. First-party executive governance discussion.
Supports: emphasis on governance, transparency, accountability, human-centered deployment and physician review/edit of ambient documentation.
<https://permanente.org/building-trust-in-health-care-through-better-ai-governance/>
17. Permanente Medicine. *How ambient AI can help identify patient social needs*. 27 August 2026. First-party follow-on innovation discussion.
Supports: future/expanded ambient-AI use for documenting social determinants of health and routing relevant information to care teams.
Boundary: Follow-on capability direction; not part of the original ambient-documentation outcome evidence.
<https://permanente.org/how-ambient-ai-can-help-identify-patient-social-needs/>
18. JAMA Network Open. *Use of Ambient AI Scribes to Reduce Administrative Burden and Professional Burnout*. 2025. External multi-clinician observational evidence.
Supports: reported reduction in burnout prevalence and documentation-related burden after ambient AI use across 263 clinicians.
<https://doi.org/10.1001/jamanetworkopen.2025.34976>
19. JAMA. *Ambient AI Scribes and the Quintuple Aim: What Is Counted—and What Matters*. 2026. Expert editorial from TPMG / Kaiser Permanente researchers.
Supports: ambient AI as rapidly emerging clinical infrastructure; evidence remains uneven across population outcomes, cost, patient experience, clinician burnout and equity; need for broader evaluation.
<https://jamanetwork.com/journals/jama/article-abstract/2847322>