Ambience

Pilot to Platform

A Framework to Evaluate Ambient AI

Abstract

Ambient listening tools have captured the attention of the market, but hype has outpaced results. Clinician burnout and documentation burdens are real, but without adoption and sustained utilization at the point of care, ambient AI remains an expensive experiment.

This guide offers a reality check. It presents a pragmatic framework to evaluate deployment health, demystify utilization metrics, and diagnose product shortcomings, from flawed clinical reasoning models to poor specialty support. It makes the case for encounter-level utilization as the first gate before the broader promise of ROI, and surfaces clear, measurable strategies employed by early adopters like the Cleveland Clinic, Houston Methodist,
and St. Luke's.

The winners in this space won't just summarize. They'll document, code, and transform healthcare's operating system from the point of care outward.

Contents

Moving Past Hype

Placing Ambient AI Under a Microscope

Ambient listening is everywhere, but results are mixed. Most systems are only using these tools in limited ambulatory deployments, with clinician satisfaction and efficiency taking center stage in 2024 and 2025. While vendors share plenty of success stories, the data, outcomes, and success measures vary significantly.

The first question every organization should ask: Are clinicians actually using the product?

To answer that, let's focus on two key metrics:

1. **Adoption:** The percentage of employed clinicians (across all specialties and settings) who become monthly active users (MAUs).

2. **Utilization:** The percentage of scheduled encounters where clinicians actively use the product, best measured directly via the EHR's native analytics modules.

So far, adoption is concentrated in primary care and family medicine. To understand why, we need to look at what's happening at the encounter level.

What does utilization look like across systems?

There's no consistent standard for how vendors report utilization. Terms like "available to X clinicians" or "X% user retention" can be misleading, giving the impression that a full health system is onboarded when, in reality, the deployment footprint may be nominal.

Fortunately, systems are sharing encounter-level vendor performance:

Stanford HEALTH CARE	Used in **55.25%** of 17,428 encounters[1]
Intermountain Health	Used in **47%** Median utilization among active users[2]
BJC HealthCare	Used in **45%** of patient encounters[3]
Providence	Used in **32%** of weekly visits[4]

These findings paint a more realistic picture: we're not seeing mass adoption, but the early musings of a new product category still finding its place in healthcare.

1. Ma SP, Liang AS, Shah SJ, et al. Ambient artificial intelligence scribes: utilization and impact on documentation time. J Am Med Inform Assoc. 2025;32(2):381-385. doi:10.1093/jamia/ocae304
2. Haberle T, Cleveland C, Snow GL, et al. The impact of Nuance DAX ambient listening AI documentation: a cohort study. J Am Med Inform Assoc. 2024;31(4):975–979. doi:10.1093/jamia/ocae022
3. Nelson M. Project Spotlight: Clinician Wellness with Abridge Ambient Notes. Healthcare Innovation Lab. Published March 7, 2025.

Utilization

Diagnosing a Deployment's Effectiveness

Why does encounter-level utilization matter?

Ambient AI has the potential to impact documentation, billing, quality, and broader workflows, but only if it's actually active during patient visits. Low utilization also dramatically drives up cost-per-encounter. For example, consider these hypothetical vendors:

Vendor 1: 80% utilization / **Vendor 2: 40% utilization**

If both vendors charge $100 per clinician per month, and each clinician sees 100 patients monthly, Vendor 2 ends up twice as expensive on a per-encounter basis. Even if Vendor 1 raised prices 50%, it would still offer significantly better value

How can I measure utilization?

Start simple: Utilization = (Encounters where the tool was used) ÷ (Total scheduled encounters) Exclude no-shows, cancellations, and cases where the tool was used for less than a minute.

Use your EHR's native analytics, like Epic's Signal or User Action Logs (UAL), to build this measurement into a dashboard. Every vendor should be helping you do this to track the health of your deployment, well beyond utilization.

Why isn't my solution being used?

Low adoption or utilization usually stems from two things:

1. **Deployment gaps**
2. **Product failure modes**

On the deployment side, some specialties, especially procedural ones like dermatology, simply don't struggle with documentation as much. These clinicians may need extra support to understand how ambient tools fit into their workflows.

That said, low encounter-level utilization is usually indicative of a deeper issue: product design. A common misconception is that ambient AI is a summary tool. That's fundamentally false. An effective system must understand the medicine being practiced. It needs clinical reasoning, not just transcription, to build useful, accurate documentation that fits the workflow and specialty.

Product Design

The Importance of Specialty-Tuned Clinical Reasoning

To work well for all clinicians, ambient platforms need clinical reasoning systems tailored to the specialties they're used in. Take oncology, where much of the documentation burden falls outside of the visit itself. That means AI needs access to longitudinal patient history and, more importantly, it must have the capability to make sense of it in the context of the visit. Let's consider an example:

General

Patient underwent lumpectomy and sentinel lymph node biopsy for breast cancer. The tumor was 6 millimeters, hormone-positive, and HER2-negative. Patient had a history of ovarian cancer in 2007, treated with carboplatin and Taxol. Patient is currently awaiting results to determine the need for chemotherapy. They have an upcoming appointment with a radiation oncologist to discuss radiation therapy options. Patient reports a sore throat and dry cough for an extended period, with a history of GERD.

We see two solutions. One that's built for oncology, and one that is not. In the "specialty-naive" solution, we're left with insufficient detail for comprehensive oncological care, like:

Issues

1. Specific dates and results of diagnostic evaluations.
2. Thorough oncological history, including previous ovarian cancer diagnosis and treatment specifics.
3. Interval history providing insight into current symptoms, clinical progression, and recovery status.
4. Essential diagnostic information (e.g., tumor size, Ki-67 index) is necessary for informed clinical decision-making and prognosis.

Subspecialty-Tuned Breast Oncology HPI

Oncological History:

Patient was diagnosed with a 7mm mass in the upper inner right quadrant of the breast on 10/11/2023 during a routine screening mammogram. A follow-up ultrasound and diagnostic mammogram on 11/17/2023 confirmed a 6mm round circumscribed mass in the upper inner right breast. A subsequent right breast ultrasound- guided biopsy revealed infiltrating ductal carcinoma. Patient underwent a localized lumpectomy and sentinel lymph node biopsy on 01/23/2024. The patient also has a history of ovarian cancer diagnosed in 2007, for which she underwent a hysterectomy and oophorectomy, followed by eight cycles of Carboplatin and Taxol chemotherapy

Interval History:

Patient is a 34 year old female presenting for follow-up after surgery for infiltrating ductal carcinoma of the right breast. She reports occasional sharp pain at the surgical site but is otherwise healing well. Patient also reports a persistent sore throat and dry cough, which she has been experiencing since recovering from COVID-19 a few weeks ago. She is also taking Prilosec daily for suspected GERD symptoms.

Diagnostic Results:

- 10/11/2023: Screening mammogram revealed a 7mm mass in the upper inner right quadrant of the breast.

- 11/17/2023: Ultrasound and diagnostic mammogram confirmed a 6mm round circumscribed mass in the upper inner right breast. Subsequent breast ultrasound-guided biopsy revealed infiltrating ductal carcinoma.

- 01/23/2024: Lumpectomy and sentinel lymph node biopsy performed. Pathology report confirmed tumor size was 8mm x 6mm x 6mm and the sentinel lymph node was negative. Tumor was hormone-positive, with positive estrogen and progesterone receptors, and negative HER2 receptors. Ki-67 index was reported as 25%, indicating an intermediate level of proliferation.

Further reinforcing the point, at Penn Medicine, clinicians reported that poorly structured notes led them to rewrite documentation themselves.[4] That erodes any perceived efficiency benefits, and ultimately, cripples utilization.

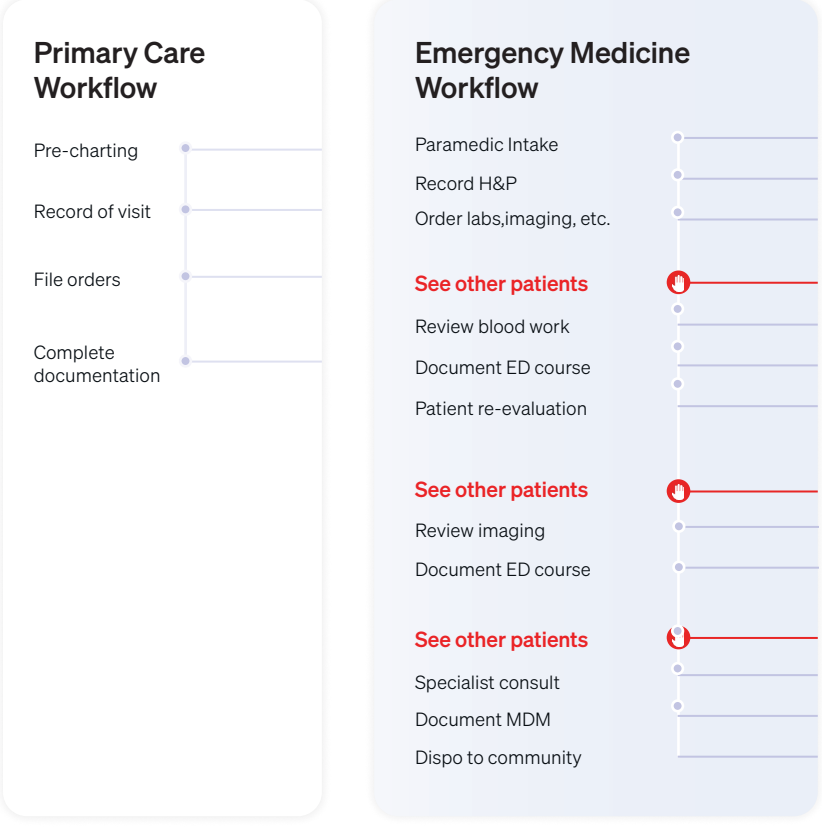
4. Duggan MJ, Schoenbaum A, Hanson CW III, et al. Clinician experiences with ambient scribe technology to assist with documentation burden and efficiency. JAMA Network Open. 2025;8(7):e24 060637. doi:10.1001/jamanetworkopen.2024.60637

Product Design

Developing Systems for Complex Workflows

Each specialty has its own rhythm and requirements. A one-size-fits-all tool will fall short. For example, emergency medicine involves fast-paced, fragmented interactions with high documentation demand. A tool that simply “listens and summarizes” can’t handle:

- Overlapping visits and multitasking
- Noise from multiple providers and settings
- Onerous EHR steps like order entry that delay throughput
- Revenue cycle awareness necessary to structure a comprehensive MDM



To succeed here, the system must do more than transcribe. It must reconcile all inputs into a defensible clinical and billing record, live, at the point-of-care.

Case Study

Defining “Good” Utilization

Because the market is early and fragmented, most health systems don’t have a baseline for what strong utilization looks like. That said, there are some promising emerging benchmarks. The Cleveland Clinic ran a structured, multi-vendor evaluation, piloting three ambient vendors side-by-side and tracking encounter-level utilization across the pilot panel.

The differences were stark, with a direct relationship emerging between utilization and clinician satisfaction:

Figure 1

Encounter-Level Utilization (Pilot)

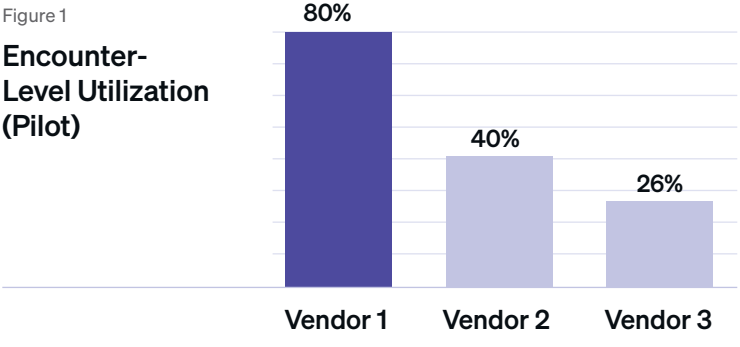
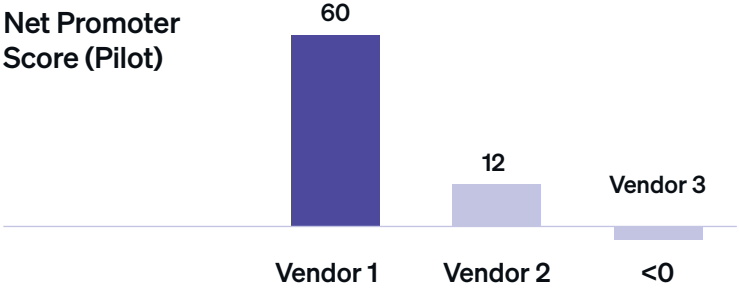


Figure 2

Net Promoter Score (Pilot)



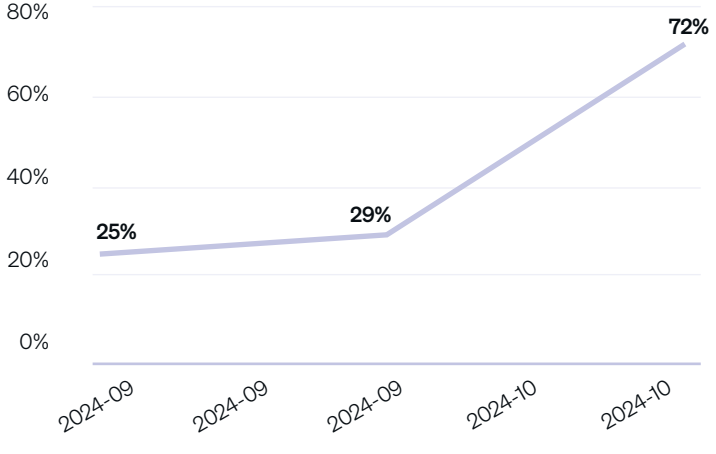
Case Study

Defining “Good” Utilization

What set Vendor 1 apart was its ability to consistently perform across more than 40 specialties (during the pilot). That success came from an underlying ensemble of specialty-tuned clinical reasoning models, and fast-moving product, engineering, and research teams that continued to enhance product performance mid-pilot. For example, tuning conducted specifically for cardiology unlocked even higher encounter-level utilization for the specialty within weeks:

Figure 3

Encounter-Level Utilization (Pilot)



After the pilot, the Cleveland Clinic moved forward exclusively with Vendor 1. In just four months, they rolled the platform out to over 4,000 ambulatory providers, sustaining >75% encounter-level utilization across a highly complex, multi-specialty environment, with the solution’s NPS rising to 80.

The Multi-Vendor Approach:

Many health systems feel pressured to choose one vendor and move quickly, creating a false sense of urgency and can lead to the wrong long-term partner. Instead, a more pragmatic strategy is a multi-vendor approach, especially for systems already mid-pilot or unsure whether to expand. Houston Methodist presents a strong example.

In 2023, they deployed an ambient tool for ambulatory care. When it came time to support emergency medicine and inpatient physicians, that existing vendor fell short. Rather than scrap what was working, Houston Methodist launched a parallel deployment. They added a specialty-tuned vendor focused on complex settings and left their ambulatory implementation untouched.[5] This allowed for:

1. Minimal disruption to satisfied users
2. Faster progress in under-served specialties
3. A clearer view of what different vendors could (and couldn't) do

Additionally, managing two ambient vendors in parallel proved surprisingly straightforward. Even though each vendor delivered different levels of performance, the underlying integration points and change management needs were similar across both. These results mirrored the Cleveland Clinic's experience, where managing multiple the five vendors was a lighter lift than expected for the IT and operational teams.

Moving From Adoption to Impact:

Clinicians are burned out and administrative overhead continues to proliferate. Ambient listening offers a glimmer of hope, and with that, the immense pressure for health systems to act, as outlined above.

Speed alone isn't strategy and acting too quickly with the wrong foundation can backfire. We believe early deployments are promising, but they're just the beginning. Real ROI comes from sustained, widespread use that unlocks the greater promise of the ambient platform, a new operating system for healthcare, working symbiotically with the system of record.

If these tools aren't aware of what happens at the point of care, their downstream potential for coding, quality, clinical research, and broader workflow transformation will never materialize.

Clinical documentation is more than a static reference point; it's the currency of healthcare. Ambient solutions are in a unique position to intercede and amplify revenue cycle performance upstream. Nearly all market participants are aligned around that vision, but few can articulate a clear path forward, and even fewer are driving real impact today.

How can I trust this technology?

Buyers are right to be skeptical. Just as utilization is impacted by the underlying clinical reasoning systems, revenue cycle accuracy and compliance depend on their own distinct system of intelligence.

The cost of cutting corners is already clear. At UW Health (Wisconsin), revenue cycle and compliance teams observed a 56% drop in ICD-10 coding accuracy after implementing an ambient tool.[6] That reduction required UW Health to divert IT resources, and if not handled, could have caused a cascade of more serious consequences, from denials to CMS investigations.

As with specialties, not all solutions are built the same. One vendor, through a deep development partnership with OpenAI, was able to outperform board-certified physicians in ICD-10 coding accuracy by 27%.[7] That same organization went as far as to audit the performance of their models across specialties and subspecialties, scoring between 90% to 100% accuracy across E/M coding suggestions, and achieving 95% accuracy for ICD-10 diagnoses mapping to HCCs, with an example from those audits available below on the next page.

Project Name: **Cardiology April Audit 2025**
Practice Reviewed: 1
Report Run Date: 04-25-2025
Number of Dates of Service Reviewed: 100

Project Detailed Report
EM Coding: Total Sample Size 100 codes

Findings	Count	%
EM/ Level appears to be correct	97	97%
EM/ Level appears to be over-coded	1	1%
EM/ Level appears to be under-coded	2	2%
EM/ Level appears to be missing, unsupported or wrong category	0	0%

Practices

Practice Name	Practice ID	%
Ambience Healthcare		97%

Providers

Practice Name	Practice ID	%
Providers, Cardiology	Ambience Healthcare	97%

The Beachhead - Outpatient CDI:

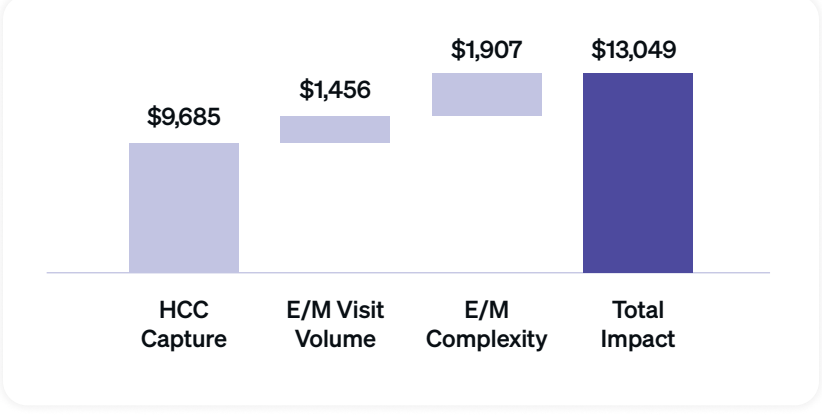
Leveraging “coding-aware” ambient revenue cycle solutions, the clearest immediate opportunity to drive impact is outpatient coding, an area where most systems under-invest. Unlike inpatient CDI, which benefits from dedicated staff and clearer financial returns, outpatient settings often go without robust coding support. Full-time CDI resources are rare, and outsourced software and services operating downstream can only be as effective as the notes they receive.
As a result, significant revenue is missed, not because providers didn’t deliver care, but because the documentation didn’t reflect it.

Case Study

"Coding-Aware" Ambient RCM in Practice

St. Luke's, the leading integrated delivery network in Idaho, began exploring ambient AI to address clinician burnout in 2023. Reid Stephan, their Chief Information Officer, realized a more ambitious opportunity to build a platform, leveraging the point-of-care access afforded by ambient listening as a wedge to enrich the data being written into Epic, and intelligently streamline longitudinal workflows. Reid's first step was to co-develop an outpatient coding and compliance suite.

Recently, St. Luke's and KLAS Research published the results achieved by their ambient partner.[8] The most shocking – direct revenue contribution:



St. Luke's also realized a 41% reduction in active documentation time, a 39% decrease in time spent documenting after-hours, and a 41% reduction in time to close charts. These efficiency metrics contribute to the compression of A/R days, improving St. Luke's cash position.

At John Muir Health (JMH), a suburban health system in the San Francisco Bay Area, The same ambient RCM platform drove a 7% improvement in coding complexity, translating to an additional $5,826 in annual revenue per clinician.[9]

Interestingly, this improvement occurred despite the presence of professional outpatient coding review services.

Results uncovered by JMH and Gartner shown below:

Gartner

Table 2: Return on Investment Value Domains, Measures and Results

Value Domains	Measures	Results
Clinician satisfaction	▪ Physician turnover rate ▪ Time spent documenting per day (or per appointment) ▪ Time spent documenting after hours ▪ Provider engagement scores	▪ Reduced PCP turnover rate by 44% (indicative cost savings of ~$3 million per year) ▪ Reduced average time spent in notes and patient instructions per day by 34 minutes ▪ Reduced provider burnout by 11%
Documentation quality	▪ Coding complexity (see Note 1)	▪ Increased coding complexity by 7% ▪ Increased annual revenue per clinician by $5,826

8. Boyter M, Blauer T. Ambience Healthcare ROI Validations 2025: How the pursuit of clinician well-being led to demonstrable ROI for St. Luke's Health System. KLAS Research. Published June 26, 2025.
9. Walk V. Case Study: Ambient Digital Scribes Accelerate Generative AI ROI for Healthcare Provider CIOs. Gartner, Inc. Published January 8, 2025.

The Checklist

10 Questions to Cut Through the Noise

As ambient solutions proliferate, it's easy to be impressed by polished demos and anecdotal wins, but real transformation comes from sustained clinical use and measurable financial performance. Whether you're mid-pilot, considering expansion, or just getting started, these 10 questions will help you pressure-test the true value and scalability of a solution.

Clinical Performance & Workflow Integration

1. **What percentage of scheduled encounters are the tool actively used in?**

   Why it matters: Encounter-level utilization directly impacts ROI and is the most honest metric of real-world performance.

2. **Does the product fully replace documentation or just assist?**

   Why it matters: Assistive tools requiring frequent rewrites erode value and stall adoption.

3. **How many distinct, specialty-tuned models are powering the system?**

   Why it matters: A single, generic model can't handle the documentation demands of diverse specialties like oncology or ED.

4. **How is the solution embedded across all care settings (ambulatory, ED, inpatient)?**

   Why it matters: Point-of-care performance varies dramatically across settings - one-size-fits-all doesn't scale.

5. **What structured pre-, in-, and post-visit integration points exist with the EHR?**

   Why it matters: Solutions must complement physician workflows, not create new burdens.

The Checklist

10 Questions to Cut Through the Noise

Financial Integrity & Revenue Impact

1. **Have query rates and misdiagnoses increased or decreased for clinicians using your solution?**

   Why it matters: "Coding-naive" solutions don't understand billing and compliance, creating issues for downstream RCM teams - make sure it isn't happening to you.

2. **How is the coding and compliance accuracy of the solution being measured?**

   Why it matters: Just like humans, this tech should be audited. Make sure your vendor partners with trusted third party auditors, and can show you the results.

3. **Can the vendor provide impact data on chart closure times, documentation lag, and other variables impacting A/R days?**

   Why it matters: These metrics don't just impact clinician satisfaction, they drive time-to-bill. Make sure they're measured empirically, using Epic's Signal and UAL data.

4. **Does the system use the diagnosis (ICD-10 selection) to drive the creation of the note, or generate notes before selecting diagnoses as an afterthought?**

   Why it matters: To properly capture and substantiate a diagnosis, the note should be derived from that decision. If done wrong, this will impact all downstream coding.

5. **Has the solution demonstrably improved financial performance at other health systems?**

   Why it matters: Ask for case studies showing revenue lift, coding complexity improvements, or clinician-level ROI. What are the dollar-value impacts.