AI Adoption and Productivity in Healthcare

Erkmen G. Aslim* Manuel Hoffmann‖ Haonan Yin†

*University of Vermont & IZA

‖University of California, Irvine

†Iowa State University

This NBER Version: *July 22, 2026*

Abstract

Artificial Intelligence (AI) has the potential to reshape work processes and productivity in the health care sector. It presents a unique opportunity to enhance the economic value of professional work by automating time–intensive peripheral tasks, allowing highly skilled practitioners to focus on their core technical expertise. We investigate the adoption, productivity, and work transformation effects of AI scribes – ambient generative AI systems that document clinical conversations – within a large academic health network. Leveraging a staggered rollout and granular data both from electronic health records and the AI scribe platform, we find that adoption of AI scribes is remarkably rapid, reaching near-universal coverage among eligible physicians within approximately a year. Physicians increase their usage post-eligibility and accept the vast majority of AI-generated notes based on audio recordings of physician-patient encounters. Eligibility for the AI scribe raises physician productivity by approximately 20 percent per physician-month, driven both by a higher volume of appointments and a shift toward more complex encounters. By reducing the marginal cost of documentation and alleviating administrative frictions, AI scribes enable a significant reallocation of time toward patient–centered care. A back-of-the-envelope calculation implies a cost-adjusted value of these productivity gains on the order of $1.6 million to $1 billion when scaled to the national physician workforce. Ultimately, this longitudinal analysis demonstrates that when AI tools are designed to address pressing professional burdens, they have the potential to effectively bridge the implementation gap and drive long–term improvements in both the quality and efficiency of care delivery.

JEL-Classification: I10, J0, O33

Keywords: Generative Artificial Intelligence, AI Scribes, Healthcare, Productivity, AI Adoption

---

Acknowledgments: For helpful comments, we thank seminar participants at the University of California, Irvine and the ESMT Berlin as well as participants at the NBER Applications of Artificial Intelligence in Healthcare, Boston, the Workshop on Applied Economics in Digital Health at the Hasso Platner Institute, Potsdam and the Conference on Health IT and Analytics, Washington D.C. We also thank Avi Goldfarb, Dan Zeltzer, and Ariel Stern for valuable feedback and comments. We are grateful for administrative support from Abridge and the academic health network. This study was reviewed and approved by the Institutional Review Board at the University of Vermont (IRB No. 00000485) and the University of California, Irvine (IRB No. 00000596).

> "We are giving agency back to clinicians, creating clarity for patients, and establishing radical efficiencies to transform healthcare from the conversation up."
>
> — Shiv Rao, MD
>
> CEO and Founder, Abridge

Artificial Intelligence (AI) is predicted to substantially transform the economy, work processes, and productivity (Hoffmann et al., 2024; Brynjolfsson, Li, and Raymond, 2025; Dillon et al., 2025; Bick, Blandin, and Deming, 2025; Hoffmann et al., 2025). In healthcare, one particularly promising application of generative AI is the transformation of clinical documentation. Ambient documentation systems, known as "AI scribes," aim to achieve this by automatically generating structured clinical notes from audio recordings of patient–physician encounters (Afshar et al., 2025). These systems use speech recognition and large language models to transcribe conversations, extract clinically relevant information, and format it into standardized notes compatible with electronic health records (EHRs). By automating documentation tasks, AI scribes may reduce administrative workload and mitigate physician burnout, with potential downstream effects on physician productivity, the quality of patient care, and clinical decision-making.

If AI scribes meaningfully reduce documentation burden, they should alter physicians' time allocation, workflow, and potentially their well-being. We study the implementation of a generative AI clinical documentation tool, characterize physician adoption and usage intensity, and estimate its effects on documentation efficiency, measured by editing effort, and physician productivity. The tool we examine is Abridge,[1] an ambient AI documentation system deployed within a large, multi-hospital academic health network integrated into its electronic health record platform. The network includes six operating hospitals, employs over 15,000 staff, and serves more than one million patients across two states, providing a sizable and geographically diverse clinical setting.

We leverage novel, high-frequency (daily) physician-level proprietary data on AI scribe adop-

[1] See Abridge AI, Inc., https://www.abridge.com.

tion, usage intensity, and output quality, linked to detailed productivity metrics at the physician–encounter level. A central advantage of our setting is the ability to construct novel physician-level linkages between AI scribe activity and electronic health record (EHR) data, allowing us to measure objective productivity outcomes such as time spent in the system, note completion timing, appointment closure rates, and documentation turnaround time. This unusually granular integration of AI usage data with operational clinical metrics enables us to observe how generative AI adoption translates into measurable changes in workflow and performance.[2]

Our identification strategy exploits within-provider variation over time in a dynamic difference-in-differences (DID) framework. In the first stage, we estimate the dynamic takeup of AI scribes, where provision of access induces a discontinuous shift in usage over time. Because providers receive access at different times, we employ a heterogeneity-robust DID estimator proposed by Callaway and Sant'Anna (2021), which accommodates staggered treatment timing by comparing early adopters to not-yet-treated providers. We further examine how usage evolves with experience, modeling outcomes as a function of tenure since provision to disentangle the role of learning from the initial adoption response.

We document three sets of results on the adoption side. First, adoption is immediate and widespread: over half of physicians generate their first AI-scribed note within one month of eligibility, rising to near-universal coverage within fourteen months, with approximately 80 to 90 percent of AI-generated text accepted without modification throughout. Second, experience deepens engagement with the tool. As physicians accumulate exposure over the first year, they steadily increase their documentation volume and record more audio per note, and they become progressively more likely to keep AI-generated notes — rising from roughly 5 percentage points in the 2–5 month window to about 12 percentage points by 6–12 months relative to a physician's first month. Third, this deepening extends to how physicians evaluate the AI's output: rather than shifting toward more critical curation, experienced physicians accept a larger share of AI-generated text

[2]For a subset of physicians, we additionally observe repeated institutional survey responses, which provide complementary measures of well-being (e.g., self-reported burnout) and time allocation across clinical and administrative tasks. These data allow us to explore mechanisms underlying the observed productivity effects.

without modification and rate the resulting notes modestly more favorably, consistent with learning to structure clinical interactions more efficiently and to treat the scribe as a collaborative partner.

On the productivity side we document three further sets of results. First, the AI scribe reduces the time physicians spend on documentation: time devoted to note writing falls sharply following eligibility and continues to decline over the first several months, freeing capacity that had previously been absorbed by administrative work. Second, physicians reallocate this recovered time toward direct patient care along two margins, they conduct more appointments, with visit volume rising by roughly five to six encounters per physician-month, and they take on a more complex case mix, with high-complexity Level 4 and Level 5 evaluation-and-management encounters increasing by about three per month. Third, these mechanisms translate into a substantial and durable increase in measured output: becoming eligible for the AI scribe raises physician productivity, as measured by work relative value units (wRVUs), by approximately 20 percent per physician-month, an effect that emerges within the first months of eligibility and persists throughout the post-period. Because ambient documentation captures the clinical content of each visit more completely, the accompanying shift toward higher-complexity encounters reflects care that was actually delivered rather than an increase in coding intensity, indicating that the measured gains correspond to real clinical work.

This paper contributes to the rapidly-evolving literature on the impacts of AI on productivity and the changing nature of professional work. While previous research has focused on AI tools that reduce the cost of core tasks— such as software development (Hoffmann et al., 2024), we document a reverse but complementary effect. In our setting, the AI scribe targets "peripheral" or bureaucratic tasks to enable a focus on core clinical work. This finding serves as a vital complement to the idea that structural frictions within EHR systems act as a significant drag on productivity (Dix, Moran, and Nguyen, 2025). We demonstrate that by improving the inflow of information, AI scribes effectively reduce these administrative frictions, which is as significant to productivity as enhancing core technical outputs.

Second, we provide empirical evidence that there are remarkably few barriers to AI adoption

when the tool is specifically designed to alleviate well–documented professional burdens. The rapid pace of adoption and the intensity of usage observed among the physician cohort in our sample serve as a powerful signal of the tool's value through revealed preferences. This suggests that the "implementation gap" often seen with new medical technologies can be bridged when the technology aligns directly with the clinician's desire to return to patient–centered care.

Third, our study offers a unique longitudinal view of the "AI–practitioner" relationship. By utilizing proprietary data on word acceptance and audio duration, we move beyond simple binary adoption metrics. We demonstrate that the value of these documentation tools does not just come from a temporary growth in encounters, but from a sustained increase in documentation volume and a long–term improvement in dictation efficiency as physicians master the tool.

1 Empirical Context

1.1 The Abridge Ambient Documentation System

The AI scribe we examine is Abridge, an ambient AI documentation system that automatically generates structured clinical notes from audio recordings of patient–physician encounters. Founded in 2018, Abridge has been deployed across more than 150 healthcare systems in the United States, including the University of Pittsburgh Medical Center (UPMC), Yale New Haven Health, Emory Healthcare, Mayo Clinic, Mount Sinai, Corewell Health, and Sutter Health, among others. The platform uses proprietary AI technology developed specifically for clinical documentation workflows. The system supports over 50 clinical specialties, 28 languages, and all care settings.

Abridge listens to the natural conversation between a physician and a patient during a clinical encounter, transcribes the audio using speech recognition, extracts clinically relevant information, and formats it into a structured clinical note compatible with the electronic health record (EHR). The system is designed to run passively in the background, requiring minimal physician interaction during the encounter.

— Figure 1 about here —

Figure 1 displays the Abridge mobile application used for recording encounters and the web-based interface for reviewing and editing AI-generated notes.[3] In addition to the mobile app, Abridge also supports web-based recording, which can be enabled by the health system for both in-person and virtual encounters.

Abridge offers two deployment modes. Abridge Integrated operates as a standalone application: the physician records the encounter on a mobile device or laptop, reviews the AI-generated note in the Abridge web interface or via mobile preview, and sends the finalized note to the patient chart via EHR integration using dotphrases.[4] Alternatively, the physician can copy individual note sections directly into the chart. Abridge Inside is fully embedded within the Epic EHR, allowing physicians to record, review, and finalize notes without leaving the system. Recording is initiated through Epic Haiku,[5] and the AI-generated draft populates directly within the patient record in Epic Hyperspace, where the physician can review, edit, and sign the note.

The Abridge web-based note review interface presents three panels: a worklist of the day's patient encounters, a note editor displaying the AI-generated clinical note in standard documentation format (History of Present Illness, Physical Exam, and Assessment & Plan), and the original conversation transcript. This layout allows physicians to verify the AI-generated content against the source dialogue, edit as needed, and transfer the finalized note to the EHR. Physicians can also rate note quality using a five-star system, which we use as a measure of AI-generated note quality in our empirical analysis.

The example note in Figure 1 illustrates how Abridge transforms unstructured patient–physician dialogue for a patient presenting with shortness of breath and chest tightness into a structured clinical note organized by HPI, Physical Exam, and Assessment & Plan.[6] We provide additional tran-

[3]Notes are also available for preview on the mobile app.

[4]Dotphrases are text shortcuts in Epic that allow users to insert pre-written or dynamically generated blocks of text into clinical notes. Abridge leverages dotphrases to push structured note sections, such as History of Present Illness (HPI), Physical Exam, Results, and Assessment and Plan, directly into the corresponding fields in the EHR.

[5]Haiku is Epic's mobile application, designed for physicians to access patient charts, review results, and manage clinical workflows from a smartphone.

[6]In the example shown, a patient presents with shortness of breath and chest tightness. The AI system organizes the

scription examples in Figure A1, where Panel (a) shows a physical exam transcription and Panel (b) shows a multilingual encounter.

1.2 System Architecture and Validation

The documentation pipeline consists of two sequential components whose design bears directly on the outcomes we measure: an automatic speech recognition system that converts encounter audio into a verbatim transcript, and a note-generation system that transforms that transcript into a structured clinical note. The platform displays both a linked-evidence panel containing the verbatim transcript and the AI-generated note derived from it. Physicians cannot edit the transcript, but they can revise the generated note before signing it into the EHR. Transcript and audio recordings are retained for partner-specific periods; at the academic health network studied here, retention is 30 days. Note generation also draws on prior EHR context, including prior visit notes, existing diagnoses, and the current visit diagnosis.

Abridge models are trained on originally generated notes and do not update automatically from physician edits or in-app interactions. Model updates require direct clinician involvement, manual prompt refinements, and formal evaluations confirming that quality improvements are statistically significant and produce no regressions on established note-quality benchmarks. Prior to any production release, Abridge conducts blinded head-to-head evaluations in which licensed clinicians assess notes from the candidate model alongside notes from the current production model, with authorship masked. To enable early stopping when evidence is conclusive, these trials use anytime-valid sequential hypothesis testing, which controls the false-positive rate regardless of stopping time. Models that pass internal evaluation then enter a staged release: an initial rollout to a small cohort of trained early adopters whose detailed feedback gates broader deployment.

During note generation, Abridge runs real-time analysis using a proprietary guardrails model.

conversation into a structured HPI narrative, a system-by-system physical exam, and a problem-oriented Assessment & Plan with specific action items, including lymph node biopsy recommendations and medication management. The transcript panel preserves the natural, conversational tone of the encounter. The interface also features a "Copy All" button for transferring note sections to the EHR and a "Mark as Done" workflow indicator.

The guardrail architecture distinguishes between two types of components: rule-based filters that handle well-defined failure modes, and trained models and large language models that handle more complex confabulation patterns. Unsupported claims are classified along two independent dimensions: the degree of transcript support (ranging from directly supported to outright contradiction, with intermediate categories for reasonable and questionable inferences) and the clinical severity of leaving the claim uncorrected (minimal, moderate, or major). When an unsupported claim is detected, an automated correction system either revises the claim to align with the transcript, removes it, or determines on the basis of the detection model's reasoning that the flag is a false positive. This guardrail model, trained on over 50,000 curated examples combining open-source hallucination-detection data and domain-specific clinical transcripts, detects 97% of confabulations in internal benchmark testing, compared with 82% for OpenAI's GPT-4o — a roughly six-fold reduction in missed cases (Liang et al., 2025). After generation, the linked-evidence feature allows clinicians to highlight any portion of the note and view the corresponding transcript passage, supporting in-workflow verification. Clinician free-text comments and star-rating quality scores provide a continuous feedback signal that informs the identification of new failure modes and guides subsequent model refinement.

The platform's automatic speech recognition (ASR) system achieves a word error rate (WER) of 12.7% and a medical term recall (MTR) of 97.0% on internal medical conversation benchmarks, figures that have continued to improve since the earlier assessment of Oberst, Liang, and Lipton (2024). These results compare favorably with leading off-the-shelf clinical ASR systems: Google Medical Conversations (WER 16.6%, MTR 96.4%), AWS Transcribe Medical (WER 17.1%, MTR 92.9%), and AssemblyAI's Universal-2 (WER 17.9%, MTR 94.6%).[7] Performance advantages are largest on new medication names, where Abridge achieves an 83% relative reduction in error compared with Google Medical Conversations ASR and an 81% relative reduction compared with OpenAI's Whisper v3.

[7]Abridge's system also outperforms DeepGram Nova-3 (WER 16.4%, MTR 96.3%) on the same benchmark.

2 Methodology

The staggered rollout of AI scribes across the health network generated plausibly exogenous variation in physician eligibility over time, enabling a difference-in-differences (DID) framework to estimate causal effects.[8] We observe the full roster of physicians practicing within the network through administrative EHR data, allowing us to construct a panel at the physician level and follow both adopters and non-adopters over time. This comprehensive coverage enables us to include physician fixed effects and estimate within-physician changes in productivity associated with eligibility and subsequent tool use. To recover static treatment effects, we estimate a pooled two-way fixed effects model:

$$y_{it} = \alpha + \beta \cdot \text{Post}_t + \gamma_i + \delta_t + \varepsilon_{it} \tag{1}$$

where y_{it} is the outcome for physician i in period t, $\text{Post}_t = \mathbf{1}[t \geq t^*]$ is an indicator equal to one after the pooled eligibility provisioning date t^*, γ_i denotes physician fixed effects, δ_t denotes month fixed effects, and ε_{it} is an idiosyncratic error term clustered at the physician level. The coefficient β captures the average within-physician change in the outcome following eligibility. To trace out dynamic effects, we estimate the following model:

$$y_{it} = \alpha + \sum_{\ell \neq -1} \beta_\ell \cdot \mathbf{1}[t - t_i^* = \ell] + \gamma_i + \delta_t + \varepsilon_{it} \tag{2}$$

where ℓ indexes event time relative to the eligibility date t_i^*, and β_ℓ captures the average treatment effect ℓ periods before or after eligibility. The period $\ell = -1$ serves as the reference. Coefficients for $\ell < 0$ provide pre-trend tests of the parallel trends assumption; coefficients for $\ell \geq 0$ trace the dynamic productivity response to eligibility. To address the heterogeneous treatment effect bias inherent in standard TWFE estimators, we follow Callaway and Sant'Anna (2021) and estimate heterogeneity-robust aggregated ATTs.[9]

[8] Importantly, physicians were not informed of the precise timing of eligibility in advance, reducing the likelihood of anticipatory behavioral responses prior to treatment.

[9] You can find an additional model and analysis on learning in Appendix B.

3 Data

We have obtained the roster of physicians of a large medical provider in the United States. It represents the universe of phycisians of this medical provider which provides near-universal coverage of patients in one state of the United States of America. We then merge the roster of physicians with data from the AI Scribe tool Abridge. This allows us to obtain granular information on AI adoption, the notes transcribed via the AI scribe as well as the audio duration of a physician-patient encounter in a session and how much of the AI generated content they accepted. [10]

— Table 1 about here —

Table 1 reports summary statistics for the estimation sample, which comprises $10,004$ physician-patient encounters across $1,868$ physicians after eligibility. AI adoption is substantial, with 71% of physician-month observations reflecting active use of the AI scribe. Conditional on the sample, physicians produce on average 29 notes per month (SD = 50.4), and log on average 506 minutes of audio per physician per month (SD = 902.9). Among AI scribe-specific engagement metrics, 29% of sessions result in the physician retaining the AI-generated note, and among those who accept AI-generated content, physicians incorporate on average 83% of the AI-generated text without modification, suggesting a high degree of trust in the tool among active users.[11]

Further, we summarize physician productivity using work relative value units (wRVUs), the standard currency through which health systems measure and compensate physician output. wRVUs enter physicians' employment contracts directly: compensation typically combines a base salary with a production bonus that is largely a function of accumulated wRVUs. Because wRVUs are assigned at the level of the billed service, they aggregate several distinct behavioral margins.[12] Productivity measured in wRVUs can therefore rise because physicians see more patients, spend

[10] We have further received access to the complete EHR records. The full-scale analysis of the records will be done prior to the NBER SI in the updated working paper version.

[11] Tables A1, A2, and A3 provide descriptive statistics for specialty (primary care or other), location (headquarters, satellites), and gender (male, female).

[12] A short, low-complexity office visit (CPT 99213, roughly 15 minutes) carries 1.30 wRVUs, whereas a longer, higher-complexity visit (CPT 99215, roughly 45 minutes) carries 2.80 wRVUs.

more time per encounter, or handle more complex cases:

$$wRVU_{it} = f(\text{volume}, \text{time}, \text{complexity}). \tag{3}$$

4 Results: AI Adoption

4.1 Main Results

Our primary analysis on AI adoption focuses on all physicians in the health care network.

— Figure 2 about here —

Figure 2 presents event-time evidence on adoption for eligible physicians and usage following eligibility within the time frame from May 2024 to Nov 2025. The figure is based on a balanced panel of 1,868 physicians and 35,492 patient–physician encounters observed before and after eligibility based on the raw data. Adoption is immediate and widespread: 56 percent of eligible physicians create their first AI-generated note within the first month of eligibility, and adoption rises steadily to near universal use within fourteen months.

Usage intensity increases in parallel. The number of AI-generated notes per physician rises quickly from approximately 20 in the first month to nearly 80 notes per month after one year. Audio recording time exhibits a similar pattern, increasing sharply upon eligibility and exceeding 1,700 minutes per physician per month by the end of the first year. Taken together, these patterns imply a remarkably stable documentation structure of roughly 20 minutes of audio per note, suggesting systematic rather than sporadic use.

Importantly, physicians accept approximately 80 to 90 percent of AI-generated text without modification throughout the post-eligibility period. This high and stable acceptance rate provides a first-order indication of substantial documentation efficiency gains and suggests that the tool is meaningfully integrated into routine clinical workflow rather than used experimentally or intermittently. Such reductions in documentation effort create scope for time reallocation, potentially

increasing physician productivity.

Table A4 shows some of the effects based on model equation 1. Since most of the adoption variation is in the first year, we have estimated the effects in that table and in all analysis henceforth for a one-year window. Consistent with Figure 2 Panel A it shows that the average adoption rate for the eligible is around 60%. The average audio duration per encounter is around 8 minutes. Most physicians decide to keep around 24% of the AI generated notes that were ultimately sent to the EHR where 80% of the AI generated content from those notes was unaltered by the physician (in line with Figure 2 Panel D). Further, in the post period where AI usage was an option there is an 8 percent chance of an individual rating an encounter note where an AI scribe was involved. For robustness, we further show the main results from Figure 2 using Callaway and Sant'anna (2021) based on model equation 2. The patterns are confirmed in the staggered difference-in-differences models with physician and year-month fixed effects, where the static average treatment effect on the treated (ATT) is positive, large in magnitude, and precisely estimated.

4.2 Heterogeneity

4.2.1 Specialty: Primary Care versus Non-Primary Care

We differentiate specialties by primary care physicians and other specialties. We define primary care, as physicians which are in family medicine and general internal medicine, i.e. clinical settings in which documentation demands are substantial and ambient AI tools are particularly well suited to routine outpatient encounters. Next, we show the same statistics as in Figure 2 with the exception of total audio recording for brevity. The total audio recording follows the usage trend for the number of notes quite closely.

— Figure 3 about here —

Figure 3 displays AI adoption, usage as measured in the number of notes, as well as the percent of AI generated text that is accepted in the notes for primary care physicians vs non primary care

physicians.[13] Panels A and B show that both groups exhibit zero adoption prior to provisioning, followed by a steep rise that plateaus close to full adoption by months 15 post-provisioning. The adoption trajectories are strikingly similar across the two subgroups, suggesting that specialty type does not meaningfully differentiate whether physicians adopt the tool once it becomes available.

The most pronounced difference between the two groups emerges in usage intensity. Panels C and D reveal that primary care physicians ramp up to approximately 80–100 mean monthly notes by months 10–20, whereas non-primary care physicians reach only around 20–25 notes per month—roughly one quarter of the primary care volume. This pattern is consistent with our characterization of primary care settings as environments with high documentation burdens and routine outpatient encounter structures that are particularly amenable to ambient AI scribing.

Despite these differences in usage intensity, Panels E and F show that both groups accept AI-generated content at similarly high rates. Acceptance rates jump from zero pre-provisioning to approximately 80–85% immediately after provisioning and remain broadly stable thereafter. This suggests that conditional on engaging with the tool, physicians across specialties find the AI-generated output largely acceptable, and that the lower usage intensity among non-primary care physicians reflects differences in encounter volume and workflow fit rather than dissatisfaction with output quality.

4.2.2 Location: Main Hospital versus Affiliate Hospitals

Next, we examine whether AI Scribe adoption and usage patterns differ between the main hospital (the "headquarters") and its affiliate hospitals (the "satellites").

— Figure 4 about here —

Figure 4 shows AI adoption, usage as measured by the number of notes, and the share of AI-generated text accepted into notes at the headquarters ($HQ = 1$) versus the satellites ($HQ = 0$).[14] Panels A and B show that both groups exhibit zero adoption prior to provisioning, followed by

[13] Figure A3 shows the staggered rollout figures based on equation 2.
[14] Figure A4 shows the staggered rollout figures based on equation 2.

a steep rise that plateaus close to full adoption by approximately 15 months post-provisioning. The adoption trajectories are remarkably similar across the two location types, suggesting that institutional setting does not meaningfully differentiate whether physicians adopt the tool once it becomes available.

Differences between the two groups are more pronounced in usage intensity. Panels C and D reveal that physicians in the headquarters ramp up to approximately 40–60 mean monthly notes by months 10–20, whereas physicians at other locations reach somewhat higher volumes of approximately 60–100 notes per month, though confidence intervals are wide in both cases. This pattern may reflect differences in patient volume, encounter complexity, or workflow structure across institutional settings.

Despite these differences in usage intensity, Panels E and F show that both groups accept AI-generated content at similarly high rates. Acceptance rates jump from zero pre-provisioning to approximately 80–85% immediately after provisioning and remain broadly stable thereafter. This suggests that conditional on engaging with the tool, physicians across both headquarters and satellite locations find the AI-generated output largely acceptable, and that any differences in usage intensity reflect workflow and volume factors rather than dissatisfaction with output quality.

4.2.3 Gender: Male versus Female Physicians

Finally, due to substantial variation in adoption of technology across gender as well as differences in the workplace across gender, we are interested whether we find differences across male and female physicians in AI Scribe adoption and usage.

— Figure 5 about here —

Figure 5 shows AI adoption, usage as measured by the number of notes, and the share of AI-generated text accepted into notes for female versus male physicians.[15] Panels A and B show that both groups exhibit zero adoption prior to provisioning, followed by a steep rise that plateaus

[15]Figure A5 shows the staggered rollout figures based on equation 2.

close to full adoption by approximately 15 months post-provisioning. The adoption trajectories are strikingly similar across genders, suggesting that gender does not meaningfully differentiate whether physicians adopt the tool once it becomes available.

Turning to usage intensity, Panels C and D reveal a broadly similar pattern across genders. Both female and male physicians ramp up to approximately 40–60 mean monthly notes by months 5–10, continuing to grow to roughly 60–100 notes by months 15–20, with overlapping confidence intervals throughout. Unlike the specialty comparison, where primary care physicians showed markedly higher usage, the gender comparison does not reveal a pronounced divergence in documentation volume, suggesting that encounter structure and specialty mix may be more important determinants of usage intensity than gender per se.

Panels E and F show that both female and male physicians accept AI-generated content at similarly high rates, jumping from zero pre-provisioning to approximately 80–85% immediately after provisioning and remaining broadly stable thereafter. The near-identical acceptance trajectories across genders further suggest that conditional on engaging with the tool, physicians find the AI-generated output equally acceptable regardless of gender, and that any future divergence in outcomes is more likely to stem from differences in workflow integration than in attitudes toward the tool's output quality.

5 Results: AI & Physician Productivity

5.1 Main Results

Our primary analysis on productivity focuses on all physicians in the health care network. We begin with the mechanisms through which the AI scribe may affect physician output: time spent on documentation, appointment volume, and the case mix of encounters.

— Figure 6 about here —

Figure 6 presents event study estimates for these three margins. Panel (a) shows a sustained

decline in time spent on note writing following eligibility, with the reduction deepening after the first three months. This pattern is consistent with physicians becoming more efficient users of the tool over time. The recovered time is capacity that physicians can reallocate to patient care, and Panels (b) and (c) show how they use it. Appointment volume rises by roughly five to six visits per physician-month after eligibility, and the number of high-complexity encounters, defined as Level 4 and Level 5 evaluation and management (E/M) visits, increases by about three per month. Both effects appear shortly after eligibility and persist throughout the post-period.

— Figure 7 about here —

These margins translate into measured productivity. Estimating the dynamic event study specification on work relative value units (wRVUs), we find that becoming eligible for the AI scribe raises physician productivity substantially. The event-study coefficients show no evidence of differential pre-trends in the months before eligibility, consistent with the parallel-trends assumption, followed by a sustained increase after eligibility. The implied intention-to-treat (ITT) effect is an increase of approximately 35 wRVUs per physician-month, or about 21 percent of the pre-eligibility mean of 226.8 monthly wRVUs. The effect emerges within the first months of eligibility and does not fade, indicating a durable change in physician output rather than a transitory adjustment.

The mechanism estimates suggest that the wRVU gains operate through two complementary channels. Part of the effect reflects volume: physicians see more patients with the time recovered from documentation. Part reflects composition: a larger share of encounters is coded at higher complexity levels. The compositional shift need not indicate upcoding. Ambient documentation captures the clinical content of a visit more completely, so encounters that were previously undercoded may now be billed at the complexity level the care in fact involved. Physicians with recovered time may also be more willing to take on complex cases. Under either interpretation, the measured gains correspond to clinical work actually performed rather than to changes in coding intensity alone.

6 Conclusion

This paper provides novel empirical evidence on the adoption, usage, and productivity effects of AI scribes in a large academic health network. Leveraging granular physician-month-level data from both the AI scribe platform and electronic health records, combined with a staggered rollout design and heterogeneity-robust difference-in-differences estimators, we document three main findings.

First, adoption is immediate and near-universal. Over half of physicians generate their first AI-scribed note within one month of eligibility, rising to near-universal coverage within fourteen months across all subgroups — regardless of specialty, location, or gender. The strikingly uniform adoption trajectories suggest that when AI tools are designed to directly address well-documented professional burdens, the implementation gap that typically characterizes new medical technologies can be effectively bridged.

Second, experience deepens how physicians engage with the technology. As exposure accumulates over the first year, physicians increase their documentation volume and audio recording duration, keep a growing share of AI-generated notes, and accept more of the AI's raw text without modification — with every margin rising monotonically and significantly between the 2–5 month and 6–12 month experience windows. Rather than a transition from passive adoption to active curation, the pattern is one of deepening reliance: physicians who use the scribe longer route more of the encounter through it and evaluate its output at least as favorably, consistent with a professionalization of the human–AI collaboration over time.

Third, these adoption dynamics translate into substantial and durable productivity gains. Physicians reallocate the time recovered from documentation toward direct patient care: appointment volume rises by roughly five to six visits per physician-month, and the number of high-complexity Level 4 and Level 5 evaluation-and-management encounters increases by about three per month. Together these margins raise measured output, as captured by work relative value units (wRVUs), by approximately 20 percent per physician-month—an effect that emerges within the first months of eligibility and persists thereafter. Because ambient documentation records the content of each

visit more completely, the accompanying shift toward higher-complexity encounters reflects care that was actually delivered rather than an increase in coding intensity.

Fourth, the productivity gains from AI scribes are not uniformly distributed. Primary care physicians, who face the heaviest documentation burdens, exhibit substantially higher usage intensity than non-primary care physicians, reaching roughly four times the monthly note volume. By contrast, gender and location produce minimal divergence in either adoption or usage trajectories, suggesting that encounter structure and specialty mix are the primary moderators of how much value physicians extract from ambient AI tools. Across all subgroups, AI acceptance rates converge rapidly to approximately 80–85%, indicating that output quality is broadly acceptable regardless of physician characteristics.

Taken together, these findings paint a picture of AI scribes as a technology that successfully reduces the marginal cost of documentation and reallocates physician time toward patient-centered care. By automating peripheral administrative tasks rather than core clinical work, AI scribes represent a complementary rather than substitutive intervention — one that amplifies the value of physician expertise rather than displacing it. Our results suggest that the productivity gains from AI in healthcare may be largest precisely where administrative frictions are most severe, pointing to a broader principle: AI tools that target well-identified professional pain points are likely to achieve both rapid adoption and sustained impact.

References

Afshar, Majid, Mary Ryan Baumann, Felice Resnik, Josie Hintzke, Anne Gravel Sullivan, Graham Wills, Kayla Lemmon, Jason Dambach, Leigh Ann Mrotek, Mariah Quinn et al. 2025. "A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being." *NEJM AI* 2 (12):AIoa2500945.

Bick, Alexander, Adam Blandin, and David Deming. 2025. "The impact of generative ai on work productivity." Tech. rep., Federal Reserve Bank of St. Louis.

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond. 2025. "Generative AI at Work." *The Quarterly Journal of Economics* 140 (2):889–942.

Callaway, Brantly and Pedro HC Sant'Anna. 2021. "Difference-in-differences with multiple time periods." *Journal of Econometrics* 225 (2):200–230.

Dillon, Eleanor W, Sonia Jaffe, Nicole Immorlica, and Christopher T Stanton. 2025. "Shifting work patterns with generative ai." Tech. rep., National Bureau of Economic Research.

Dix, Rebekah, Kelsey Moran, and Thi Mai Anh Nguyen. 2025. "Costs of Technological Frictions: Evidence from EHR (Non-) Interoperability." Tech. rep., Working Paper.

Hoffmann, Manuel, Sam Boysel, Frank Nagle, Sida Peng, and Kevin Xu. 2024. "Generative AI and the Nature of Work." *Revise & Resubmit at Management Science* .

Hoffmann, Manuel, Sam Boysel, Frank Nagle, and Kevin Xu. 2025. "The Generative AI - Gender Puzzle." *Working Paper* .

Liang, Davis, Michael Oberst, Chenhao Tan, and Zachary C. Lipton. 2025. "The Science of Confabulation Elimination: Toward Hallucination-Free AI-Generated Clinical Notes." Whitepaper, Abridge AI, Pittsburgh, PA.

Oberst, Michael, Davis Liang, and Zachary C. Lipton. 2024. "Pioneering the Science of AI Evaluation." Whitepaper, Abridge AI, Pittsburgh, PA.

Figure 1: Abridge Mobile Application and Note Review Interface

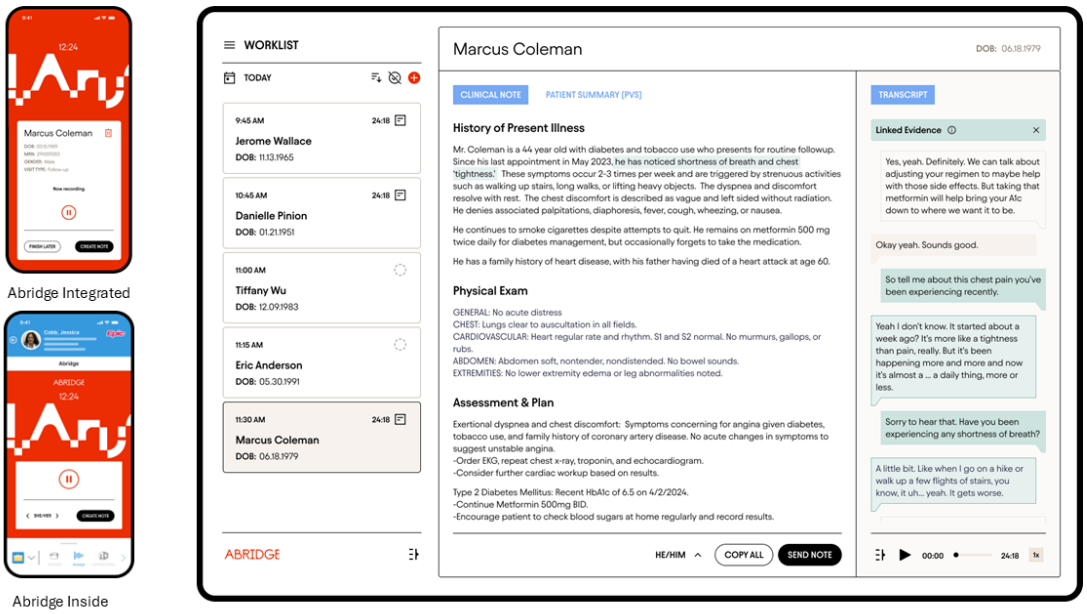
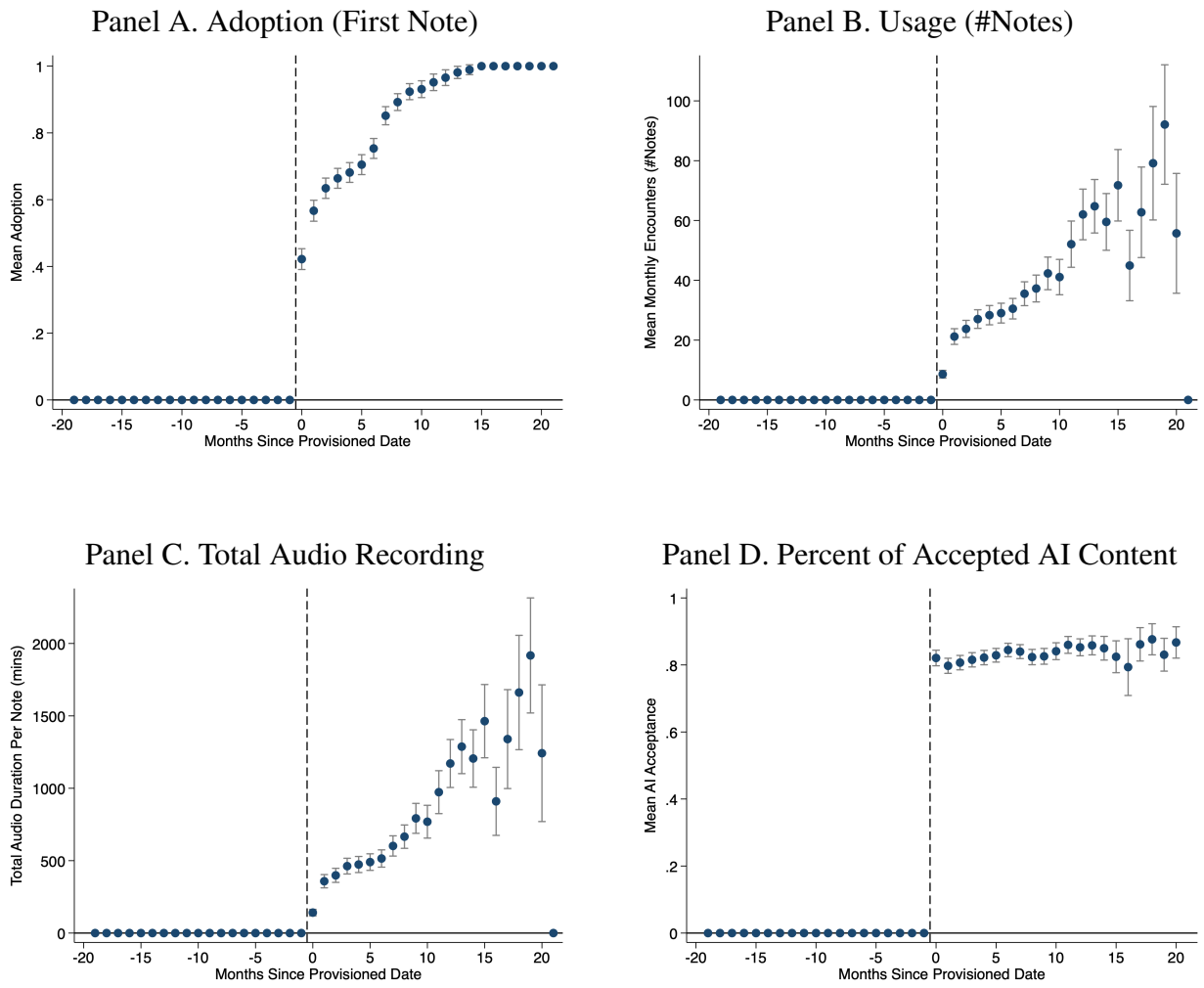


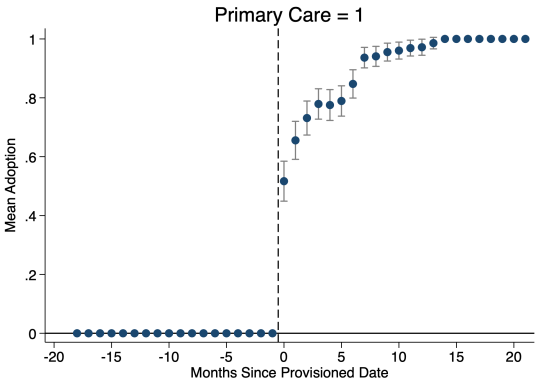
Figure 2: Staggered Rollout: AI Scribes Adoption and Usage



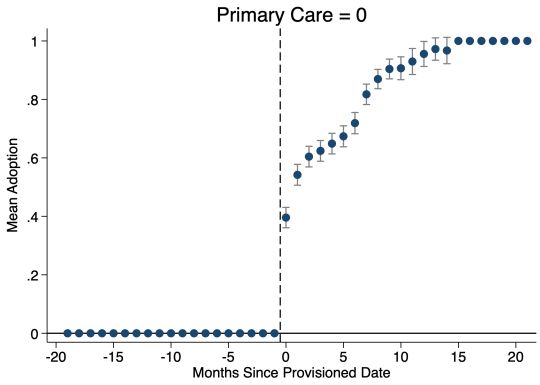
Notes: Panel A illustrates initial adoption of AI scribes, defined as the first AI-generated note following physician eligibility. Panel B reports the number of clinical notes generated using the AI scribe. Panel C displays total audio recording duration per physician. Panel D presents the percentage of AI-generated text accepted without modification, measured as the share of AI-produced content incorporated directly into the final clinical note. This metric serves as a proxy for documentation efficiency. The figures are based on a time period of May 2024 to Nov 2025. 95 percent confidence intervals are shown, with standard errors clustered at the physician level.

Figure 3: Dynamics for Primary Care vs. Non Primary Care

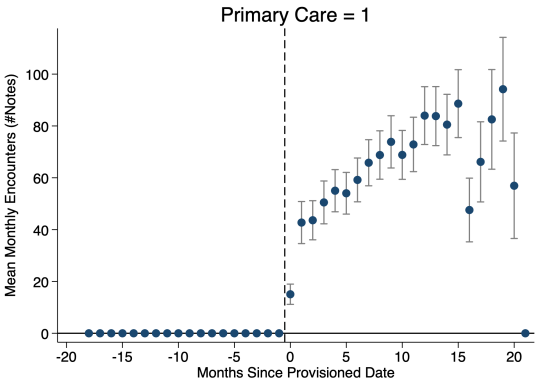
Panel A. Adoption – Primary Care



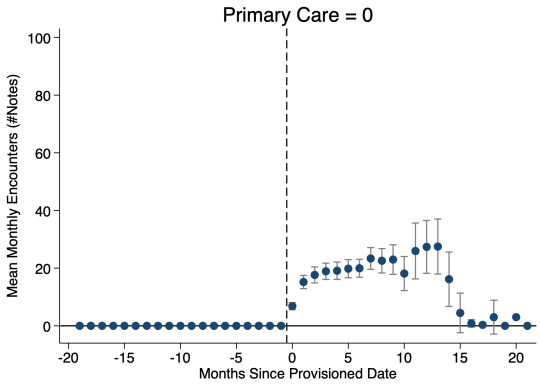
Panel B. Adoption – Non Primary Care



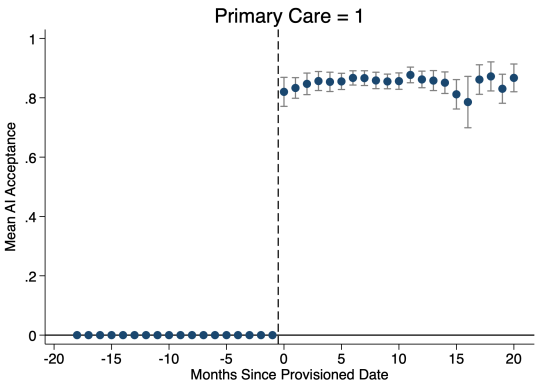
Panel C. Usage (#Notes) – Primary Care



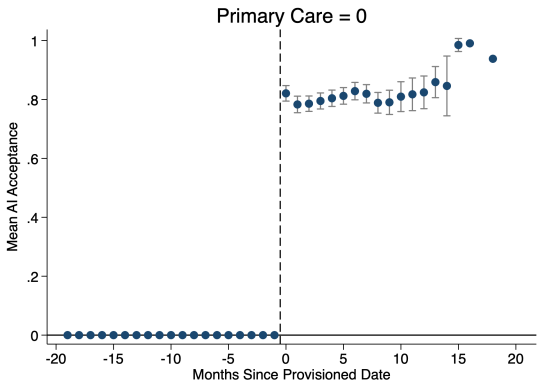
Panel D. Usage (#Notes) – Non Primary Care



Panel E. AI Acceptance – Primary Care



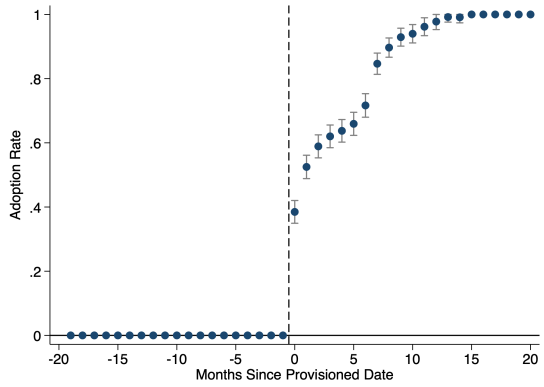
Panel F. AI Acceptance – Non Primary Care



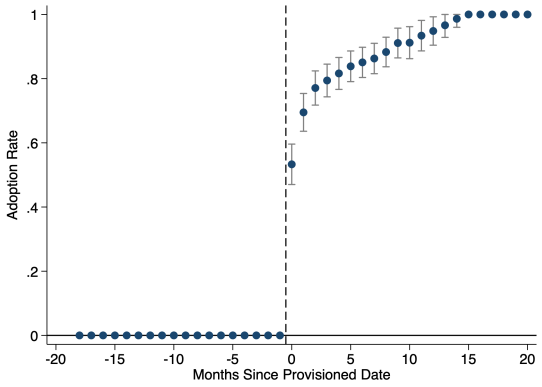
Notes: Panel A and B illustrate AI scribe adoption and usage for primary care and non-primary care physicians, respectively. Panels C and D report mean monthly physician-patient encounters. Panels E and F present the percentage of AI-generated content accepted without modification. The figures are based on a time period of May 2024 to Nov 2025. 95 percent confidence intervals are shown, with standard errors clustered at the physician level.

Figure 4: Dynamics for Headquarters vs. Satellites

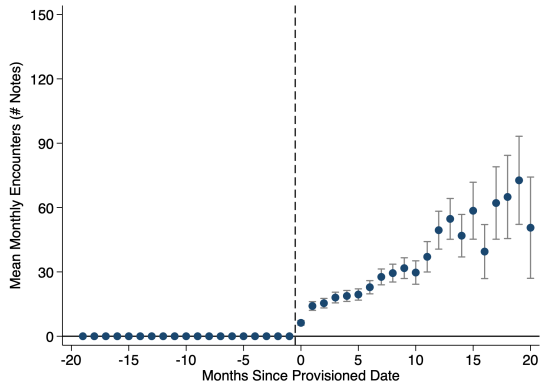
Panel A. Adoption – Headquarters



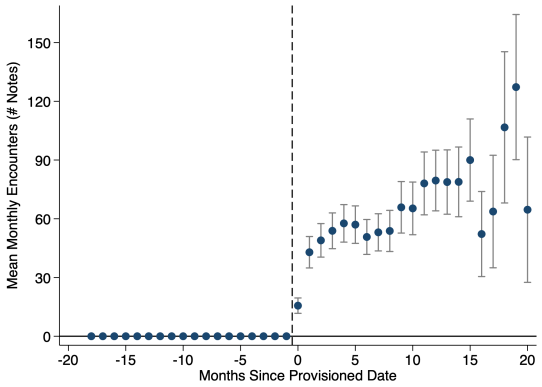
Panel B. Adoption – Satellites



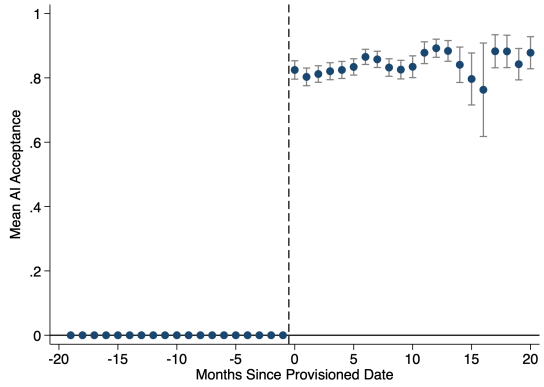
Panel C. Usage (#Notes) – Headquarters



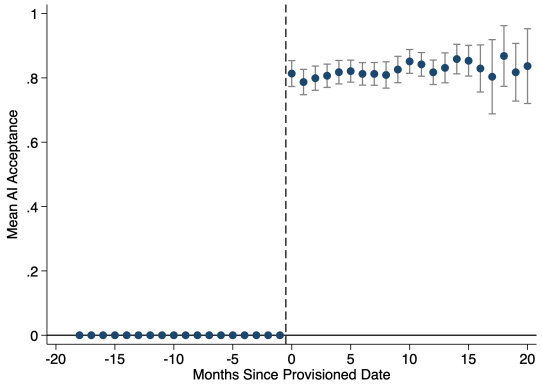
Panel D. Usage (#Notes) – Satellites



Panel E. AI Acceptance – Headquarters



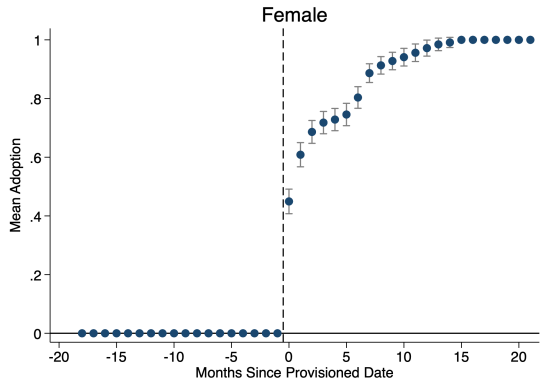
Panel F. AI Acceptance – Satellites



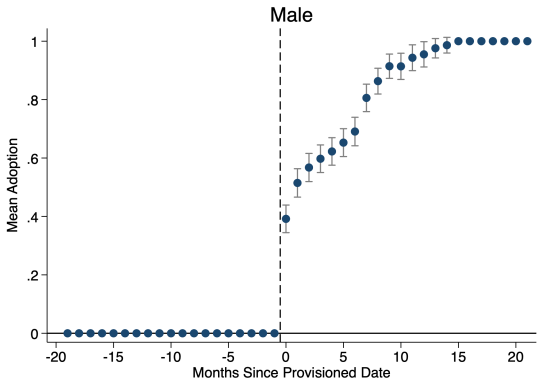
Notes: Panels A and B illustrate AI scribe adoption for headquarters and other locations, respectively. Panels C and D report mean monthly physician-patient encounters. Panels E and F present the percentage of AI-generated content accepted without modification. The figures are based on a time period of May 2024 to Nov 2025. 95 percent confidence intervals are shown, with standard errors clustered at the physician level.

Figure 5: Dynamics for Female vs. Male Physicians

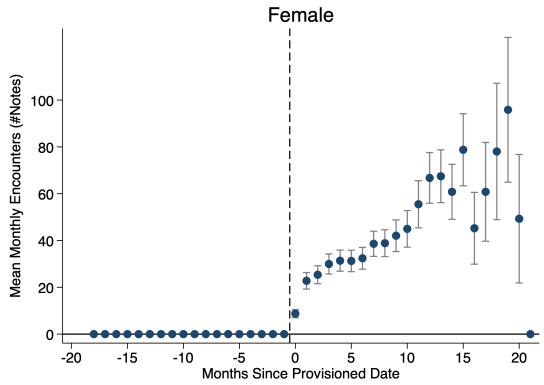
Panel A. Adoption – Female



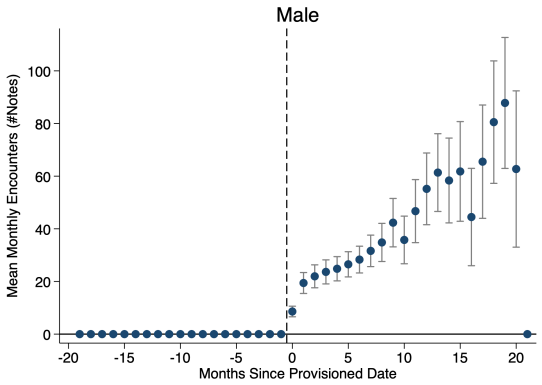
Panel B. Adoption – Male



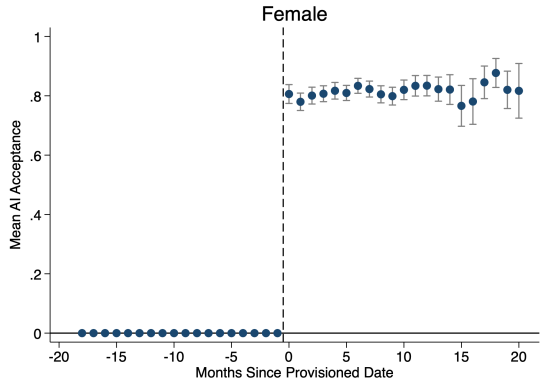
Panel C. Usage (#Notes) – Female



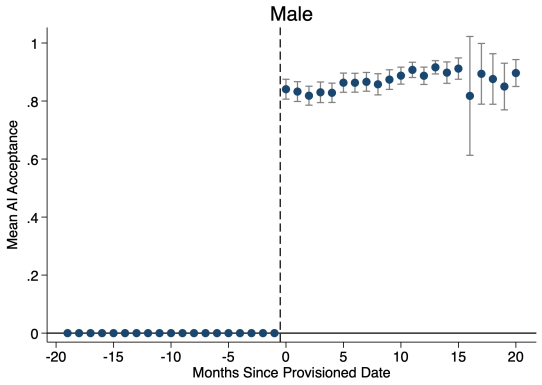
Panel D. Usage (#Notes) – Male



Panel E. AI Acceptance – Female



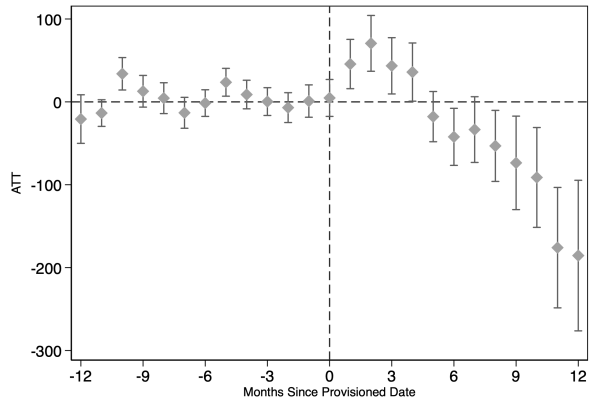
Panel F. AI Acceptance – Male



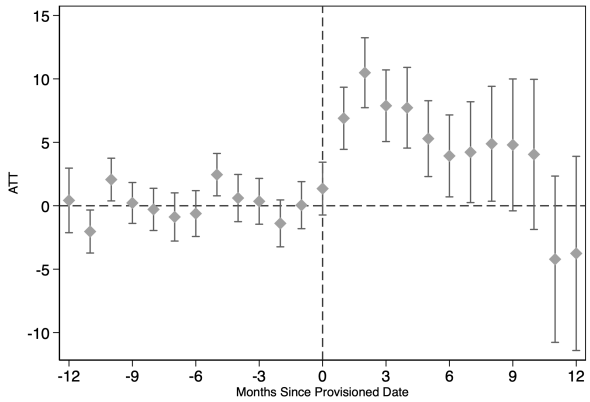
Notes: Panels A and B illustrate AI scribe adoption for female and male physicians, respectively. Panels C and D report mean monthly physician-patient encounters. Panels E and F present the percentage of AI-generated content accepted without modification. The figures are based on a time period of May 2024 to Nov 2025. 95 percent confidence intervals are shown, with standard errors clustered at the physician level.

Figure 6: Effect of AI Scribe Eligibility on Documentation Time, Throughput, and Case Mix

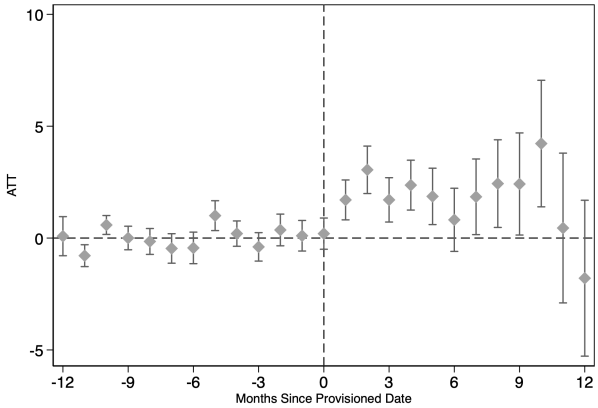
(a) Time spent on note writing



(b) Appointment volume

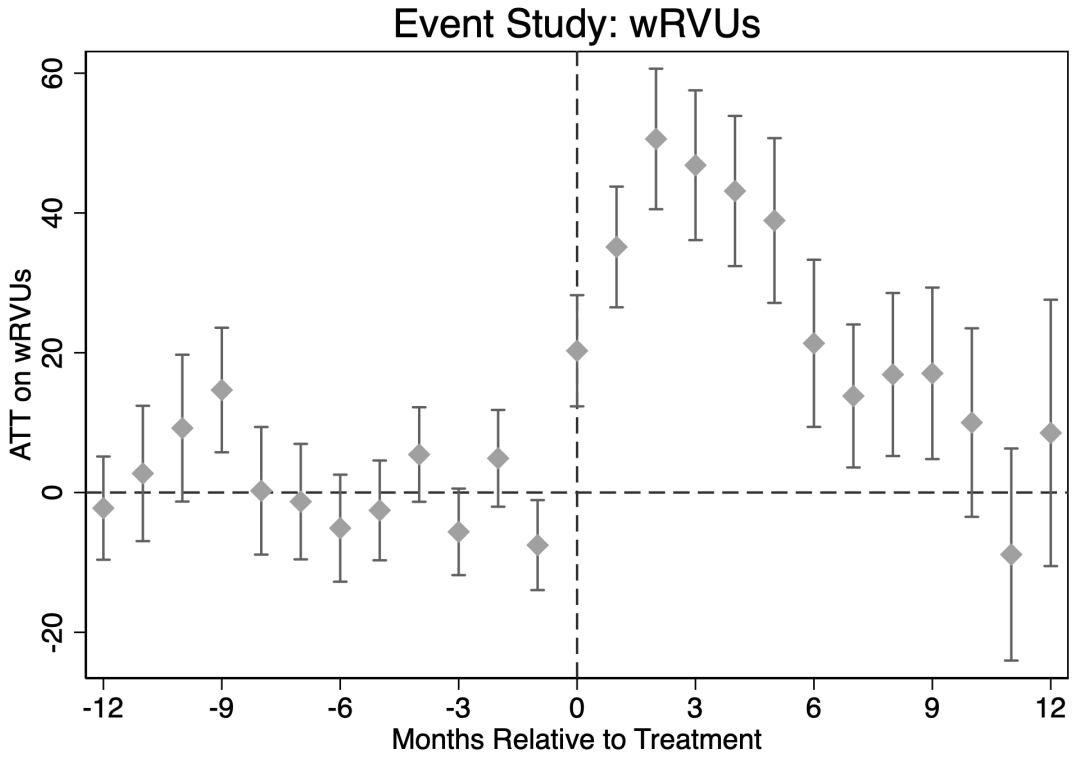


(c) Encounter complexity



Notes: This figure displays the effect of AI Scribe eligibility on documentation time, throughput, and case mix. Panel (a) shows time spent on note writing, panel (b) shows the number of appointments, and panel (c) shows encounter complexity. Complexity is measured using Evaluation and Management (E/M) coding levels (Levels 1–5), where Level 4 indicates higher complexity encounters involving multiple chronic conditions, exacerbations, or significant diagnostic and treatment decisions, and Level 5 indicates the highest complexity encounters involving severe illness, high risk of morbidity, or extensive medical decision making. The outcome in panel (c) is the number of Level 4 and Level 5 encounters. Coefficients are estimated from an event study specification using the staggered difference-in-differences design, with the month prior to eligibility as the reference period. Bands denote 95% confidence intervals based on standard errors clustered at the physician level.

Figure 7: Effect of AI Scribe Eligibility on Physician Productivity



Notes: This figure displays the effect of AI Scribe eligibility on physician productivity, as measured by work relative value units (wRVUs). Coefficients are estimated from an event study specification using the staggered difference-in-differences design, with the month prior to eligibility as the reference period. Bands denote 95% confidence intervals based on standard errors clustered at the physician level.

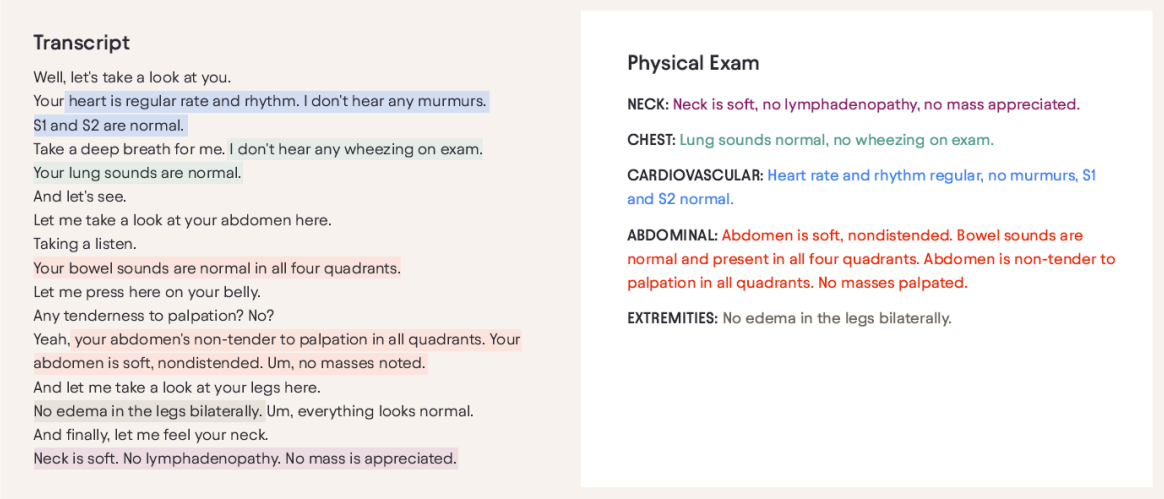
Table 1: Descriptive Statistics

	Mean	SD	Min	Max
AI Adoption	0.71	0.45	0.00	1.00
Notes (Monthly)	29.10	50.35	0.00	420.00
Audio Duration (Min/Physician)	506.64	902.92	0.00	6554.73
Accepted AI Content	0.83	0.17	0.01	1.00
Physician Observations	1,868			

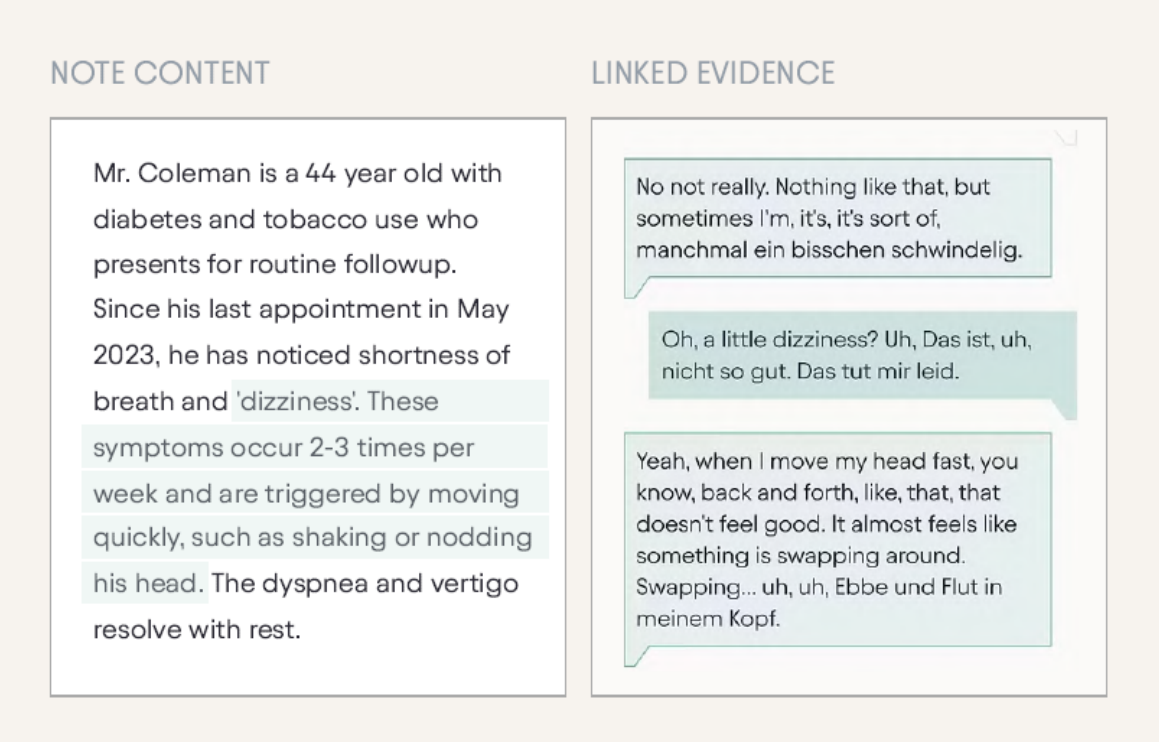
Notes: This table shows descriptive statistics at the physician-month level. AI adoption is an indicator equal to one if the physician used the AI scribe in a given month. Notes (monthly) is the count of AI-scribed notes per physician per month. Audio duration is total recorded audio in minutes per physician per month. Accepted AI Content is the share of AI-generated content accepted by the physician, conditional on use. The sample comprises 1,868 unique physicians observed over multiple months.

A Online Appendix: Tables and Figures

Figure A1: Abridge Ambient Documentation

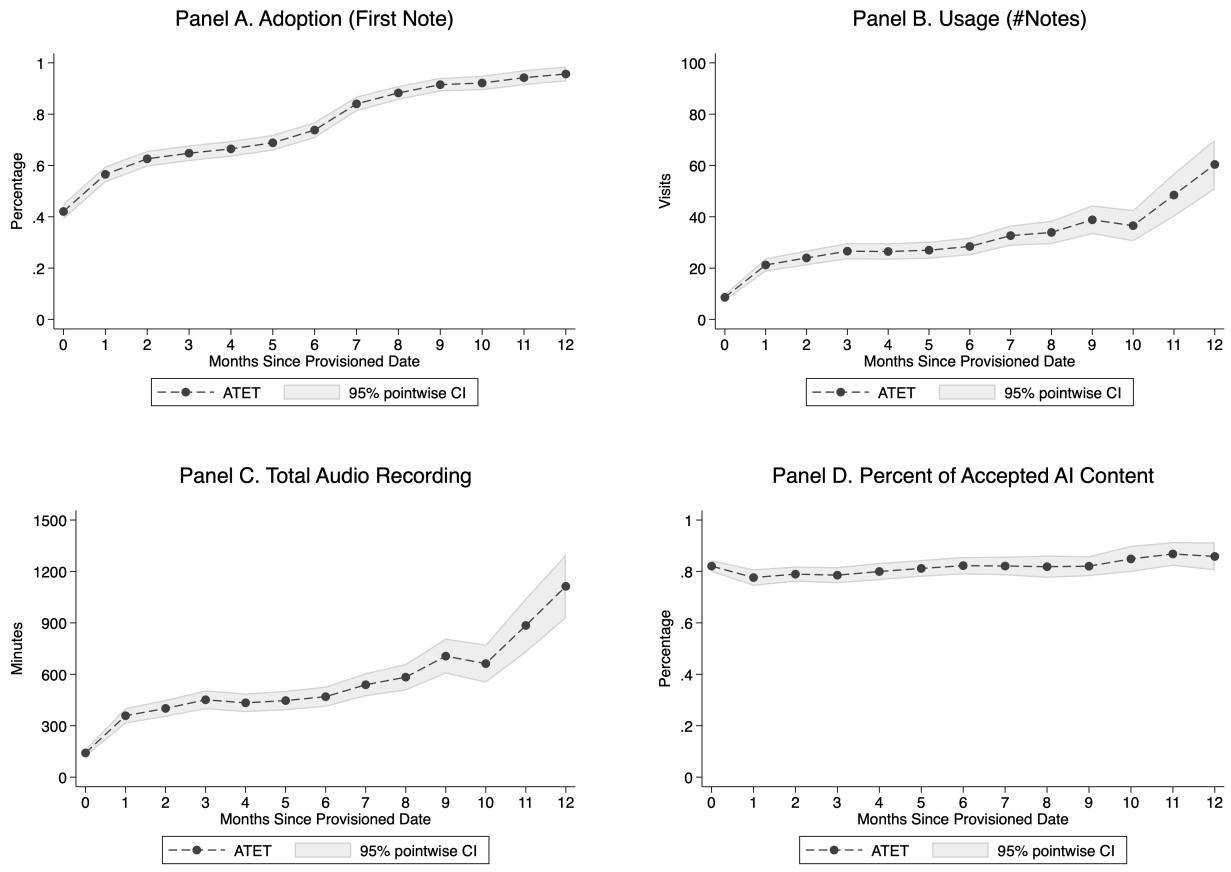


(a) Physical Exam Transcription



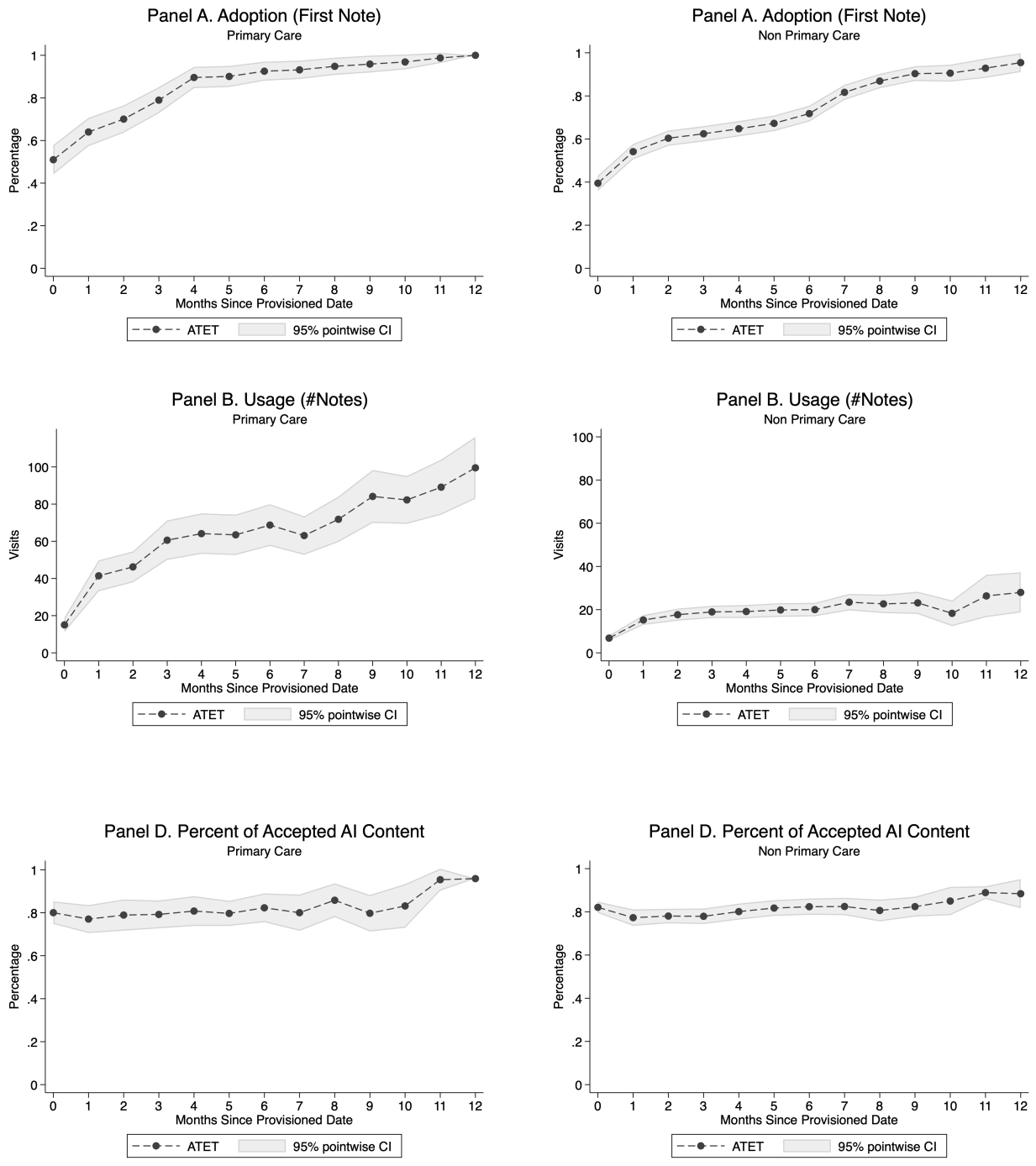
(b) Multilingual Encounter Transcription

Figure A2: AI Scribes Adoption and Usage via Callaway and Sant'Anna (2021)



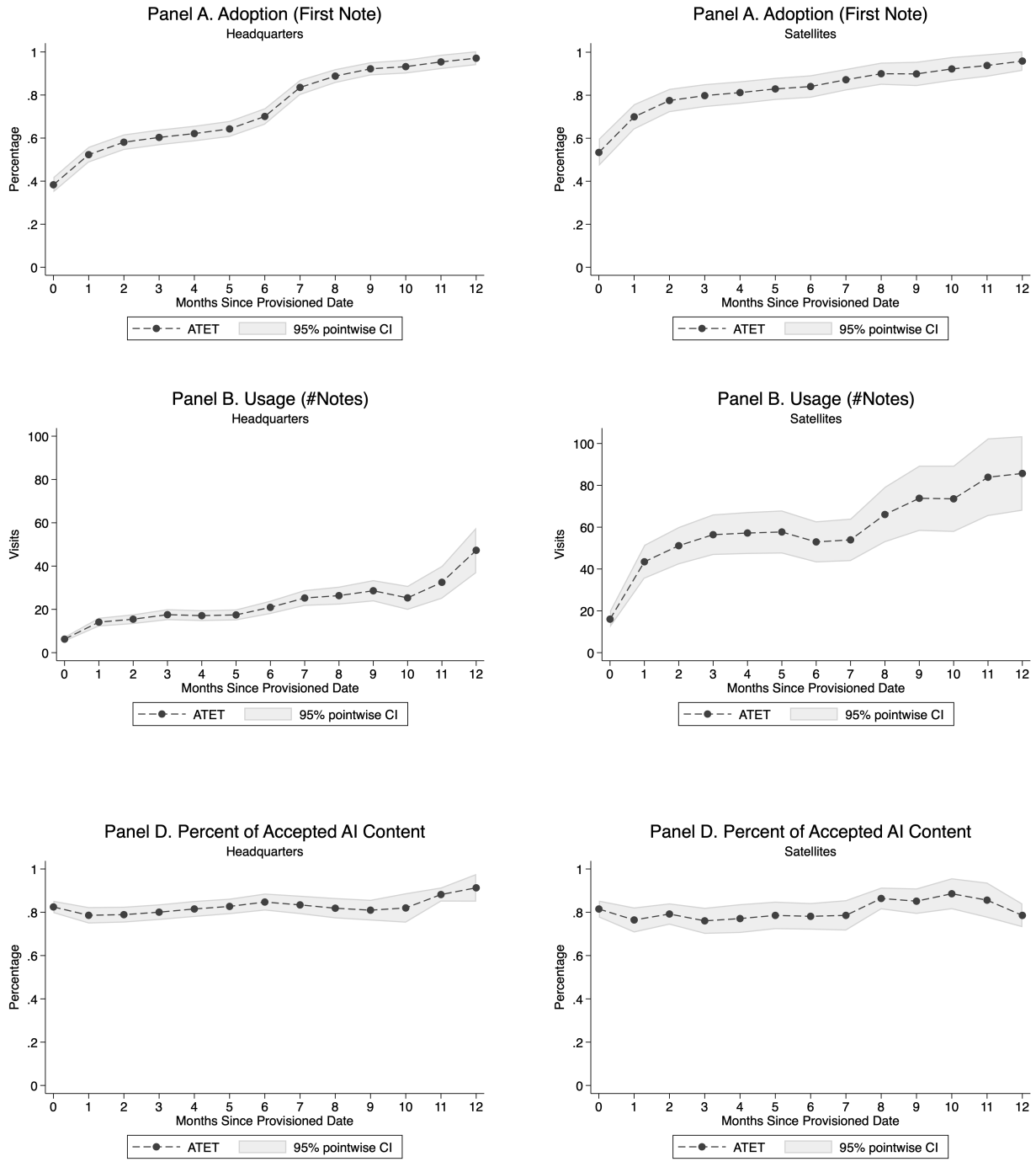
Notes: Panel A illustrates initial adoption of AI scribes, defined as the first AI-generated note following physician eligibility. Panel B reports the number of clinical notes generated using the AI scribe. Panel C displays total audio recording duration per physician. Panel D presents the percentage of AI-generated text accepted without modification, measured as the share of AI-produced content incorporated directly into the final clinical note. This metric serves as a proxy for documentation efficiency. All panels cover the 12 months before and the 12 months after each physician's provisioning date. The underlying data span May 2024 through November 2025. 95 percent confidence intervals are shown, with standard errors clustered at the physician level.

Figure A3: Dynamics for Primary Care vs. Non Primary Care



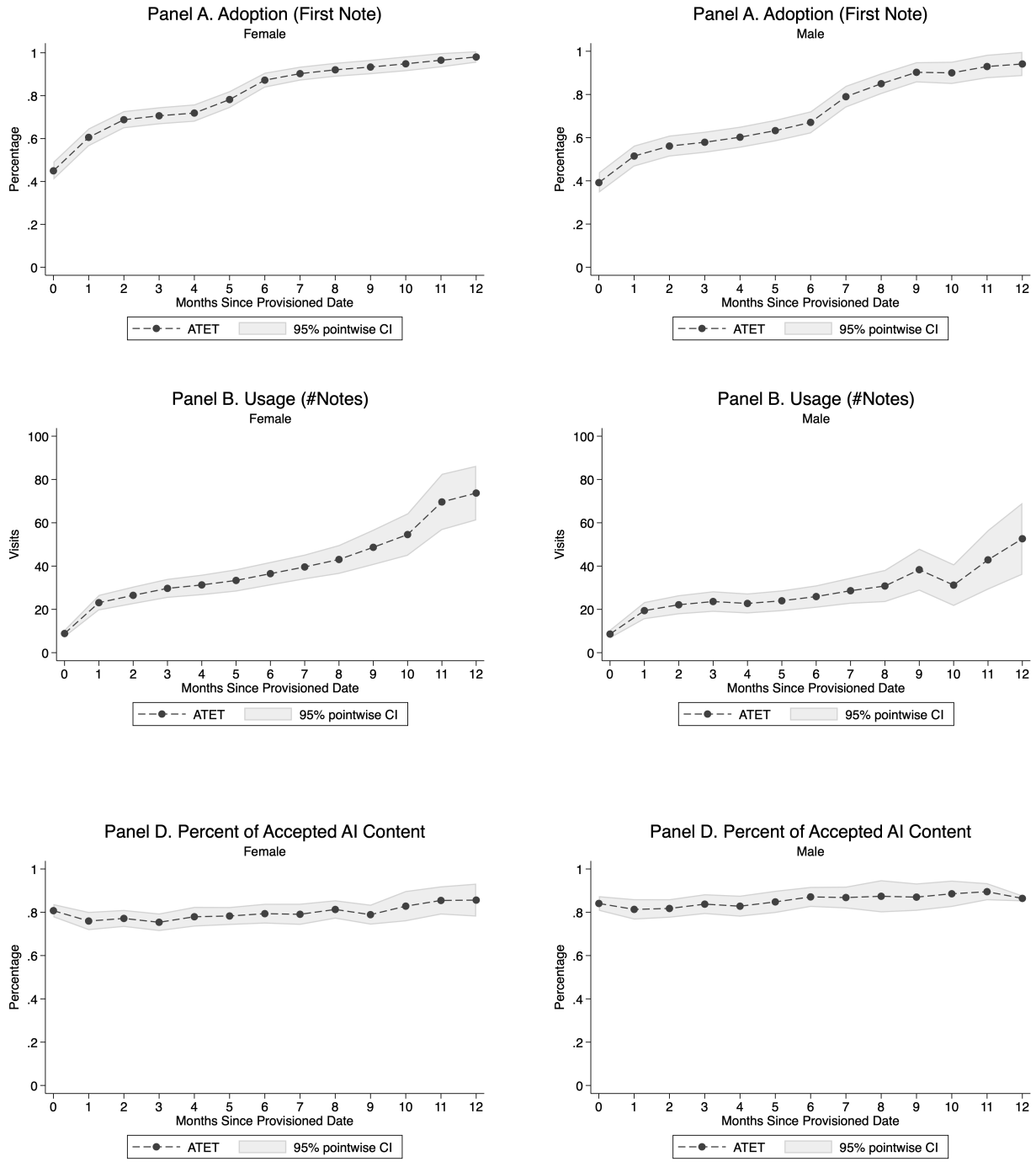
Notes: Panels A and B illustrate AI scribe adoption for primary care and non-primary care physicians, respectively. Panels C and D report mean monthly physician-patient encounters. Panels E and F present the percentage of AI-generated content accepted without modification. All panels cover the 12 months before and the 12 months after each physician's provisioning date. The underlying data span May 2024 through November 2025. 95 percent confidence intervals are shown, with standard errors clustered at the physician level.

Figure A4: Dynamics for Headquarters vs. Satellites



Notes: Panels A and B illustrate AI scribe adoption for headquarters and satellite locations, respectively. Panels C and D report mean monthly physician-patient encounters. Panels E and F present the percentage of AI-generated content accepted without modification. All panels cover the 12 months before and the 12 months after each physician's provisioning date. The underlying data span May 2024 through November 2025. 95 percent confidence intervals are shown, with standard errors clustered at the physician level.

Figure A5: Dynamics for Female vs. Male Physicians



Notes: Panels A and B illustrate AI scribe adoption for female and male physicians, respectively. Panels C and D report mean monthly physician-patient encounters. Panels E and F present the percentage of AI-generated content accepted without modification. All panels cover the 12 months before and the 12 months after each physician's provisioning date. The underlying data span May 2024 through November 2025. 95 percent confidence intervals are shown, with standard errors clustered at the physician level.

Table A1: Descriptive Statistics: Specialty

	Primary Care				Non-Primary Care			
	Mean	SD	Min	Max	Mean	SD	Min	Max
AI Adoption	0.82	0.38	0.00	1.00	0.66	0.47	0.00	1.00
Notes (Monthly)	56.32	64.21	0.00	344.00	18.09	38.38	0.00	420.00
Audio Duration (Min/Physician)	1093.91	1266.05	0.00	6554.73	269.15	548.67	0.00	4596.59
Keep AI Notes	0.41	0.49	0.00	1.00	0.24	0.43	0.00	1.00
Accepted AI Content	0.86	0.14	0.05	1.00	0.80	0.18	0.01	1.00
Avg Note Rating	3.42	1.26	1.00	5.00	3.39	1.17	1.00	5.00
Observations	2631				6506			

Table A2: Descriptive Statistics: Location

	UVM				Non-UVM			
	Mean	SD	Min	Max	Mean	SD	Min	Max
AI Adoption	0.67	0.47	0.00	1.00	0.81	0.40	0.00	1.00
Notes (Monthly)	20.33	36.19	0.00	247.00	52.25	70.95	0.00	420.00
Audio Duration (Min/Physician)	385.58	737.61	0.00	6514.49	826.08	1179.24	0.00	6554.73
Keep AI Notes	0.24	0.43	0.00	1.00	0.40	0.49	0.00	1.00
Accepted AI Content	0.83	0.16	0.01	1.00	0.81	0.17	0.02	1.00
Avg Note Rating	3.36	1.15	1.00	5.00	3.49	1.31	1.00	5.00
Observations	6626				2511			

Table A3: Descriptive Statistics: Gender

	Female				Male			
	Mean	SD	Min	Max	Mean	SD	Min	Max
AI Adoption	0.75	0.43	0.00	1.00	0.66	0.47	0.00	1.00
Notes (Monthly)	31.48	51.61	0.00	420.00	26.21	48.48	0.00	410.00
Audio Duration (Min/Physician)	567.44	934.37	0.00	6514.49	430.13	854.91	0.00	6554.73
Keep AI Notes	0.32	0.47	0.00	1.00	0.25	0.44	0.00	1.00
Accepted AI Content	0.81	0.17	0.01	1.00	0.85	0.16	0.05	1.00
Avg Note Rating	3.39	1.22	1.00	5.00	3.44	1.23	1.00	5.00
Observations	5223				3804			

Table A4: How Eligibility Shapes AI Scribe Adoption and Use

	(1) Adoption	(2) Avg Audio Duration	(3) AI Acceptance Rate	(4) Note Rating
$Post$	0.590***	8.003***	0.807***	0.085***
	(0.014)	(0.294)	(0.009)	(0.005)
Observations	34151	34151	27639	34151
R-squared	0.795	0.605	0.974	0.320

Notes: This table shows the estimated effects of AI scribe eligibility on physician adoption and use, based on the two-way fixed effects model in Equation 1. Each column reports the coefficient on $Post$, an indicator equal to one after the AI scribe eligibility date, from a separate regression. The outcomes are: adoption, an indicator for whether the physician used the AI scribe (column 1); average audio duration per encounter in minutes (column 2); AI acceptance, which is the share of AI-generated content left unaltered by the physician (column 3); and note rating, an indicator for whether the physician rated an encounter note involving the AI scribe (column 4). All models include physician and month fixed effects, and the estimation sample is restricted to the first 12 months. Standard errors, clustered at the physician level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B Online Appendix: Learning Dynamics

To better understand learning dynamics, we estimate a separate model that replaces event time with experience - defined as time elapsed since the physician first adopted the tool:

$$y_{it} = \alpha + \sum_{b \neq 0} \theta_b \cdot \mathbf{1}[\text{Experience}_{it} = b] + \gamma_i + \delta_t + \varepsilon_{it} \tag{4}$$

where $\text{Experience}_{it} = b$ indicates that physician i is in experience bin b at time t, and θ_b captures the average outcome in bin b relative to the reference bin $b = 0$. Physician fixed effects γ_i absorb time-invariant heterogeneity across clinicians, and time fixed effects δ_t control for seasonality. The coefficients θ_b trace the within-physician productivity trajectory as experience with the tool accumulates, isolating learning from selection into adoption. The rapid adoption of AI scribes raises a natural question: does experience with the tool further reshape clinical workflow? To examine dynamic effects, we estimate models with physician and year-month fixed effects that exploit variation in exposure duration following eligibility. We define physician experience relative to the first month post-eligibility and group physicians into 1 month, 2–5 months, and 6–12 months of AI exposure.

— Table B1 about here —

Table B1 provides a comprehensive overview of how physician interaction with AI scribes evolves over the first year of use, tracing four distinct dimensions of practice as experience accumulates. Relative to a physician's first month of eligibility (the omitted reference category), the estimates report how each outcome shifts once a physician has accrued 2–5 months and then 6–12 months of exposure. The analysis begins with usage and technical engagement (columns 1–2). As experience accumulates, physicians steadily increase their documentation volume, generating approximately 4.8 additional notes in the 2–5 month window and roughly 9.8 additional notes by 6–12 months relative to their first month (column 1). Column (2) shows that average audio duration rises in tandem, increasing by about 0.99 minutes per note in the 2–5 month cohort and by 2.97 minutes

by 6–12 months. Rather than a one-off calibration spike that later reverses, the pattern is one of deepening engagement: as physicians grow more comfortable with the system, they route progressively more of the clinical encounter through it. Columns (3) and (4) capture how physicians evaluate the AI's output. The AI acceptance rate (column 3), i.e. the share of raw AI-generated text accepted without editing, rises rather than declines with experience, increasing by 4.0 percentage points at 2–5 months and 10.8 percentage points by 6–12 months. The note rating propensity margin in column (4) moves in the same positive direction, improving modestly from 2.8 percentage points at 2–5 months to 3.6 percentage points at 6–12 months. Together these columns indicate that experienced physicians not only accept more of the AI's text but also evaluate the resulting notes somewhat more. There are multiple plausible explanations for these observed dynamics. An efficiency story suggests that the AI scribe reduces the marginal cost of documentation, enabling physicians to sustain higher note volumes as they learn it. A learning story is supported by the joint rise in audio duration and acceptance: physicians provide progressively more input to the system and, in turn, accept more of its output as they master it as a collaborative partner. Finally, a compositional change may be at play, where the types of encounters or the profile of physicians utilizing the tool shift as it moves from early adopters to the broader clinical population. Taken together, this progression suggests that experience with AI scribes is complementary rather than substitutive: across all four margins, deeper exposure is associated with greater rather than diminished reliance on AI scribes.

B.1 Specialty: Primary Care versus Non-Primary Care

For experience, Table B2 shows that the general patterns documented in Table B1 hold across both specialty groups, but with pronounced differences in magnitude. Primary care physicians exhibit substantially larger usage increases as experience accumulates, generating roughly 23.5 additional notes at 2–5 months and 35.5 by 6–12 months, compared with only 0.7 and 4.2 additional notes for non-primary care physicians which is consistent with their higher baseline documentation burden. The same ordering appears on the technical margin: average audio duration rises

with experience for both groups, but far more steeply for primary care physicians (+3.83 then +6.74 minutes) than for non-primary care physicians (+0.37 then +2.15 minutes). Acceptance of AI content follows suit: primary care physicians become markedly more likely to accept raw AI text without editing (18.1 then 30.5 percentage points), whereas the corresponding gains for non-primary care physicians, though positive and significant, are roughly an order of magnitude smaller (1.5 then 7.3 percentage points). The average likelihood to engage in an encounter rating improve for both groups; for primary care physicians the effect is strongest early (0.117 at 2–5 months, attenuating to 0.073 by 6–12 months), while for non-primary care physicians it grows with experience (0.008 to 0.027). Every estimate is positive and statistically significant at the one percent level, and the consistently larger primary care coefficients are in line with these physicians deriving greater documentation leverage from the AI scribe as they gain experience.

B.2 Location: Headquarters versus Satellites

For experience, Table B3 shows that the general patterns documented in Table B1 hold across both headquarters and satellite sites, but with notable differences in magnitude. In contrast to a simple remote-demand story, headquarters physicians exhibit the larger usage increases as experience accumulates, generating roughly 7.0 additional notes at 2–5 months and 12.8 by 6–12 months, compared with 3.6 and 8.0 additional notes for satellite physicians. The same ordering holds on the technical margin: average audio duration rises with experience for both groups, but more steeply at headquarters (+2.06 then +5.73 minutes) than at satellite sites (+0.53 then +1.30 minutes). Acceptance of AI content follows suit: headquarters physicians become more likely to accept raw AI text without editing (7.6 then 19.6 percentage points), while the corresponding gains for satellite physicians, though positive and significant, are smaller (2.7 then 6.4 percentage points). The propensity to engage in rating an encounter also improve for both groups and strengthen with experience, again by more at headquarters (0.045 then 0.060) than at satellite sites (0.017 then 0.020). Every estimate is positive and statistically significant at the one percent level, and the uniformly larger headquarters coefficients indicate that physicians at the main site deepen their reliance on

the AI scribe more markedly than those at remote locations as experience grows.

B.3 Gender: Male versus Female Physicians

For experience, Table B4 shows that the general patterns documented in Table B1 hold across both female and male physicians, but with notable differences in magnitude. Female physicians exhibit the larger usage increases as experience accumulates, generating roughly 5.5 additional notes at 2–5 months and 11.5 by 6–12 months, compared with 4.0 and 8.0 additional notes for male physicians. The same ordering holds on the technical margin: average audio duration rises with experience for both groups, but more steeply for female physicians ($+1.24$ then $+3.62$ minutes) than for male physicians ($+0.69$ then $+2.20$ minutes). The pattern carries over to AI text acceptance, where both groups now show significant, experience increasing gains: female physicians become more likely to accept raw AI text without editing (5.1 then 13.2 percentage points), while male physicians exhibit the same positive pattern at somewhat smaller magnitudes (2.8 then 8.2 percentage points). The propensity to engage in an encounter rating likewise improve for both groups and strengthen with experience, again by slightly more for female physicians (0.034 then 0.043) than for male physicians (0.021 then 0.028). Every estimate is positive and statistically significant at the one percent level, and the consistently larger female coefficients indicate that female physicians deepen their engagement with the AI scribe somewhat more than their male counterparts as experience accumulates.

Table B1: Heterogeneous Effect of Experience on Usage

	(1) Usage (# Notes)	(2) Avg Audio Duration	(3) AI Acceptance Rate	(4) Note Rating
Exp. 2–5 Months	4.767***	0.985***	0.040***	0.028***
	(0.467)	(0.094)	(0.003)	(0.003)
Exp. 6–12 Months	9.809***	2.971***	0.108***	0.036***
	(0.736)	(0.173)	(0.006)	(0.003)
Observations	34151	34151	27639	34151
R-squared	0.611	0.511	0.608	0.296

Notes: This table shows how physician use of the AI scribe evolves with experience over the first year of eligibility. Each column reports estimates from a separate regression of the outcome on experience-bin indicators, with physician and year-month fixed effects. Experience is measured relative to the physician's first month of eligibility, which is the omitted reference category; the reported coefficients give the average shift in each outcome once a physician has accrued 2–5 months and then 6–12 months of exposure. The outcomes are: usage, the number of notes generated (column 1); average audio duration per note in minutes (column 2); AI acceptance rate, which is the share of raw AI-generated text accepted without editing (column 3); and note rating, which is the propensity to rate an encounter note involving the AI scribe (column 4). Columns 1–2 are in levels (notes and minutes); columns 3–4 are changes in propensities or shares. The estimation sample is restricted to the first 12 months. Standard errors, clustered at the physician level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B2: Experience Effects - Primary Care versus Non-Primary Care

	(1) Usage (# Notes)	(2) Avg Audio Duration	(3) AI Acceptance Rate	(4) Note Rating
Panel A: Primary Care				
Exp. 2–5 Months	23.483***	3.827***	0.181***	0.117***
	(2.262)	(0.427)	(0.018)	(0.014)
Exp. 6–12 Months	35.512***	6.743***	0.305***	0.073***
	(3.074)	(0.578)	(0.025)	(0.012)
Observations	5611	5611	4054	5611
R-squared	0.675	0.632	0.753	0.364
Panel B: Non-Primary Care				
Exp. 2–5 Months	0.701***	0.366***	0.015***	0.008***
	(0.109)	(0.052)	(0.002)	(0.001)
Exp. 6–12 Months	4.237***	2.152***	0.073***	0.027***
	(0.481)	(0.160)	(0.005)	(0.003)
Observations	28540	28540	23585	28540
R-squared	0.474	0.419	0.489	0.211

Notes: This table shows how physician use of the AI scribe evolves with experience over the first year of eligibility, separately for primary care physicians (Panel A) and non-primary care physicians (Panel B). Each column reports estimates from a separate regression of the outcome on experience-bin indicators, with physician and year-month fixed effects. Experience is measured relative to the physician's first month of eligibility, which is the omitted reference category; the reported coefficients give the average shift in each outcome once a physician has accrued 2–5 months and then 6–12 months of exposure. The outcomes are: usage, the number of notes generated (column 1); average audio duration per note in minutes (column 2); AI acceptance rate, which is the share of raw AI-generated text accepted without editing (column 3); and note rating, which is the propensity to rate an encounter note involving the AI scribe (column 4). Columns 1–2 are in levels (notes and minutes); columns 3–4 are changes in propensities or shares. The estimation sample is restricted to the first 12 months. Standard errors, clustered at the physician level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B3: Experience Effects - Headquarters versus Satellites

	(1) Usage (# Notes)	(2) Avg Audio Duration	(3) AI Acceptance Rate	(4) Note Rating
Panel A: Headquarters				
Exp. 2–5 Months	6.951***	2.058***	0.076***	0.045***
	(0.763)	(0.203)	(0.007)	(0.006)
Exp. 6–12 Months	12.827***	5.727***	0.196***	0.060***
	(1.063)	(0.358)	(0.012)	(0.006)
Observations	14365	14365	9355	14365
R-squared	0.597	0.479	0.637	0.270
Panel B: Satellites				
Exp. 2–5 Months	3.593***	0.534***	0.027***	0.017***
	(0.589)	(0.088)	(0.004)	(0.003)
Exp. 6–12 Months	8.045***	1.302***	0.064***	0.020***
	(0.996)	(0.141)	(0.006)	(0.003)
Observations	19786	19786	18284	19786
R-squared	0.620	0.586	0.636	0.331

Notes: This table shows how physician use of the AI scribe evolves with experience over the first year of eligibility, separately for physicians at headquarters (Panel A) and satellite sites (Panel B). Each column reports estimates from a separate regression of the outcome on experience-bin indicators, with physician and year-month fixed effects. Experience is measured relative to the physician's first month of eligibility, which is the omitted reference category; the reported coefficients give the average shift in each outcome once a physician has accrued 2–5 months and then 6–12 months of exposure. The outcomes are: usage, the number of notes generated (column 1); average audio duration per note in minutes (column 2); AI acceptance rate, which is the share of raw AI-generated text accepted without editing (column 3); and note rating, which is the propensity to rate an encounter note involving the AI scribe (column 4). Columns 1–2 are in levels (notes and minutes); columns 3–4 are changes in propensities or shares. The estimation sample is restricted to the first 12 months. Standard errors, clustered at the physician level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B4: Experience Effects - Female versus Male Physicians

	(1) Usage (# Notes)	(2) Avg Audio Duration	(3) AI Acceptance Rate	(4) Note Rating
Panel A: Female				
Exp. 2–5 Months	5.525***	1.244***	0.051***	0.034***
	(0.682)	(0.142)	(0.005)	(0.004)
Exp. 6–12 Months	11.457***	3.621***	0.132***	0.043***
	(1.085)	(0.258)	(0.009)	(0.005)
Observations	18394	18394	14823	18394
R-squared	0.618	0.510	0.619	0.312
Panel B: Male				
Exp. 2–5 Months	3.960***	0.690***	0.028***	0.021***
	(0.639)	(0.119)	(0.004)	(0.004)
Exp. 6–12 Months	7.973***	2.201***	0.082***	0.028***
	(0.987)	(0.222)	(0.008)	(0.004)
Observations	15347	15347	12508	15347
R-squared	0.604	0.510	0.593	0.262

Notes: This table shows how physician use of the AI scribe evolves with experience over the first year of eligibility, separately for female physicians (Panel A) and male physicians (Panel B). Each column reports estimates from a separate regression of the outcome on experience-bin indicators, with physician and year-month fixed effects. Experience is measured relative to the physician's first month of eligibility, which is the omitted reference category; the reported coefficients give the average shift in each outcome once a physician has accrued 2–5 months and then 6–12 months of exposure. The outcomes are: usage, the number of notes generated (column 1); average audio duration per note in minutes (column 2); AI acceptance rate, which is the share of raw AI-generated text accepted without editing (column 3); and note rating, which is the propensity to rate an encounter note involving the AI scribe (column 4). Columns 1–2 are in levels (notes and minutes); columns 3–4 are changes in propensities or shares. The estimation sample is restricted to the first 12 months. Standard errors, clustered at the physician level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.