From Conversation to Chart: An Analysis of Clinician Edits to Ambient AI Draft Notes

Yawen Guo MISM[1], Di Hu MS[1], Yiliang Zhou MS[1], Tianchu Lyu PhD[1], Sairam Sutari[2], Steven Tam MD[3], Emilie Chow MD[3], Danielle Perret MD[4], Deepti Pandita MD[3] , Kai Zheng PhD[1]

1. Department of Informatics, University of California, Irvine, Irvine, CA, USA
2. Institute for Clinical and Translational Science, University of California, Irvine, Irvine, CA, USA
3. UCI Health, University of California, Irvine, Irvine, CA, USA
4. Department of Physical Medicine & Rehabilitation, University of California, Irvine, Irvine, CA, USA

Corresponding Author: Yawen Guo

Mailing Address: 6136 Donald Bren Hall, Irvine, CA 92697, United States

E-mail: yaweng5@uci.edu

Telephone: (412) 641-0453

Keywords: ambient listening; artificial intelligence; documentation; electronic health records[E05.318.308.940.968.625.500]; clinician editing behavior; workflow efficiency

Structured Abstract:

Objective

Ambient artificial intelligence (AI) tools are increasingly adopted in clinical practices. This study investigated whether and how clinicians edit AI-generated drafts and the linguistic differences between AI drafts and clinician-finalized notes entered into the electronic health records.

Materials and Methods

This retrospective study analyzed real-world data from ambulatory clinics at a large academic health system. We quantified clinicians' editing behavior by using the Myers diff algorithm to compare AI drafts and final documentation. We then applied statistical and linguistic analysis to study factors associated with the frequency/intensity of editing across note sections (e.g., history of present illnesses vs. assessment and plan), turnaround time, clinician characteristics (e.g., speciality), and encounter types (e.g., in-person vs. telehealth).

Results

Across 20,198 notes drafted by ambient AI, 84.4% were edited by clinicians before signing. While rates of unedited notes differed by note section and care setting, the dominant source of variation was individual clinician practice style rather than specialty-level norms. Notes signed after 24 hours are associated with fewer edits. The final versions became shorter, and exhibit slightly higher lexical diversity and readability. Editing is most intensive in the assessment and plan section, compared with other note sections, and is higher in specialty than primary care.

Conclusion and Discussion

A majority of AI-drafted clinical notes were edited by clinicians, even though the editing rate varies across note sections, medical specialties, and individual clinicians. Future research is needed to fully analyze this editing behavior to inform improvement in AI-assisted clinical documentation to achieve better documentation quality, efficiency, and clinician satisfaction.

Background and Significance

Clinical documentation is essential to outpatient care but remains a major driver of after-hours "pajama time" and clinician burnout. [1–3] Ambient artificial intelligence (AI) documentation systems have emerged to capture visit conversations and draft notes that integrate into the electronic health record (EHR) for clinician to review. [4,5] Prior studies report that clinicians are generally satisfied with ambient AI tools, citing reduced documentation burden and improvements in workload, work–life integration, and patient engagement. [6–10] Despite this promise, clinicians still spend substantial time reviewing, refining, and signing off notes, and analysis on time savings alone do not reveal how clinicians adopt and modify AI drafts in routine practice. Most evaluations of ambient AI documentation to date rely on before-and-after implementation comparisons, assessing time-based workflow efficiency outcomes such as time spent on note writing. [11–13] Early evidence suggests that following deployment, clinicians spend less time on documentation while AI-assisted notes become longer, raising concerns about note bloat, in which increased length may dilute clinically important content and increase downstream review burden. [6,14]

In contrast, routine editing behavior in real-world practice remains poorly characterized. It is still unclear whether clinicians sign ambient AI-drafted notes without changes, how they revise draft, and how editing behaviors vary by different encounter characteristics. Empirical evidence is also limited in how ambient AI relates to note completion timeliness within everyday ambulatory workflows. Understanding these patterns is essential because documentation practices and information needs differ substantially across clinical settings such as specialties and visit types,

and ambient AI documentation should support workflow-specific needs without shifting effort into additional review and correction work. [2,15,16]

To address this gap, we analyzed a large dataset of paired ambient AI–drafted note and clinician-finalized text from a large academic health system, where two commercially available ambient AI documentation systems are integrated into ambulatory EHR workflows. Our analyses addressed three questions: (1) how often AI drafts are submitted without edits and whether no-edit behavior reflects individual clinician writing style versus specialty norms, as well as how it varies by note section, encounter type, and specialty; (2) among edited note sections, how clinician-edited text differs from the AI draft in linguistic characteristics, including length, lexical diversity, and readability, and how the magnitude of edits varies by note sections and care settings; and (3) how turnaround time relates to edit intensity at a token level, and which note sections, encounter types, and specialties are associated with more editing and longer turnaround. Together, these analyses provide a detailed picture of how clinicians interact with ambient AI documentation in routine outpatient practice.

Method

Setting, Data Source, and Unit of Analysis. A pilot deployment of ambient AI documentation was initiated at University of California, Irvine Medical Center (UCI Health) in 2023. UCI Health is an academic health system with ambulatory clinics across multiple sites in Southern California. Participating physicians from selected primary care and specialty clinics used two commercially available ambient AI tools integrated with the Epic EHR (DAX Copilot and Abridge) [17,18] . Clinicians accessed the tools through Haiku mobile application or through a browser-based interface connected to Epic. In all workflows, AI-generated text was reviewed,

edited as needed, and signed within the EHR, and clinicians retained full responsibility for the final documentation. Clinicians received onboarding and workflow guidance as part of the pilot rollout, with ongoing technical support available during the study period. Consistent with institutional procedures for the pilot, clinicians informed patients at the visit that audio-assisted note drafting might be used, that the resulting documentation would be clinician-reviewed before filing, and that identifiable audio was not retained. Liability and regulatory considerations were managed through institutional compliance and IT review, and the EHR audit trail preserved both the initial AI-generated draft and the final signed version. Because vendor attribution was not available in the analytic dataset under institutional data governance and reporting policies, analyses are presented in pooled form across the two tools.

We extracted all ambient AI note sections from Epic between September 26, 2024, and August 5, 2025. In routine clinical use, ambient AI generates draft text at the note-section level, and documentation is completed and signed at the encounter note level. Clinicians can selectively choose one or more AI-drafted note sections to insert into an EHR note. Four types of note sections can be generated by the ambient AI tools, including History of Present Illness (HPI), Assessment & Plan (A&P), Physical Exam, and Results. We analyzed workflow outcomes that are naturally defined per encounter (e.g., turnaround time and encounter-level unedited use) at the note level by aggregating note section-level data within each unique note ID, and then conducted section-level analyses to characterize where edits concentrate across note-section types.

Prevalence of Unedited Note Sections by Clinical Context (Descriptives and χ^2 Tests). To quantify how often ambient AI drafts were accepted as-is, we first computed a note-level no-edit indicator (1 = all AI-generated sections within a note were signed without edits; 0 = otherwise)

and summarized the prevalence of unedited use. We then conducted a section-level drill-down by defining a binary section-level indicator (1 = no-edit, 0 = edited) and stratifying unedited prevalence by note section, care setting (primary vs specialty care), and encounter type. For each dimension, we used χ^2 tests of independence to compare the distribution of unedited vs edited sections across categories, reporting χ^2 statistics, degrees of freedom, and p-values.[19]

Unedited Use as Personal Style vs Specialty Norms (Mixed-Effects Logistic Regression). To distinguish individual documentation style from broader specialty norms, we fit a mixed-effects logistic regression model restricted to clinicians who have used ambient AI to draft over 100 note sections.[20] The binary outcome indicated whether a section was submitted without edits. Fixed effects included note section, encounter type, and clinician group (specialty vs primary care). We specified random intercepts for individual clinician and specific medical specialty (e.g., cardiology, neurology, etc.) and fit a binomial mixed-effects model (BinomialBayesMixedGLM).[21] Including clinician group as a fixed effect captures broad differences between primary and specialty care, while the specialty random intercept captures residual variation across specific specialties beyond that grouping. From this model we obtained fixed-effect log-odds estimates and variance components, and computed intraclass correlation coefficients (ICC) to estimate the proportion of explainable variation in unedited behavior attributable to individual clinicians versus specialties.

Text-Based Linguistic Metrics Before and After Editing. For note sections that were edited, we computed eight metrics for both AI draft and final version: character count, word count, sentence count, unique word count, type–token ratio (unique words divided by total words), average word length, Flesch Reading Ease, and Flesch–Kincaid Grade Level.[22] Within each note section we calculated a before–after difference for each metric by subtracting the AI draft from the

clinician-final text (Δ = Final − AI). To quantify overall editing, we used paired t-tests to test whether the mean Δ differed from zero.[23] To compare editing magnitude between primary-care and specialty-care clinicians, we compared Δ values using independent-samples t-tests and Mann–Whitney U tests.[24,25] To examine variation by note section, we compared Δ across HPI, A&P, Physical Exam, and Results using one-way ANOVA with Tukey's HSD post-hoc tests for pairwise contrasts.[26]

Quantifying Edit Intensity with Myers Diff. For all edited sections, we aligned the AI draft and final note using a token-level Myers diff algorithm.[27] This approach measures edits at the token level, where each token corresponds to a word or punctuation unit. We defined total edit intensity as the total number of tokens changed, calculated as the sum of inserted, deleted, and replaced tokens.

Factors Associated With Longer Turnaround Time (Binomial Logistic Regression). To identify which notes were more likely to have turnaround over 24 hours, we fit a binomial generalized linear model with a logit link using the binary turnaround indicator as the outcome, and encounter type, clinician specialty, and the transformed number of AI-generated note sections included in each note as predictors. Encounter types and specialties with fewer than 20 notes, or with 0% or 100% turnaround prevalence over 24 hours, were collapsed into pooled "Other" categories.[28]

Turnaround Time and Editing Intensity (Mixed-Effects Linear Model). A binary turnaround indicator was coded as 1 when time from AI draft generation to note signing exceeded 24 hours, and 0 otherwise. Because signing occurs at the note level and the draft and signing dates were identical across sections within a note, we conducted turnaround analyses at the note level by aggregating section-level edits within each note. To test whether later submissions received more

edits, we first compared total note-level edit intensity (sum of inserted, deleted, and replaced tokens across all sections in the note) between notes signed within 24 hours and those signed after 24 hours using descriptive statistics and a Mann–Whitney U test. We restricted the dataset to clinicians who have used ambient AI to draft over 100 note sections. We then fit a mixed-effects model with total edit intensity per note as the outcome and turnaround time indicator as the primary predictor, adjusting for encounter type, clinician group (primary vs specialty care), and the transformed number of AI-generated note sections included in each note as fixed effects, and including an individual clinician-level random intercept.[29]

Factors Associated With Heavier Editing Intensity (OLS With Cluster-Robust SEs). To examine which note sections, encounter types, and specialties were associated with heavier editing, we fit a regression model with log-transformed edit intensity as the outcome, and note section, encounter type, and clinician specialty as categorical predictors, using clinician-clustered robust standard errors to account for within-clinician correlation (i.e., clustering on clinician to allow correlation among sections authored by the same clinician).[30] Encounter types and specialties with fewer than 50 sections were pooled into "Other" categories.

Ethical Oversight and Data Security. The work was approved by the University of California, Irvine Institutional Review Board (IRB #7234). All data processing occurred within HIPAA-compliant secure environments, including UCI Health's Protected Virtual Computing Environment (PVCE) and Amazon Web Services (AWS) HIPAA-aligned infrastructure.

Results

Dataset Characteristics

The dataset included 23,911 unique notes containing 48,460 AI-drafted note sections, authored by 225 clinicians for 16,520 patients (Table 1). We summarized characteristics at the note-section level because clinicians can selectively insert and edit individual AI-generated smart elements, and a single note may contain multiple AI-drafted sections. Nearly half of sections were HPI (49.8%), followed by A&P (34.0%), Results (9.3%), and Physical Exam (6.9%). Most sections were written by physicians (MD/DO, 91.3%), with additional contributions from nurse practitioners and physician assistants (7.6%). Progress notes accounted for 97.6% of sections; the remainder were consults, History and Physical (H&P) notes, or procedure notes. Most encounters were office visits (88.5%), with 9.5% from video or telemedicine visits and 2.0% from other ambulatory visits (integrative medicine, procedure, OB, erroneous encounters, tumor board). Primary-care specialties contributed roughly half of all sections; the rest came from a broad mix of medical and surgical specialties .

Table 1. Study Cohort and Dataset Characteristics

Category	Subcategory	Value
General	Study period	2024-09-26 – 2025-08-05
	Total notes (N)[a]	23,911
	Total note sections	48,460
	Unique clinicians (N)	225

	Unique patients (N)	16,520
Section name (per note section) [b]	HPI	24,119 (49.8%)
	A&P	16,492 (34.0%)
	Results	4,521 (9.3%)
	Physical Exam	3,328 (6.9%)
Clinician license (per note section) [b]	MD, DO	44,221 (91.3%)
	NP, PA	3,673 (7.6%)
	OD, MD-PhD, LVN	25 (<0.1%)
Note type (per note section) [b]	Progress Note	47,291 (97.6%)
	Consults / History and Physical (H&P) / Procedure Note	1,169 (2.4%)
Encounter type (per note section) [b]	Office Visit	42,900 (88.5%)
	Video Visit / Telemedicine	4,614 (9.5%)
	Other encounter types [c]	946 (2.0%)

Clinician specialty (per note section) [b]	Family Practice	12,053 (24.9%)
	Internal Medicine	6,444 (13.3%)
	Geriatric Medicine	4,103 (8.5%)
	Nurse Practitioner	1,396 (2.9%)
	Physician Assistant	1,362 (2.8%)
	Rheumatology	5,913 (12.2%)
	Orthopaedic Surgery	4,478 (9.2%)
	Hand Surgery	3,995 (8.2%)
	Neurology	1,567 (3.2%)
	Ob/Gyn	962 (2.0%)
	Gastroenterology	866 (1.8%)
	Physical Medicine and Rehab	573 (1.2%)
	Neurosurgery	535 (1.1%)

	Cardiology	483 (1.0%)
	Other specialty (<1.0% each) [d]	2,013 (4.2%)

[a] To reduce the risk of sensitive content, we excluded VIP patients, confidential/sensitive encounters or notes, sensitive departments (e.g., psychiatry/behavioral therapy/substance use/addiction), emergency department–origin encounters, psychotherapy notes, and any notes explicitly marked sensitive.

[b] Unless otherwise noted, values are n (%) of note sections (N = 48,460).

[c] Other encounter types = Integrative Medicine, Procedure Visit, OB Visit, Erroneous Encounter, Tumor Board.

[d] Other specialty (<1.0% each) = Ophthalmology, Epilepsy, Otolaryngology, Hospitalist, Endocrinology, Dermatology, Colon and Rectal Surgery, Integrative Medicine, Emergency Medicine, Infectious Diseases, Optometry, General Surgery, Psychiatry, Urology, Vascular Neurology, Hematology/Oncology, Orthopedics, Vascular Surgery, Head and Neck Surgery, Surgical Oncology.

How Often Are AI Drafts Signed Without Edits?

At the note level, 15.6% of 20,198 notes containing one or multiple AI-generated note sections were signed without edits to the AI-generated content. Across these notes, there were 48,460 AI-generated note sections, of which 14,670 (30.3%) were signed without edits. Prevalence unedited use differed by note section ($\chi^2(3) = 5{,}685.6$, $p < 0.001$), ranging from 13.2% for A&P sections to 32.8% for HPI, 46.5% for Physical Exam, and 67.2% for Results. Primary-care clinicians were more likely than specialty-care clinicians to submit drafts without edits (33.4% vs

27.8%; $\chi^2(1) = 172.3$, $p < 0.001$). No-edit rates also varied by encounter type ($\chi^2(7) = 173.2$, $p < 0.001$): office visits had a 30.8% no-edit rate, compared with 25.7% for video visits, 34.6% for telemedicine encounters (Table 2). We conducted a supplementary specialty-level analysis across all clinician specialties and found that some high-volume specialties (e.g., ophthalmology) had relatively high no-edit rates, whereas others (e.g., neurology, endocrinology, epilepsy, and physical medicine and rehabilitation) edited the majority of AI drafts (Supplementary Table S1).

Table 2. Section-Level No-Edit Rates by Clinician Group, Encounter Type, and Note Section

Dimension	Category	Total sections, n	No-edit sections, n (%)	χ^2 (df)	P value
Clinician group	Primary Care	25,358	8,473 (33.4%)	172.3(1)	<0.001
	Specialty Care	21,385	5,942 (27.8%)		
Encounter type	Office Visit	42,900	13,205 (30.8%)	173.2(7)	<0.001
	Video Visit	2,997	769 (25.7%)		
	Telemedicine	1,617	559 (34.6%)		
	Integrative Medicine	614	77 (12.5%)		

	Procedure Visit	215	39 (18.1%)		
	OB Visit	103	16 (15.5%)		
	Erroneous Encounter	8	5 (62.5%)		
	Tumor Board	6	0 (0.0%)		
Note section	HPI	24,119	7,905 (32.8%)	5685.6(3)	<0.001
	A&P	16,492	2,179 (13.2%)		
	Results	4,521	3,040 (67.2%)		
	Phys Exam	3,328	1,546 (46.5%)		

Is "no-edit" behavior driven by individual writing styles or specialty norms?

Unedited submissions were driven more by individual clinician style than by specialty norms. In the mixed-effects logistic regression, specialty-care clinicians had slightly lower log-odds of signing AI note drafts without edits than primary-care clinicians ($\beta_Specialty = -0.17$, $p < 0.001$), consistent with the descriptive differences. Random-effects variance was concentrated at the clinician level (Var = 3.24; ICC = 0.50) with negligible variance attributable to specialty (Var = 0.001; ICC ≈ 0.00), indicating that no-edit behavior varied primarily across individual clinicians rather than across specialties after adjustment for note section and encounter type (Supplementary Table S2).

Text-Based Linguistic Differences Between AI Drafts and Final Note Sections

Paired comparisons showed small but statistically significant differences between AI drafts and final note sections across most linguistic metrics (Table 3). Final notes had fewer sentences (15.64 vs 14.43; $\Delta = -1.21$; $p < 0.001$) and slightly fewer words (194.67 vs 194.10; $\Delta = -0.57$; $p < 0.001$), with higher Flesch Reading Ease (41.12 vs 42.28; $\Delta = +1.16$; $p < 0.001$) and modest increases in lexical diversity (type–token ratio 0.646 to 0.662; $\Delta = +0.016$; $p < 0.001$). Compared with specialty care, primary care showed larger before–after changes in most metrics (all $p < 0.001$), except average word length ($p = 0.516$).

Table 3. Linguistic differences comparing AI drafts and clinician-final note sections, overall and by clinician group (primary vs specialty care).

Metric	AI draft, mean	Final note, mean	Mean difference (Final − AI)	Mean difference (Final − AI), primary care	Mean difference (Final − AI), specialty care	P value (primary vs specialty)
Average word length (characters)	5.402	5.31	-0.092	-0.087654	-0.092744	p = 0.516
Character	1278.955	1263.486	-15.47	-435.527418	29.298454	p < 0.001

count						
Flesch–Kincaid grade level	10.881	10.937	0.056	-0.064397	0.068871	$p < 0.001$
Flesch Reading Ease	41.119	42.277	1.158	1.990755	1.069707	$p < 0.001$
Sentence count	15.642	14.432	-1.209	-5.269254	-0.776472	$p < 0.001$
Type–token ratio	0.646	0.662	0.016	0.047222	0.012697	$p < 0.001$
Unique word count	116.497	118.82	2.323	-28.80191	5.639525	$p < 0.001$
Word count	194.666	194.097	-0.57	-58.7939	5.635355	$p < 0.001$

Across note sections, changes in linguistic metrics were most pronounced in A&P, followed by HPI, with the smallest changes in Results and Physical Exam. Compared with AI drafts, A&P showed the largest decreases in character count, word count, and sentence count, whereas HPI

showed intermediate changes and Results and Physical Exam showed minimal and often positive changes (Figure 1).

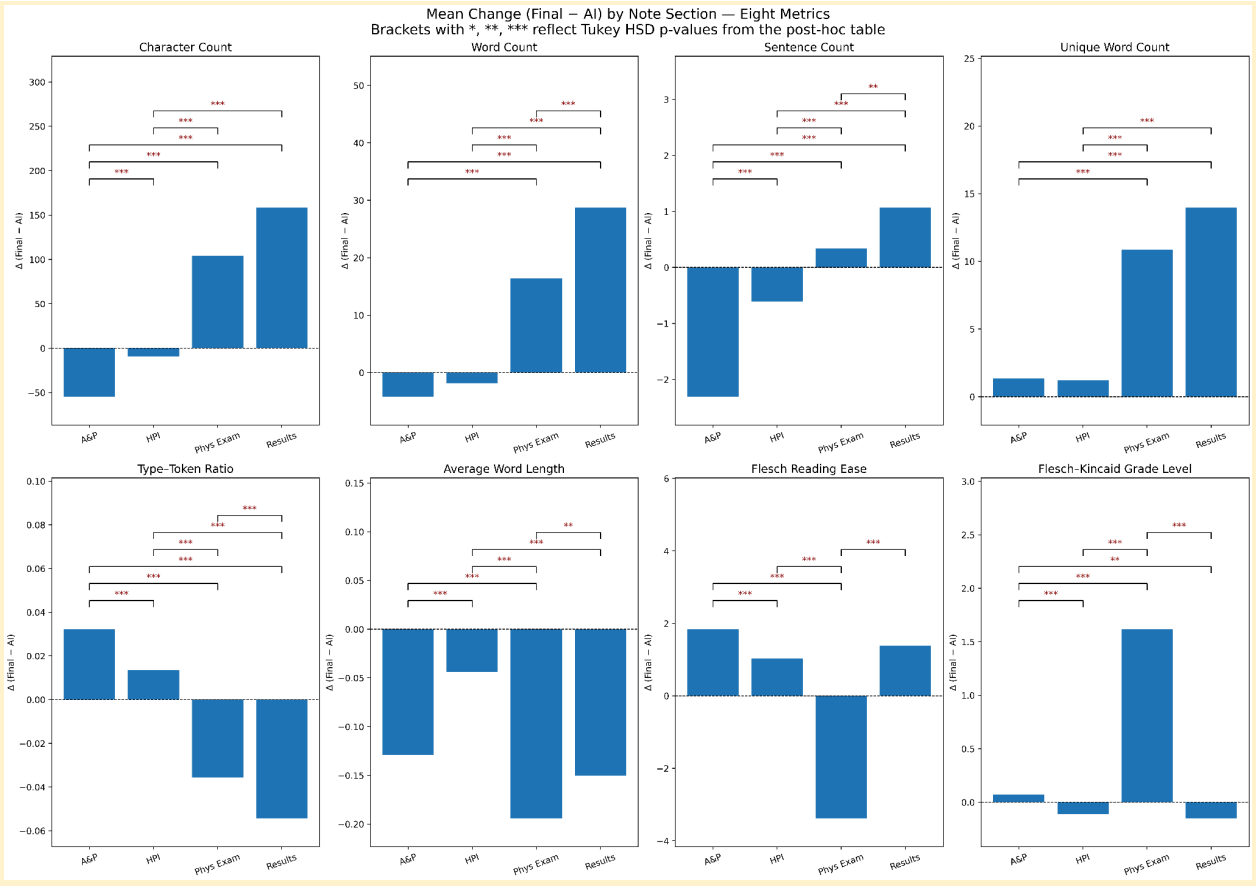

Figure 1. Note Section-Level Comparison of Edits Across Eight Linguistic Metrics

*, **, *** indicate $p < 0.05$, $p < 0.01$, $p < 0.001$.

Which notes are more likely to have turnaround over 24 hours?

Across notes that were edited before signed (n = 20,198), 3,233 (16.0%) had turnaround over 24 hours. The number of AI-generated note sections within one note was associated with turnaround over 24 hours (log[1+sections]: $\beta = 0.99$, $p < 0.001$), indicating that notes containing more AI-generated sections were more likely to be signed more than 24 hours after draft generation.

The encounter type also showed systematic differences. Relative to integrative-medicine visits (reference), OB visits had higher odds of turnaround over 24 hours ($\beta = 1.66$, $p = 0.001$), while telemedicine visits had markedly lower odds ($\beta = -0.92$, $p < 0.001$) and video visits similarly showed shorter turnaround ($\beta = -0.80$, $p = 0.001$); office visits were directionally lower but not statistically significant ($\beta = -0.42$, $p = 0.054$). At the specialty level, family practice clinicians were less likely to have a turnaround over 24 hours ($\beta = -1.83$, $p < 0.001$). Several specialties had longer turnaround time, including dermatology ($\beta = 1.31$, $p < 0.001$), geriatrics ($\beta = 1.87$, $p < 0.001$), infectious diseases ($\beta = 2.68$, $p < 0.001$), and colon and rectal surgery ($\beta = 1.17$, $p = 0.001$) (Supplementary Table S4).

Do notes signed after 24 hours have more or fewer edits?

To examine whether notes signed later had more or fewer edits, we compared total note-level edit intensity between notes signed within 24 hours versus after 24 hours. Among edited notes with valid turnaround times (n = 18,315), 15,310 notes (83.6%) were signed within 24 hours of draft generation and 3,005 (16.4%) were signed after 24 hours. Long-turnaround (>24h) was associated with lower edit loads after adjustment for encounter type, clinician group, and the log-transformed number of AI-generated note sections per note (log[1+sections]). Compared with ≤24h, >24h submissions had a reduction in log edit load ($\beta = -0.216$, $p < 0.001$). The clinician random intercept variance was estimated near zero, indicating minimal residual clustering by individual clinician behavior after these adjustments. (Supplementary Table S3).

Which sections, encounter types, and specialties show more edits?

This analysis included 33,699 edited note sections from 215 clinicians. In the adjusted model, edit intensity differed strongly by note section: compared with A&P, edit intensity was lower for HPI ($\beta = -0.50$, $p = 0.002$), physical exam ($\beta = -1.34$, $p < 0.001$), and results ($\beta = -1.23$, $p <$

0.001). Specialty remained associated with edit intensity relative to cardiology, with higher edit intensity in epilepsy ($\beta = 1.74$, $p < 0.001$), neurology ($\beta = 1.23$, $p < 0.001$), and physical medicine and rehabilitation ($\beta = 0.57$, $p = 0.032$), and lower edit intensity in family practice ($\beta = -0.76$, $p = 0.005$), geriatrics ($\beta = -0.74$, $p = 0.004$), hand surgery ($\beta = -0.72$, $p < 0.001$), hospitalist medicine ($\beta = -1.69$, $p < 0.001$), integrative medicine ($\beta = -0.99$, $p = 0.002$), internal medicine ($\beta = -0.50$, $p = 0.022$), obstetrics/gynecology ($\beta = -0.70$, $p < 0.001$), ophthalmology ($\beta = -1.19$, $p = 0.003$), and orthopaedic surgery ($\beta = -0.49$, $p = 0.016$). Encounter type was not a significant predictor of edit intensity in the adjusted model (all $p \geq 0.17$) (Supplementary Table S5).

Discussion

This study provides empirical evidence on ambient AI documentation usage and edit patterns among clinicians in routine ambulatory care. By linking AI drafts to clinician-final text and quantifying edits, we describe where clinicians accept AI drafts as-is, where they invest effort to revise, and how these patterns relate to turnaround time and care settings, offering a practical evaluation framework that can be reused for operational monitoring and model improvement.[31]

We found that most notes containing AI drafts (84.4%) were edited by clinicians before signing. Even though no-edit rates differed across note sections and specialties, our findings suggested substantial clinician-level heterogeneity with minimal residual clustering by specialty after accounting for section and encounter types. This suggests that the tendency to sign AI-drafted sections without edits is primarily a clinician-level phenomenon. Health systems could therefore target implementation support and workflow coaching to clinicians with

unusually high or low rates of unedited submission, rather than applying training or configuration changes uniformly by specialty. [32,33] These findings also reinforce that ambient AI remains a human-in-the-loop technology, where clinician review and editing are an essential part of safe use and accountability.

When edits occurred, changes were generally modest and consistent, with the greatest editing concentrated in A&P, followed by HPI. This gradient aligns with A&P as the most information-rich note section and the primary site of clinical reasoning, risk management, and decision-making.[34] In later practice, this suggests prioritizing A&P and HPI for model refinement, such as clearer problem organization, and more specific and actionable plans with fewer generic statements, while concentrating implementation monitoring and quality checks on note sections where clinicians invest the most editing effort.[35]

Besides, notes signed more than 24 hours after AI draft generation had lower edit intensity, suggesting that prolonged turnaround is not solely explained by extensive text revision. Longer turnaround may reflect other workflow factors (e.g., clinic scheduling, team documentation practices). Accordingly, efforts to reduce prolonged turnaround may benefit from workflow supports such as end-of-visit signing prompts and tighter integration between the ambient AI tool and EHR workflows. [36,37] In addition, because clinical notes support billing and compliance, longer AI-generated drafts should not be assumed to meet specialty-specific documentation requirements; implementation should include billing/compliance guardrails and specialty-specific guidance alongside quality monitoring.

This study has several limitations. Including only clinical notes from ambulatory settings within a single academic health system may limit generalizability to other institutions and clinical settings. Observed editing patterns may be shaped by local implementation factors,

including system configuration, workflow integration, and clinician training and support. Because these factors vary across health systems, they may influence both the amount and type of clinician editing. Besides, our analyses quantified the frequency and intensity of edits rather than the clinical correctness of the final notes or downstream care outcomes; therefore, we could not determine whether high no-edit use reflected high-quality drafts or missed errors, or whether higher edit levels reflected necessary clinical corrections or individual documentation style.[38] Finally, although two ambient solutions were piloted at our institution, institutional data governance and reporting policies precluded vendor-stratified analyses, so results are presented in pooled form.

Future work should link editing patterns to independent assessments of note quality and safety, and distinguish stylistic edits of individual clinician from clinically impactful changes to support targeted model refinement and note quality monitoring. Furthermore, prospective studies that evaluate vendor-specific performance and systematically document local implementation details could better uncover how vendor and deployment factors contribute to observed editing patterns. Overall, our proposed methods suggest an approach for health systems deploying clinical AI applications beyond ambient documentation, such as AI-drafted patient message replies: routinely track no-edit rates, edit intensity, and turnaround time to identify clinician- and content-specific friction points, and tailor implementation support to individual workflow variation rather than assuming uniform specialty-level patterns. [2,39,40]

Conclusion

In this large-scale evaluation of two EHR-integrated ambient AI documentation systems in ambulatory settings, we found that clinicians edited the majority of AI drafted notes before

signing and without edits appeared to reflect individual clinician practice style more than specialty-level norms. Documentation turnaround time was not explained by higher editing burden, as notes signed more than 24 hours after AI draft generation tended to have lower edit intensity. Section level drill-down analysis further showed that when edits occurred, revisions were generally modest and concentrated in information-dense note sections, particularly the A&P, with specialty care clinicians editing more than primary care. Together, these findings provide a comprehensive understanding of clinician editing behavior for ambient AI drafts to inform targeted model refinement, clinician-facing implementation support, and ongoing governance of ambient AI documentation in real-world practice.

FUNDING STATEMENT

This study did not receive any external funding.

COMPETING INTERESTS STATEMENT

The authors have no competing interests to declare.

CONTRIBUTORSHIP STATEMENT

Yawen Guo (YG) led the study and contributed to the study concept and design, data acquisition, analysis, and drafting of the manuscript. Di Hu (DH) contributed to the study design and manuscript revision. Yiliang Zhou (YZ), Tianchu Lyu (TL) and Sairam Sutari (SS) provided data and administrative support and contributed to manuscript revision. Danielle Perret (DP), Emilie Chow (EC), Steven Tam (ST), and Deepti Pandita (DP) provided guidance through the quality improvement pilot project. Kai Zheng (KZ) provided overall supervision and guidance

throughout the project. All authors critically reviewed the manuscript for important intellectual content and approved the final version.

DATA AVAILABILITY STATEMENT

The underlying dataset contains protected health information (PHI) and is not publicly available. Summary statistics and analysis outputs are provided in the Supplementary Materials.

Reference

[1] Baumann LA, Baker J, Elshaug AG. The impact of electronic health record systems on clinical documentation times: A systematic review. Health Policy 2018;122:827–36. https://doi.org/10.1016/j.healthpol.2018.05.014.

[2] Hassan H, Zipursky AR, Rabbani N, et al. Special Topic on Burnout: Clinical Implementation of Artificial Intelligence Scribes in Healthcare: A Systematic Review. Appl Clin Inform 2025:a-2597-2017. https://doi.org/10.1055/a-2597-2017.

[3] Albrecht M, Shanks D, Shah T, et al. Enhancing clinical documentation with ambient artificial intelligence: a quality improvement survey assessing clinician perspectives on work burden, burnout, and job satisfaction. JAMIA Open 2024;8:ooaf013. https://doi.org/10.1093/jamiaopen/ooaf013.

[4] Hundal J, Jain M, McCollom J. Ambient Artificial Intelligence in Health Care Documentation: A Review of Tools, Integration, and Clinical Implications. AI Precis Oncol 2025.

[5] Gebauer S. Benchmarking And Datasets For Ambient Clinical Documentation: A Scoping Review Of Existing Frameworks And Metrics For AI-Assisted Medical Note Generation. medRxiv 2025:2025–01.

[6] Guo Y, Wang J, Hu D, et al. Evaluating ambient artificial intelligence documentation: effects on work efficiency, documentation burden, and patient-centered care. J Am Med Inform Assoc 2025:ocaf180.

[7] Balloch J, Sridharan S, Oldham G, et al. Use of an ambient artificial intelligence tool to improve quality of clinical documentation. Future Healthc J 2024;11:100157. https://doi.org/10.1016/j.fhj.2024.100157.

[8] 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 JAMIA 2024;31:975–9. https://doi.org/10.1093/jamia/ocae022.

[9] Tierney AA, Gayre G, Hoberman B, et al. Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation. NEJM Catal 2024;5. https://doi.org/10.1056/CAT.23.0404.

[10] Tierney AA, Gayre G, Hoberman B, et al. Ambient Artificial Intelligence Scribes: Learnings after 1 Year and over 2.5 Million Uses. NEJM Catal 2025;6. https://doi.org/10.1056/CAT.25.0040.

[11] Kakaday R, Herrera EZ, Coskey O, et al. The STREAMLINE Pilot – Study on Time Reduction and Efficiency in AI-Mediated Logging for Improved Note-Taking Experience. Appl Clin Inform 2025:a-2559-5791. https://doi.org/10.1055/a-2559-5791.

[12] Ma SP, Liang AS, Shah SJ, et al. Ambient artificial intelligence scribes: utilization and impact on documentation time. J Am Med Inform Assoc 2025;32:381–5. https://doi.org/10.1093/jamia/ocae304.

[13] Cao DY, Silkey JR, Decker MC, et al. Artificial intelligence-driven digital scribes in clinical documentation: Pilot study assessing the impact on dermatologist workflow and patient encounters. JAAD Int 2024;15:149–51. https://doi.org/10.1016/j.jdin.2024.02.009.

[14] Guo Y, Hu D, Wang J, et al. Ambient Listening in Clinical Practice: Evaluating EPIC Signal Data Before and After Implementation and Its Impact on Physician Workload. In: Househ MS, Tariq ZUA, Al-Zubaidi M, et al., editors. Stud. Health Technol. Inform., IOS Press; 2025. https://doi.org/10.3233/SHTI250921.

[15] van Buchem MM, Boosman H, Bauer MP, et al. The digital scribe in clinical practice: a scoping review and research agenda. NPJ Digit Med 2021;4:57. https://doi.org/10.1038/s41746-021-00432-5.

[16] Chishtie J, Sapiro N, Wiebe N, et al. Use of Epic Electronic Health Record System for Health Care Research: Scoping Review. J Med Internet Res 2023;25:e51003. https://doi.org/10.2196/51003.

[17] Generative AI for Clinical Conversations | Abridge n.d. https://www.abridge.com/ (accessed May 5, 2025).

[18] Microsoft Dragon Copilot | Microsoft Cloud for Healthcare n.d. https://www.microsoft.com/en-us/health-solutions/clinical-workflow/dragon-copilot (accessed May 5, 2025).

[19] Cochran WG. Some Methods for Strengthening the Common χ 2 Tests. Biometrics 1954;10:417. https://doi.org/10.2307/3001616.

[20] Mixed Effects Logistic Regression | Stata Data Analysis Examples n.d. https://stats.oarc.ucla.edu/stata/dae/mixed-effects-logistic-regression/ (accessed December 13, 2025).

[21] Josef Perktold, Skipper Seabold, Kevin Sheppard, et al. statsmodels/statsmodels: Release 0.14.2 2024. https://doi.org/10.5281/ZENODO.593847.

[22] Kincaid JP, Fishburne Jr, Robert P. R, et al. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel: Fort Belvoir, VA: Defense Technical Information Center; 1975. https://doi.org/10.21236/ADA006655.

[23] Xu M, Fralick D, Zheng JZ, et al. The Differences and Similarities Between Two-Sample T-Test and Paired T-Test. Shanghai Arch Psychiatry 2017;29:184–8. https://doi.org/10.11919/j.issn.1002-0829.217070.

[24] Ross A, Willson VL. Independent Samples T-Test. Basic Adv. Stat. Tests, Rotterdam: SensePublishers; 2017, p. 13–6. https://doi.org/10.1007/978-94-6351-086-8_3.

[25] McKnight PE, Najab J. Mann‐Whitney U Test. In: Weiner IB, Craighead WE, editors. Corsini Encycl. Psychol. 1st ed., Wiley; 2010, p. 1–1. https://doi.org/10.1002/9780470479216.corpsy0524.

[26] Nanda A, Mohapatra DrBB, Mahapatra APK, et al. Multiple comparison test by Tukey's honestly significant difference (HSD): Do the confident level control type I error. Int J Stat Appl Math 2021;6:59–65. https://doi.org/10.22271/maths.2021.v6.i1a.636.

[27] The Myers Difference Algorithm n.d. https://www.nathaniel.ai/myers-diff/ (accessed December 13, 2025).

[28] Blizzard L, Hosmer W. Parameter Estimation and Goodness-of-Fit in Log Binomial Regression. Biom J 2006;48:5–22. https://doi.org/10.1002/bimj.200410165.

[29] Oberg AL, Mahoney DW. Linear Mixed Effects Models. In: Ambrosius WT, editor. Top. Biostat., vol. 404, Totowa, NJ: Humana Press; 2007, p. 213–34. https://doi.org/10.1007/978-1-59745-530-5_11.

[30] Colin Cameron A, Miller DL. A Practitioner's Guide to Cluster-Robust Inference. J Hum Resour 2015;50:317–72. https://doi.org/10.3368/jhr.50.2.317.

[31] Sloss EA, Abdul S, Aboagyewah MA, et al. Toward Alleviating Clinician Documentation Burden: A Scoping Review of Burden Reduction Efforts. Appl Clin Inform 2024;15:446–55. https://doi.org/10.1055/s-0044-1787007.

[32] Mullankandy S, Mukherjee S, Ingole BS. Applications of AI in Electronic Health Records, Challenges, and Mitigation Strategies. 2024 Int. Conf. Comput. Appl. ICCA, Cairo, Egypt: IEEE; 2024, p. 1–7. https://doi.org/10.1109/ICCA62237.2024.10927863.

[33] Varghese J. Artificial Intelligence in Medicine: Chances and Challenges for Wide Clinical Adoption. Visc Med 2020;36:443–9. https://doi.org/10.1159/000511930.

[34] Stupp D, Barequet R, Lee I-C, et al. Structured Understanding of Assessment and Plans in Clinical Documentation 2022. https://doi.org/10.1101/2022.04.13.22273438.

[35] Stetson PD, Bakken S, Wrenn JO, et al. Assessing Electronic Note Quality Using the Physician Documentation Quality Instrument (PDQI-9). Appl Clin Inform 2012;3:164–74. https://doi.org/10.4338/aci-2011-11-ra-0070.

[36] Nguyen OT, Kunta AR, Katoju S, et al. Electronic Health Record Nudges and Health Care Quality and Outcomes in Primary Care: A Systematic Review. JAMA Netw Open 2024;7:e2432760. https://doi.org/10.1001/jamanetworkopen.2024.32760.

[37] Lindsay MR, Lytle K. Implementing Best Practices to Redesign Workflow and Optimize Nursing Documentation in the Electronic Health Record. Appl Clin Inform 2022;13:711–9. https://doi.org/10.1055/a-1868-6431.

[38] Luo Y-F, Sun W, Rumshisky A. A Hybrid Normalization Method for Medical Concepts in Clinical Narrative using Semantic Matching. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci 2019;2019:732–40.

[39] Hu D, Guo Y, Cho HN, et al. When AI Writes Back: Ethical Considerations by Physicians on AI-Drafted Patient Message Replies 2025. https://doi.org/10.48550/ARXIV.2508.13217.

[40] Hu D, Guo Y, Zhou Y, et al. A systematic review of early evidence on generative AI for drafting responses to patient messages. Npj Health Syst 2025;2:27. https://doi.org/10.1038/s44401-025-00032-5.