A multimodal framework for assessing patient-physician engagement in ambient AI scribe evaluations

Basam Alasaly, MS[1*], Matthew Hill[1], Sriharsha Mopidevi[1], Kevin B. Johnson, MD, MS[1,3], Anna Schoenbaum, DNP, MS, RN-BC[5], Srinivas Sridhara, PhD, MHS[5], Srinath Adusumalli, MD, MSHP, MBMI[2,3,5], Matthew J. Duggan, MD, MBA[5,6], Kaustav P. Shah, MD[2,3,4]

[1]*Division of Biomedical Informatics, Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, US*

[2]*Department of Medicine, University of Pennsylvania, Philadelphia, PA*

[3]*Leonard Davis Institute, University of Pennsylvania, Philadelphia, PA*

[4]*Corporal Michael J. Crescenz Medical Center, Veterans Affairs, Philadelphia, PA*

[5]*University of Pennsylvania Health System, Philadelphia, PA*

[6]*Department of Emergency Medicine, Hospital of the University of Pennsylvania, Philadelphia, PA*

**Corresponding Author E-mail: basam.alasaly@pennmedicine.upenn.edu*

Abstract: Ambient artificial intelligence (AI) scribes can listen to a conversation between a clinician and patient and automatically generate a clinical document. Early evaluations have largely relied on surveys and electronic health record-derived metrics, which may not fully capture how these tools affect in-room communication, attention, and engagement. We evaluated video-recorded and professionally transcribed primary care encounters from a primary care network in the Philadelphia area. Through structured group consensus, we defined communication and engagement metrics that could be operationalized from multimodal encounter data, including physician egocentric field-of-view categories, speaking-time distribution, silence, device use, visit flow, and expressive utterances, such as laughter. We then assessed these metrics in a pilot sample of encounters with and without documented ambient AI scribe use. The pilot sample included 35 primary care encounters: 14 with ambient AI scribe use and 21 without ambient AI scribe use. The pipeline generated encounter-level measures from available audiovisual recordings and transcripts, including speaking-time distribution, silence, laughter, transcript-labeled phone, tablet, or computer use, visit-phase classifications, and physician egocentric video frame categories. Comparing encounters with versus without documented ambient AI scribe use, PCP speaking time was 53.0% versus 49.8% of analyzed audio time, patient speaking time was similar (34.2% in both groups), silence 12.2% versus 15.2%, PCP laughter 0.44 versus 0.21 laughs/min, PCP phone and computer use 26.3% versus 29.4%, plan/counseling time 39.1% versus 42.8%, and computer-only physician egocentric frames 11% versus 28%. These measures were summarized descriptively by ambient AI scribe use status without inferential testing. Multimodal audiovisual encounter data can support descriptive measurement of patient-physician communication and engagement during ambulatory visits. This proof-of-concept framework may complement surveys and electronic health record-derived metrics in future evaluations of ambient AI scribes; the pilot was not designed to estimate a causal effect of ambient-scribe use.

Keywords: Ambient Artificial Intelligence; AI Scribe; Patient-Physician Communication; Multimodal Data; Audiovisual Analysis; Primary Care.

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

1. Introduction

Ambient artificial intelligence (AI) scribes listen to the conversation between a patient and clinician and generate clinical documentation [1]. Initial evaluations of ambient AI tools have used surveys to assess burnout, cognitive load, and documentation burden - largely finding sustained improvements in these domains- and EHR efficiency metrics to measure clinician documentation time, with mixed findings trending towards mild time savings [2-8]. Reducing documentation demands during patient encounters may alter the style and flow of visits by changing how clinicians divide attention between patients, computers, and other clinical tasks [9-11]. However, prior evaluations of ambient AI, relying on surveys and EHR data, are limited in their ability to objectively assess patient-physician communication and engagement during the encounter.

Better patient-physician communication can improve quality of care [12]. Nearly a third of patient encounters may contain contextual errors – failures to identify or adapt the care plan to a patient's relevant life circumstances – representing missed opportunities for prevention and chronic disease management [13, 14]. Since the implementation of EHRs, clinicians' capacity for undivided attention on their patients has been replaced by multitasking and frequent task-switching, limiting the ability for deep work and sustained focus and raising risk for these errors [15, 16]. Untethering clinicians from documentation and EHR review with AI scribes could improve attention and ability to address contextual factors in real time.

In addition to its impact on the quality of care, patient satisfaction is directly related to patient-clinician communication, through improved attention to patients' social and emotional needs [17]. Eye contact and gaze patterns are a primary channel through which this attentiveness is conveyed and perceived by patients [18, 19]; conversely, physicians who devote more visual attention to the EHR screen during visits exhibit reduced patient-directed gaze and are rated as less attentive, independent of the clinical content delivered [20]. In addition to improving rates of clinical effectiveness through increased medication adherence [21], improved engagement between patients and physicians can have positive consequences for health systems such as reduced malpractice risk for primary care physicians (PCPs) who use humor and laughter [22]. AI scribes have been associated with improvements on standardized patient satisfaction surveys [23], and qualitative studies have described clinician-perceived changes in patient engagement [24], but quantitative evaluations of non-verbal markers of engagement with ambient AI scribe use remain limited.

Assessments linking patient-physician communication and engagement to quality, adherence, and satisfaction often rely on patient encounters obtained through direct observation by independent researchers, recorded visit audio or video with post hoc analysis [25], or standardized patients. While these methods are essential to objectively study engagement and communication, there are also significant practical challenges in data collection. First, patients desire privacy and institutions must comply with federal, state, and local privacy statutes and best practices. Additionally, introducing observers into a busy ambulatory workflow can cause disruptive delays, compromising operational efficiency. There also may be bias from direct observation by a third

party (e.g., the Hawthorne effect) [26]. Lastly, there are technical limitations, especially with large volumes of frame-by-frame video data.

These impediments have contributed to the underuse of recorded visit data in clinical AI evaluation [27, 28]. As a result, published assessments of recent health AI solutions, including ambient AI scribes, have relied largely on post-visit surveys, clinician-reported outcomes, or secondary data from electronic health records rather than direct audiovisual observation. Although these studies have helped shape the conversation about documentation burden and clinician burnout, they leave important questions unanswered about how ambient AI affects patient-clinician communication, patient engagement, visit dynamics, and the downstream health and financial consequences of these changes. Prior studies used video recordings to examine the early impact of EHRs on visit flow and generated practical recommendations for exam room design, including optimal computer placement [27, 29, 30]. Extending this approach to ambient AI scribe research would provide a more complete understanding of how these tools affect the patient-physician relationship and the clinical encounter itself. Prior observational and recorded-visit research has demonstrated the value of direct audiovisual analysis for characterizing patient-physician communication and clinical workflow [25, 27]. Our framework extends this approach by integrating audiovisual, transcript-derived, and survey measures within a common communication framework for ambient AI scribe evaluation.

In this study, we proposed a multimodal framework to quantify patient-physician communication and engagement during primary care encounters. We assessed whether recorded visits could support reproducible measurement of visual-attention proxies, speaking-time distribution, silence, device use, visit flow, rapport, and laughter. We applied the framework to encounters with and without documented ambient AI scribe use to illustrate how multimodal encounter data can complement surveys and EHR-derived metrics. This descriptive pilot was not designed to estimate the causal effect of ambient-scribe use.

2. Methods

2.1 Study setting and data source

This study leveraged the Observer repository, a multimodal dataset of primary care visits collected at Penn Medicine [27]. As part of the study, PCPs and patients wore egocentric (first-person perspective) video cameras during their visit, capturing each party's field of view during the clinical encounter. A third camera provided a view of the room.

All patients and clinicians included in this analysis provided written informed consent and authorization for the collection and research use of encounter audiovisual and associated data. This secondary analysis was conducted by authorized REDUCE study personnel within Penn-approved secure computing environments. Recordings involving family members or caregivers were handled in accordance with the protocol's consent/release and deidentification requirements.

2.2 Patient-physician engagement framework

After reviewing audiovisual data from primary care encounters, as well as existing literature on sociotechnical ethnography [28-31], we organized the audiovisual and survey measures in this study using the pathways framework proposed by Street and colleagues [32]. In this framework, communication processes including nonverbal behavior, affect, and attentional availability link to proximal outcomes (e.g., rapport and trust) and reflect the immediate relational products of communication. Proximal outcomes link to intermediate outcomes (e.g., trust in system and commitment to treatment) and to health outcomes (survival, functional status).

Our audiovisual measures operationalize communication and proximal outcomes across verbal, nonverbal, and attentional channels, while our post-visit surveys capture intermediate outcomes. This structure allows encounter-level audiovisual differences to be interpreted as candidate pathways through which ambient-scribe use may relate to intermediate outcomes, including patient satisfaction and clinician experience. It does not establish that ambient-scribe use caused those outcomes. Table 1 displays the metrics in the patient-physician engagement framework and maps each metric to its corresponding Street pathway category and communication channel.

To operationalize the parameters in Table 1, we constructed a software pipeline to process audio and video data from each encounter in the Observer repository. Each pipeline component maps to one or more Street pathway categories in Table 1: transcript-derived speaking time and visit-phase segmentation to information exchange and to making decisions and enabling self-management; laughter and affect labels to fostering the healing relationship and to responding to emotions; silence to managing uncertainty; and physician egocentric field-of-view categories and device use to nonverbal immediacy and attentional availability. Figure 1 summarizes the multimodal processing framework.

2.3 Transcript-derived measures and visit-phase classification

Encounter recordings were professionally reviewed and transcribed by Datagain Inc. (Parkland, FL, USA), a professional transcription vendor with authorized access to the audiovisual recordings. Transcript files included timestamps, speaker-role identification, affect labels including laughter, and manually coded indicators of concurrent phone, tablet, or computer use. For each encounter, we calculated speaking time for the patient, PCP, and other participants, along with total silence time. Speaking-time and silence percentages were calculated for each encounter as a proportion of total analyzed audio time, defined as the sum of PCP, patient, and other speaking time and silence, so that the four categories sum to 100% within each encounter. Total analyzed audio time was used as the denominator rather than recording length because utterances sharing a start timestamp with the following utterance were assigned a minimum duration of 0.5 seconds, which caused summed speaking and silence time to slightly exceed recording length. Device-use percentages and laughter rates used total encounter duration as the denominator because these measures can overlap with speaking time.

Table 1. Patient-physician engagement framework mapped to the Street et al. (2009) communication pathways, operationalized measures, and multimodal data sources.

Framework element			***Data source***							
Street pathway/category	**Communication parameter**	**Operationalized metric**	**Physician egocentric video**	**Patient egocentric video**	**Room video**	**Visit audio**	**Transcript**	**Encounter Note***	**EHR Audit log**	**Survey**
Information exchange	Visit flow; conversational balance	Visit phase segmentation; PCP/patient/other speaking-time distribution			○	●	●		○	
Fostering the healing relationship	Rapport; laughter/joy	Rapport-building utterances; PCP and patient laughter rate	○	○		○	●			
Responding to emotions	Affect; silence; laughter	Affect labels; laughter events; silence distribution	○	○		●	●			
Making decisions and enabling self-management	Patient participation	Plan/counseling time; patient speaking proportion				●	●	○		
Managing uncertainty	Visit pacing; deliberation/task-switching	Plan/counseling time; silence time				●	●		○	
Nonverbal immediacy and attentional availability	Physician field-of-view categories (visual-attention proxy); patient attention	Person only, computer only, both, neither; PCP/patient device use	●	○	○		●		○	
Proximal interaction outcomes	Patient and clinician-reported experience	Patient satisfaction, clarity, listening; PCP visit goals and patient engagement	○	○						●

● Operationalized in this pilot.

○ Available in the Observer repository and relevant to the parameter but not used for it in this pilot (e.g., physician egocentric video was used for field-of-view categories but not to detect laughter or affect). Blank cells indicate modality-parameter pairings not identified as primary measurement sources.

***** The signed encounter note was used in this pilot to ascertain ambient AI scribe exposure status via the standardized ambient-technology footnote.

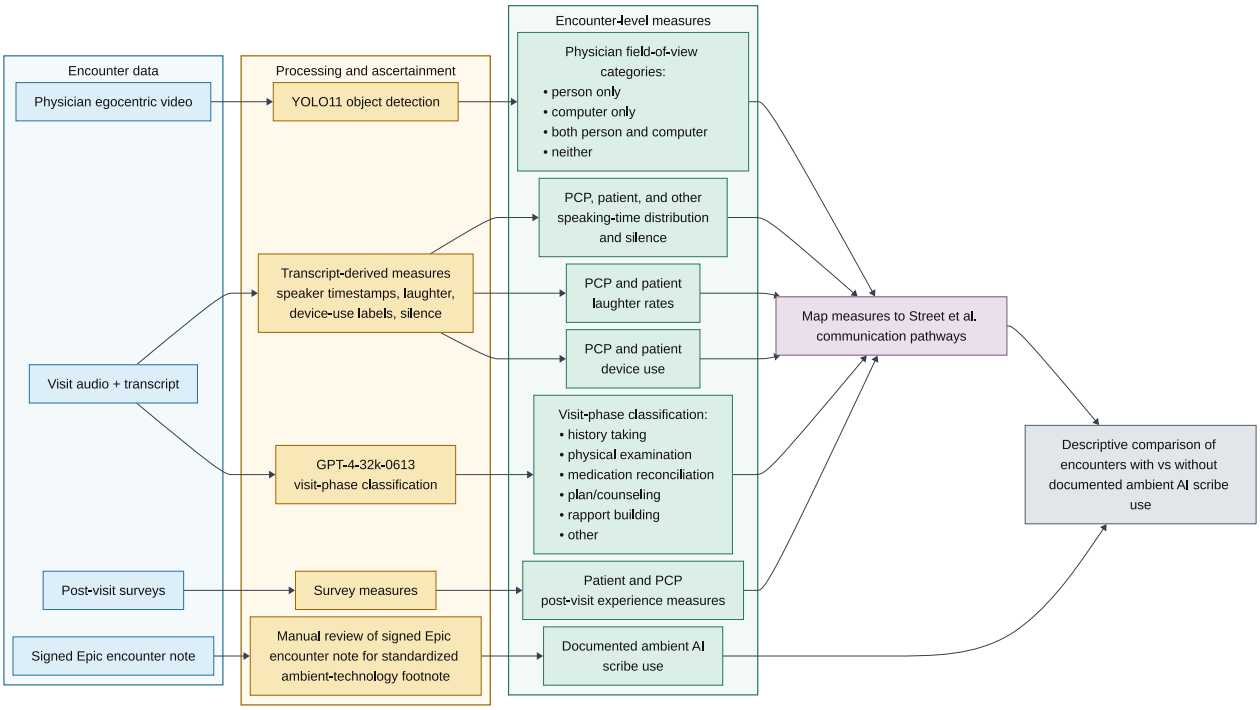


Figure. 1. Multimodal processing framework for encounter-level measurement of patient-physician communication and engagement.

2.3.1 Visit Segmentation

Transcript utterances were classified into visit phases using OpenAI GPT-4-32k-0613 through the Penn Databricks environment in April 2025. Only deidentified transcript content was submitted for language-model processing. Categories included history taking, physical examination, medication reconciliation, plan/counseling, rapport building, and other. Transcripts were segmented into chunks of approximately 50 consecutive utterances for classification, preserving local conversational context within each segment. Visit-phase duration was calculated from the time attributed to each classified phase and divided by total encounter duration; because each point in the encounter timeline was attributed to one of the six visit phases, the phase percentages sum to 100%. Model parameters included temperature = 0 and top_p = 0.95. Two clinician reviewers independently labeled 854 utterances from three transcripts to assess concordance with automated classification (Appendix A7).

2.4 Physician egocentric video analysis

Encounters were recorded using a room camera and first-person perspective cameras worn by the PCP and patient. An automated, publicly available object-detection algorithm (Ultralytics YOLO11 [33]) was applied to images from the PCP camera within Penn-approved secure computing infrastructure. Detection focused on person and computer object classes, with computer represented by laptop and monitor-related model categories. Frames were classified into four mutually exclusive field-of-view categories: person only, computer only, both person and

computer, or neither. Frame counts were divided by the total frames to estimate the proportion of each encounter represented by each category. These egocentric field-of-view categories were treated as proxies for visual attention rather than direct eye-tracking measures. The pretrained object-detection model was used without task-specific retraining; formal frame-level validation against manually annotated physician egocentric video was not performed in this proof-of-concept analysis. Appendix A provides additional details (Supplementary Material, https://doi.org/10.5281/zenodo.23048887).

2.5 Post-visit surveys and documented ambient AI scribe use

Patients completed surveys asking about their satisfaction with the visit, and whether their PCP communicated clearly and listened to them. PCPs completed surveys assessing their satisfaction with achieving visit goals and with patient engagement/comprehension of the medical concepts discussed. The full surveys are available from the corresponding author on request.

Encounter-level ambient AI scribe use was manually ascertained by BA through review of the signed Epic encounter note for the standardized footnote documenting use of ambient technology. Encounters containing the footnote were categorized as having documented ambient-scribe use; encounters without the footnote were categorized as not having documented use. Three PCPs at a single academic health system contributed encounters both with and without documented ambient-scribe use, and the sample was limited to these PCPs. Ambient-scribe access and use occurred through routine institutional practice and were not assigned by the study team. Encounters took place between April 9, 2024, and February 13, 2025. This design supported a descriptive comparison of encounters with and without documented use but not a randomized or causal analysis.

2.6 Statistical analysis

For continuous audio-, video-, transcript-, and survey-derived variables, encounter-level means and standard deviations were calculated by documented ambient AI scribe use status. The analysis was descriptive; no inferential statistical testing or causal estimation was performed.

3. Results

The pilot sample included 35 unique patient-PCP dyads, including 14 encounters with ambient AI scribe use and 21 without ambient AI scribe use. These encounters were conducted by 3 PCPs who each had visits with and without ambient AI scribe use in the Observer repository. For two encounters without ambient AI scribe use, audio-only data were available due to technical challenges with video recording. The average age of patients in this sample was 54 years and patients had a range of responses across self-reported health status, race/ethnicity and educational backgrounds (Supplementary Table 1).

The pipeline successfully generated encounter-level measures from the available audiovisual recordings, supporting the feasibility of applying automated processing methods to multimodal

primary care encounter data. Table 2 presents encounter-level measures organized by patient-physician engagement framework element. The PCP speaking time was 53.0% of analyzed audio time in visits with ambient AI scribe use and 49.8% in visits without ambient AI scribe use (Table 2). The amount of silence time was 12.2% with ambient AI scribe use and 15.2% in visits without ambient AI scribe use. Patient speaking time was 34.2% both with and without ambient AI scribe use.

PCP laughter rate was 0.44 laughs per minute with ambient AI scribe use compared to 0.21 without ambient AI scribe use. Patient laughter rate was 0.15 laughs per minute in encounters with ambient AI scribe use and 0.19 without ambient AI scribe use. Patient cell phone usage was 20.3 seconds (1.8%) in encounters without ambient AI scribe use compared to 0.4 seconds (<0.1%) in encounters with ambient AI scribe use. PCP computer usage was 352 seconds (29.1%) in encounters without ambient AI scribe use compared to 328 seconds (25.0%) in encounters with ambient AI scribe use.

For visit-phase classification, agreement between the two clinician reviewers was 78.0% (Cohen's $\kappa = 0.70$). Restricted to the 665 utterances on which both clinician reviewers agreed, LLM-assigned labels matched the reviewer consensus in 65.7% of utterances ($\kappa = 0.53$) (Appendix A7). Time attributed to history taking was 27.2% with ambient AI scribe use compared to 28.2% without ambient AI scribe use (Table 2). Plan and counseling was 39.1% with ambient AI scribe use vs. 42.8% without ambient AI scribe use. Time spent building rapport was 18.6% with ambient AI scribe use vs. 16.2% without ambient AI scribe use. Time spent on physical examination was 6.5% with ambient AI scribe use vs. 3.6% without ambient AI scribe use.

Computer-only frames constituted 28% of physician egocentric video in encounters without documented ambient AI scribe use and 11% in encounters with documented use (Table 2). Person-only frames constituted 38% with documented ambient AI scribe use and 28% without documented use. Frames containing both a person and computer constituted 44% with documented use and 40% without documented use. These categories represent automated estimates of the physician camera's field of view and should not be interpreted as direct eye-tracking measures.

Patient satisfaction was rated at 9.4 (SD: 0.7) in encounters with ambient AI scribe use, compared to 8.9 (SD: 1.5) in encounters without ambient AI scribe use. PCP satisfaction with achieving visit goals was rated at 8.5 (SD: 0.7) in encounters with ambient AI scribe use compared to 8.2 (SD: 1.1) in encounters without ambient AI scribe use.

Table 2. Encounter-level communication and engagement measures in primary care encounters with and without ambient AI scribe use, mapped to patient-physician engagement framework elements.

Operationalized metric	Street pathway/category							With ambient AI scribe use, mean (SD)	Without ambient AI scribe use, mean (SD)
	Information exchange	Fostering the healing relationship	Responding to emotions	Making decisions and enabling self-management	Managing uncertainty	Nonverbal immediacy and attentional availability	Proximal interaction outcomes		
PCP speaking time (% of analyzed audio)	●							53.0 (9.5)	49.8 (10.3)
Patient speaking time (% of analyzed audio)	●			●				34.2 (8.1)	34.2 (7.7)
Other speaking time (% of analyzed audio)	●							<1 (1.3)	<1 (2.5)
Silence time (% of analyzed audio)			●		●			12.2 (2.3)	15.2 (6.7)
History taking (% of encounter) *	●							27.2	28.2
Plan/counseling (% of encounter) *				●	●			39.1	42.8
Rapport (% of encounter) *		●						18.6	16.2
Medication reconciliation (% of encounter)*	●							8.4	7.4
Physical examination (% of encounter)*	●							6.5	3.6
Other visit phase (% of encounter) *	●							0.2	1.8
PCP laughter rate (laughs/min)		●	●					0.44 (0.33)	0.21 (0.21)
Patient laughter rate (laughs/min)		●	●					0.15 (0.13)	0.19 (0.29)
PCP cell phone use (% of encounter)						●		1.3 (1.6)	0.3 (0.8)
PCP computer use (% of encounter)						●		25.0 (11.7)	29.1 (12.4)
Patient cell phone use (% of encounter)						●		<0.1 (0.1)	1.8 (3.1)
Patient computer use (% of encounter)						●		0.5 (1.3)	0.1 (0.2)
Person-only frames (% of physician egocentric video)						●		38 (24.7)	28 (12.3)
Computer-only frames (% of physician egocentric video)						●		11 (15.9)	28 (19.0)
Both person and computer frames (% of physician egocentric video)						●		44 (31.0)	40 (18.6)
Neither person nor computer in frame (% of physician egocentric video)						●		7 (16.1)	3 (3.8)
Patient satisfaction (1-10 scale)							●	9.4 (0.7)	8.9 (1.5)
PCP satisfaction with achieving visit goals (1-10 scale)							●	8.5 (0.7)	8.2 (1.1)

● indicates the framework element(s) in Table 1 to which each metric maps.

* Mean % of encounter time, reported without SD; the six visit phases sum to 100%. Silence was calculated separately and may overlap visit phases.

Note: Video measures reflect 33 encounters (14 with, 19 without; two audio-only); all others reflect 35. Speaking and silence are % of analyzed audio and sum to 100%.

4. Discussion

This study developed and piloted a multimodal framework for quantifying patient-physician communication and engagement during primary care encounters. In this proof-of-concept analysis, an automated audiovisual processing pipeline generated encounter-level measures of egocentric field of view, speaking-time distribution, silence, device use, visit segmentation, rapport, and laughter, each mapped to a Street pathway category (Table 1 and Table 2). The findings support the technical feasibility of extracting these communication parameters from multimodal encounter data; they do not establish a causal effect of ambient-scribe use.

This proof-of-concept analysis complements prior survey-based and qualitative studies suggesting that ambient AI scribes may reduce cognitive burden and improve perceived patient engagement [4, 24]. In this pilot sample, the framework summarized computer-only frame time, person-only frame time, speaking-time distribution, silence, device use, visit-phase segmentation, and laughter rate across encounters with and without ambient AI scribe use, demonstrating how multimodal audiovisual data can characterize field-of-view-based proxies for visual attention and clinician affect across encounter types. The ability to build strong patient relationships is a commonly cited reason for entering medicine, particularly in ambulatory specialties [34], yet since the advent of the EHR era, clinicians spend less time with their patients and more effort multitasking on a computer screen.

Clinician burnout is a public health crisis with many root causes including EHR design [35], documentation burden [36], understaffing, increasing patient workload and complexity, and decreasing reimbursement [9, 37, 38]. While much has been written about drivers of burnout, there have been few sustained interventions that provide meaningful burnout reduction. In randomized studies, ambient AI scribes have shown notable improvements in burnout across multiple institutions [4, 39, 40]. However, these burnout improvements have not always been accompanied by reductions in EHR efficiency metrics, suggesting that ambient AI may reduce burnout through mechanisms beyond decreased EHR engagement alone [5]. It is important to understand the potential mechanisms of this reduction both to assist in the iterative design of ambient AI scribes and to facilitate other simultaneous burnout interventions. Objective measures of in-room communication may help clarify whether changes in documentation burden are accompanied by changes in patient-facing attention, silence, or visit flow.

Video data can help characterize physician field-of-view context and candidate proxies for visual attention during clinical encounters. Reviewing video of primary care visits from patient and physician perspectives by the study team shows many aspects of care delivery that are never captured through traditional evaluation – the patient watching their doctor type away on their computer who then pulls out their phone to occupy themselves. As more healthcare physical spaces are equipped with video cameras, these data can be used to improve patient and clinician experience with careful balance of clinician and patient agency, privacy, and data security [41, 42].

Audio data can similarly help quantify encounter dynamics. In this pilot sample, the framework summarized silence time and patient speaking time across encounters with and without documented ambient AI scribe use. Silence may occur during typing, EHR interaction, reflection, physical examination, or other activities; this pilot did not directly attribute each silent interval to a specific activity. Future work should assess, at larger scale and with aligned behavioral or EHR data, how silence and speaking-time distribution relate to care quality and patient experience.

The impact of ambient AI on primary care encounter dynamics is unknown and likely evolving [1]. EHRs changed the flow of a typical primary care visit by allowing for pre-visit longitudinal chart review and concurrent history-taking and order entry [43]. Ambient AI scribes may lead to a similar shift in clinician workflows and present another opportunity to shape those workflows in service of the patient. Although our work is preliminary, the framework summarized time attributed to building rapport and physical examination, elements of the visit requiring human interaction that may improve the quality of the physician-patient relationship [44]. As AI tools become increasingly integrated into primary care workflows including triage, history-taking, and post-encounter summarization, the balance between care delivery activities performed by human clinicians may continue to shift [45, 46]. PCP time is likely most valuable in uniquely human activities including those requiring physical contact and discussions with nuanced risks and benefits or requiring elicitation of patient values.

These pilot results have several important limitations. The analysis included only 35 encounters from three PCPs at one academic health system, and clinicians received ambient-scribe licenses through routine institutional practice rather than random assignment. The two encounter groups differed in patient and visit characteristics, and no inferential statistical testing was performed; observed differences should therefore be treated as descriptive and not causal. Ambient-scribe status was derived from a standardized signed-note footnote, so absent or inconsistent documentation could produce exposure misclassification. The same pilot sample contributed to development and assessment of parts of the analytical workflow and therefore does not constitute an independent validation cohort. Automated audio, transcript, and video processing may introduce measurement error. Language-model concordance was assessed in only three transcripts; agreement with the individual clinician labels was moderate ($\kappa = 0.42$ and 0.48), and agreement with consensus labels on reviewer-agreed utterances was $\kappa = 0.53$, with the weakest concordance in the "other" category; visit-phase percentages should therefore be interpreted as approximate. Formal frame-level validation of the YOLO classifications against a manually annotated reference set was beyond the scope of this pilot and should be performed before these measures are used for definitive comparative or causal evaluation. Two encounters also lacked video data.

A last important caveat to this work is that ambient AI technology is still early in development. As ambient AI scribes shift into a clinical intelligence layer that can assist with orders, diagnosis entry, coding, and decision support, more time may shift away from physicians interacting with their computer towards patient-facing activities. Because the encounters took place between April 9, 2024, and February 13, 2025, at an early-adopting institution, the findings should be interpreted in the context of rapidly evolving ambient AI technology.

Future work should incorporate EHR audit-log data to compare audiovisual measures of patient and computer engagement with keystroke-level evidence of concurrent EHR use. Additional analyses should also consider pre- and post-encounter activities and the roles of nurses, medical assistants, family members, and other participants in the primary-care team. Applying the framework to larger and more diverse encounter samples will be necessary to assess reliability, refine measurement, and evaluate how well observations generalize across clinical settings and workflows.

5. Conclusion

This study developed and piloted a multimodal framework for the sociotechnical analysis of patient-physician communication and engagement during ambient AI scribe evaluations, mapping each measure to a Street et al. communication pathway so that encounter-level observations can be interpreted as candidate mechanisms rather than isolated metrics. In a pilot of 35 primary care encounters, an automated pipeline generated measures of physician egocentric field of view, speaking-time distribution, silence, device use, visit flow, rapport, and laughter, supporting the feasibility of reproducible, encounter-level measurement of in-room interaction from recorded visits. These findings are descriptive and establish technical feasibility rather than a causal effect of ambient-scribe use. Applied to larger and more diverse samples alongside surveys and EHR audit-log data, this approach can help clarify how ambient AI scribes and other clinical technologies reshape the encounter itself.

Ethics Approval and Consent to Participate

The University of Pennsylvania Institutional Review Board approved the REDUCE study (Protocol 853189); this secondary analysis used deidentified data collected with written informed consent under that protocol.

Acknowledgements

This work was supported by the National Library of Medicine and the NIH Office of the Director under project number 5DP1LM014558-03 (Former Number: 1DP1OD035237-01) for the REDUCE project, entitled “Helping Doctors Doctor: Using AI to Automate Documentation and ‘De-Autonomate’ Health Care.” The funders had no role in study design, data collection, analysis and interpretation of data, or writing of the manuscript.

Competing Interests and Disclosures

The views expressed are those of the authors and do not necessarily reflect the position or policy of the VA or the US government.

All authors declare no financial or non-financial competing interests.

Author Contributions:

BA: software, coding, analysis, methodology, writing – original draft

MH: software, coding, analysis, methodology, writing – original draft

SM: software, coding

KBJ: conceptualization, methodology, writing – review and editing, funding acquisition

AS: resources

SS: resources

SA: resources

MJD: methodology, review of transcripts for LLM concurrence, writing – review and editing

KPS: conceptualization, methodology, review of transcripts for LLM concurrence, validation, writing – original draft, writing – review and editing

All authors read and approved the final manuscript.

Supplementary material: Appendices A–B and supplementary tables and figures are available at https://doi.org/10.5281/zenodo.23048887.

Data availability:

The data analyzed during the current study are part of the Observer repository. Access is restricted because the dataset contains audiovisual recordings, transcripts, visit metadata, survey responses, and related clinical encounter data. Data requests can be submitted through the Observer repository at https://observer.med.upenn.edu/dataset. Copies of the patient and PCP satisfaction surveys are available from the corresponding author on request.

Code availability:

Code for video and audio analysis can be accessed at the following GitHub repository: https://github.com/kbjohnson-penn/ambient-audiovisual-analysis

References

1. Shah, K.P. and K.B. Johnson, *The Ambient AI Scribe Revolution—Early Gains and Open Questions.* JAMA Network Open, 2025. **8**(10): p. e2534982.
2. Liu, T.-L., et al., *Does AI-Powered Clinical Documentation Enhance Clinician Efficiency? A Longitudinal Study.* NEJM AI, 2024. **1**(12): p. AIoa2400659.
3. Tierney, A.A., et al., *Ambient artificial intelligence scribes to alleviate the burden of clinical documentation.* NEJM Catalyst Innovations in Care Delivery, 2024. **5**(3): p. CAT. 23.0404.
4. Olson, K.D., et al., *Use of ambient AI scribes to reduce administrative burden and professional burnout.* JAMA Network Open, 2025. **8**(10): p. e2534976.
5. Afshar, M., et al., *A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being.* NEJM AI, 2025. **2**(12): p. AIoa2500945.
6. Duggan, M.J., et al., *Clinician experiences with ambient scribe technology to assist with documentation burden and efficiency.* JAMA Network Open, 2025. **8**(2): p. e2460637–e2460637.
7. You, J.G., et al., *Ambient documentation technology in clinician experience of documentation burden and burnout.* JAMA network open, 2025. **8**(8): p. e2528056.
8. Rotenstein, L.S., et al., *Changes in Clinician Time Expenditure and Visit Quantity With Adoption of Artificial Intelligence–Powered Scribes: A Multisite Study.* JAMA, 2026.
9. Abraham, C.M., K. Zheng, and L. Poghosyan, *Predictors and outcomes of burnout among primary care providers in the United States: a systematic review.* Medical Care Research and Review, 2020. **77**(5): p. 387–401.
10. Holmgren, A.J., et al., *Electronic health record usability, satisfaction, and burnout for family physicians.* JAMA Network Open, 2024. **7**(8): p. e2426956–e2426956.
11. Owens, L.M., et al., *Effect of ambient voice technology, natural language processing, and artificial intelligence on the patient–physician relationship.* Applied Clinical Informatics, 2024. **15**(04): p. 660–667.
12. Stewart, M.A., *Effective physician-patient communication and health outcomes: a review.* CMAJ: Canadian medical association journal, 1995. **152**(9): p. 1423.
13. Weiner, S.J. and A. Schwartz, *Contextual errors in medical decision making: overlooked and understudied.* Academic Medicine, 2016. **91**(5): p. 657–662.
14. Weiner, S.J., et al., *Contextual errors and failures in individualizing patient care: a multicenter study.* Annals of Internal Medicine, 2010. **153**(2): p. 69–75.
15. Kissler, M.J., K. Kissler, and M. Burden, *Toward a medical "ecology of attention".* New England journal of medicine, 2021. **384**(4): p. 299–301.
16. Chen, Y., J. Adler-Milstein, and C.A. Sinsky, *Measuring and maximizing undivided attention in the context of electronic health records.* Applied Clinical Informatics, 2022. **13**(04): p. 774–777.
17. DiMatteo, M.R., *A social-psychological analysis of physician-patient rapport: toward a science of the art of medicine.* Journal of social issues, 1979. **35**(1): p. 12–33.
18. Gorawara-Bhat, R. and M.A. Cook, *Eye contact in patient-centered communication.* Patient Educ Couns, 2011. **82**(3): p. 442–7.
19. Haase, R.F. and D.T. Tepper, *Nonverbal components of empathic communication.* Journal of counseling psychology, 1972. **19**(5): p. 417.
20. Asan, O., et al., *How physician electronic health record screen sharing affects patient and doctor non-verbal communication in primary care.* Patient Educ Couns, 2015. **98**(3): p. 310–6.
21. Bartlett, E.E., et al., *The effects of physician communications skills on patient satisfaction; recall, and adherence.* Journal of chronic diseases, 1984. **37**(9-10): p. 755–764.

22. Levinson, W., et al., *Physician-patient communication: the relationship with malpractice claims among primary care physicians and surgeons.* Jama, 1997. **277**(7): p. 553–559.
23. Davis, E., et al., *Ambient AI Documentation and Patient Satisfaction in Outpatient Care: Retrospective Pilot Study.* JMIR AI, 2026. **5**: p. e78830.
24. Shah, S.J., et al., *Physician perspectives on ambient AI scribes.* JAMA Network Open, 2025. **8**(3): p. e251904.
25. Coleman, T., *Using video-recorded consultations for research in primary care: advantages and limitations.* Family practice, 2000. **17**(5): p. 422–427.
26. Pringle, M. and C. Stewart-Evans, *Does awareness of being video recorded affect doctors' consultation behaviour?* The British Journal of General Practice, 1990. **40**(340): p. 455.
27. Johnson, K.B., et al., *Observer: creation of a novel multimodal dataset for outpatient care research.* Journal of the American Medical Informatics Association, 2026. **33**(2): p. 424–433.
28. Fitzpatrick, G. and G. Ellingsen, *A review of 25 years of CSCW research in healthcare: contributions, challenges and future agendas.* Computer Supported Cooperative Work (CSCW), 2013. **22**(4): p. 609–665.
29. Ackerman, M., et al., *Designing healthcare that works: a sociotechnical approach*. 2017: Academic Press.
30. Blomberg, J. and H. Karasti, *Reflections on 25 years of ethnography in CSCW.* Computer supported cooperative work (CSCW), 2013. **22**(4): p. 373–423.
31. Savage, J., *Ethnographic evidence: The value of applied ethnography in healthcare.* Journal of research in Nursing, 2006. **11**(5): p. 383–393.
32. Street Jr, R.L., et al., *How does communication heal? Pathways linking clinician–patient communication to health outcomes.* Patient education and counseling, 2009. **74**(3): p. 295–301.
33. Jocher, G. and J. Qiu, *Ultralytics YOLO11*. 2024.
34. Clark, W., et al., *Improving physicians' relationships with patients.* Journal of general internal medicine, 1999. **14**(Suppl 1): p. S45.
35. Asgari, E., et al., *Impact of electronic health record use on cognitive load and burnout among clinicians: narrative review.* JMIR Medical Informatics, 2024. **12**: p. e55499.
36. Gardner, R.L., et al., *Physician stress and burnout: the impact of health information technology.* Journal of the American Medical Informatics Association, 2019. **26**(2): p. 106–114.
37. Guille, C. and S. Sen, *Burnout, depression, and diminished well-being among physicians.* New England Journal of Medicine, 2024. **391**(16): p. 1519–1527.
38. Amoafo, E., et al., *What are the significant factors associated with burnout in doctors?* Occupational medicine, 2015. **65**(2): p. 117–121.
39. Shah, S.J., et al., *Ambient artificial intelligence scribes: physician burnout and perspectives on usability and documentation burden.* Journal of the American Medical Informatics Association, 2025. **32**(2): p. 375–380.
40. Misurac, J., L.A. Knake, and J.M. Blum, *The effect of ambient artificial intelligence notes on provider burnout.* Applied clinical informatics, 2025. **16**(02): p. 252–258.
41. Simma, B., et al., *Video recording in the delivery room: current status, implications and implementation.* Pediatric Research, 2024. **96**(3): p. 610–615.
42. Cheikh Youssef, S., et al., *Evolution of the digital operating room: the place of video technology in surgery.* Langenbeck's archives of surgery, 2023. **408**(1): p. 95.
43. Zheng, K., R.M. Ratwani, and J. Adler-Milstein, *Studying workflow and workarounds in electronic health record–supported work to improve health system performance.* Annals of internal medicine, 2020. **172**(11_Supplement): p. S116–S122.

44. Duan, L., E.M. Mukherjee, and D.G. Federman, *The physical examination: a survey of patient preferences and expectations during primary care visits.* Postgraduate medicine, 2020. **132**(1): p. 102–108.
45. McMahon, B. and D. McInerney, *Right care, right place, first time: how AI is improving national virtual front doors*. 2025, Massachusetts Medical Society. p. AIpc2401260.
46. Allen, L.N., et al., *Artificial intelligence in primary care: frameworks, challenges, and guardrails.* The Lancet Primary Care, 2025.