

Sociotechnical Ethnography in Health Informatics Bridging Technology, Practice, and Human Experience

Kevin B. Johnson, MD, MS

David L. Cohen University Professor of Pediatrics, Informatics, Engineering and Communication
Perelman School of Medicine, School of Engineering and Applied Science, Annenberg School
of Communication

University of Pennsylvania, Philadelphia, PA, USA

Email: kevin.johnson1@pennmedicine.upenn.edu

Andrea L. Hartzler, PhD

Department of Biomedical Informatics and Medical Education, School of Medicine
University of Washington, Seattle, WA, USA

Email: andrea@uw.edu

Karen O'Connor, MS

Associate Director

Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine,
University of Pennsylvania, Philadelphia, PA, USA

Email: karoc@pennmedicine.upenn.edu

Artificial intelligence (AI) and digital health technologies are now integrated into clinical care at scale, from ambient AI scribes that listen to the visit to secure messaging and patient portals that move clinical conversations outside the exam room. Whether these tools succeed depends largely on how they fit the social and organizational settings where care happens. This session brings together work that combines computational methods with ethnographic and human-centered approaches to study how technologies reshape clinical workflow, patient–clinician communication, care delivery, and evaluation practices that conventional metrics miss. Across three themes, the studies draw on diverse multimodal data, including encounter recordings, clinical notes, EHR audit logs, messaging, imaging, stakeholder interviews, and text from many sources (e.g., social media, biomedical literature) to: (1) observe clinical encounters through multimodal data, (2) characterize communication and coordination beyond the exam room, and (3) question whose perspectives and forms of performance count. The session concludes by introducing the Observer Consortium, a federated effort to support shared standards, governance, and collaboration around multimodal clinical data for sociotechnical AI research.

Keywords: sociotechnical systems, ethnography, multimodal data, ambient AI, clinical communication, digital traces, value sensitive design

1. Introduction

Health information technology has long been understood as part of sociotechnical systems, not a set of standalone tools. Its effects arise from the interactions among software, people, workflows, organizational policies, and the physical environment (Berg, 1999; Sittig and Singh, 2010). Ethnographic and qualitative studies have repeatedly shown that technologies that perform well in isolation can have unintended

consequences when used in practice. Examples include fragmenting clinical communication, shifting work between roles, and changing what gets recorded and what gets lost (Ash, Berg and Coiera, 2004). These insights have mostly come from fieldwork that is detailed and rich in context but hard to scale.

Two developments make sociotechnical ethnography more pressing and, at the same time, open new avenues for addressing it. First, AI is moving into the center of the clinical encounter. Ambient AI scribes using large-language models (LLMs) now listen to patient–clinician conversations and generate visit summaries (Tierney et al., 2024). Second, digital infrastructure now produces fine-grained “breadcrumbs” detailing clinical activities. EHR audit logs, secure messaging metadata, and multimodal recordings of encounters can give passive, scalable views of the activities that ethnographers previously needed to observe directly (Rule, Chiang and Hribar, 2020; Johnson et al., 2026). Together, these developments call for methods that pair computational scale with the contextual sensitivity of ethnography, so that we can see how technology changes care, assess who benefits or is left out, and propose new or different innovations to further improve clinical interactions and outcomes.

This session at the 2027 Pacific Symposium on Biocomputing (PSB), titled “Sociotechnical Ethnography in Health Informatics: Bridging Technology, Practice, and Human Experience,” brings together research at the intersection of computational and ethnographic approaches. The accepted papers fall into three themes: (1) observing the clinical encounter through multimodal data, (2) communication and coordination beyond the exam room, and (3) whose voice and whose performance counts in the data and metrics that shape clinical AI. The session also describes the Observer Consortium, an initiative to build a federated network of scientists and multimodal clinical datasets from partner institutions across the United States and possibly worldwide. Our goal is to encourage dialogue among computer scientists, informaticians, social scientists, and clinicians toward technologies that are technically robust, fairly implemented, and grounded in the realities of clinical practice.

2. Sociotechnical Ethnography in Health Informatics

2.1. *Observing the Clinical Encounter Through Multimodal Data*

The clinical visit is where many of the most consequential sociotechnical changes are now taking place. Two papers in this session use the Observer repository, a dataset of recorded primary care encounters, to study how technology shapes what happens in the room and to demonstrate the multimodal insights made possible by richly captured visit data (Johnson et al., 2026).

Alasaly et al. (K. Johnson, co-author with COI*) apply multimodal observation to ambient AI scribes. Their framework measures patient–physician communication and engagement using physician- and patient-worn first-person (egocentric) video, room audio, timestamped transcripts, and post-visit surveys. Each measure is mapped to the pathways Street et al. (2009) proposed for how communication affects health outcomes. The pipeline uses object detection to classify each frame of the physician’s video by gaze target (person, computer, both, or neither). It also uses LLM-assisted labeling of visit phases and transcript-derived measures of device use, talk time, silence, and laughter. The authors applied the framework to 35 primary care encounters: 14 with documented ambient scribe use and 21 without. Comparing encounters with and without the scribe, physicians gazing at the computer made up 11% versus 28% of physician video frames, the amount of silence time was 12.8% versus 15.8%, and the physician laughter rate was 0.44 versus 0.21 laughs per minute. The LLM’s phase labels matched clinician consensus on 65.7% of utterances (Cohen’s kappa $\kappa = 0.53$). While the results are descriptive and do not establish causality, they show that multimodal data from recorded visits can capture shifts in attention and rapport that surveys and EHR metrics miss (Alasaly et al., 2027).

Raj et al. (A. Hartzler co-author with COI*) ask a related question about documentation: do the feelings expressed in a clinical conversation carry into the clinical note? Using 89 outpatient visits from the

* COI papers reviews were coordinated, and inclusion decision were made, by PSB organizers and not session chairs

Observer repository, they apply TRACE, a sentiment method that scores text on five dimensions of how people perceive one another (sociability, morality, agency, ability, and status), each with a positive and negative valence. They score provider speech, patient speech, and the matching note separately. Note sentiment tracked provider speech on seven of the ten dimensions (Spearman correlation coefficient $r_s \approx 0.25\text{--}0.39$, Benjamini–Hochberg adjusted $p < 0.05$), most strongly for morality, agency, and ability. However, note sentiment was not significantly associated with patient speech on any dimension. Sentiment language was also used more broadly in conversation than in documentation: 17.4% of lexicon terms appeared in transcripts versus 11.6% in notes. These findings suggest the note is a provider-framed record that largely omits the patient’s emotional voice, a gap that becomes more pronounced as ambient AI scribes take over documentation (Raj et al., 2027).

Together, these papers show that multimodal encounter data can turn ethnographic concepts such as attentional availability, rapport, and the translation from conversation to documentation into measures that are reproducible and grounded in theory. The same data sources that make ambient documentation possible can also support systematic evaluation of how these systems reshape clinical attention, communication, and representation in the medical record.

2.2. Communication and Coordination Beyond the Exam Room

A growing share of clinical communication happens asynchronously among clinicians via secure messaging and through patient portal messaging between patients and clinicians. Two papers examine this less visible layer of care from two directions: large-scale analysis of digital traces and in-depth elicitation of clinicians’ values.

Osweiler et al. move from individual clinicians’ EHR use to care team coordination. They combine EHR audit logs, secure chat messaging metadata, and other clinical data (ADT events, patient demographics, billing codes) to characterize team communication across 6,737 inpatient hospitalizations at 12 hospitals. Using relational event models fit to each patient’s care team, they find that team messaging is strongly reciprocal. A clinician was 171 times more likely to reply to someone who had recently messaged them (median hazard 171.84), while repeat messages to the same recipient were 66% less likely. A one-standard-deviation increase in team reciprocity was associated with a 3% shorter length of stay. “Turn-receiving” patterns, in which a message recipient immediately messages a third person, were associated with a 4% longer stay. Repeated messaging was associated with about 1% more team EHR time per day. These findings show that fine-grained digital traces can reveal coordination behaviors linked to patient throughput and clinician workload that cannot otherwise be observed at scale (Osweiler et al., 2027).

Slater et al. study a high-stakes form of asynchronous communication: patient portal messages that disclose suicidal thoughts or behaviors. Using value-sensitive design (VSD), they interviewed 16 healthcare professionals, most of whom were nurses, from five clinics at an academic medical center. They analyzed the interviews with open coding to identify barriers and facilitators in the current system, and with coding guided by VSD and the Systems Engineering Initiative for Patient Safety (SEIPS) 2.0 human factors framework (Holden et al., 2013). Participants described a complex sociotechnical system spanning nurses, call-center staff, family members, mobile crisis teams, and police. Messages of concern were most often “cryptic” (vague hopelessness) rather than explicit, so assessment depended on follow-up to establish additional context for response. Patients also had no way to mark a message as urgent, so they used subjects such as “depression” instead. The values participants endorsed most were safety, collaboration, therapeutic alliance, personalized risk assessment, and timeliness/urgency. From these, the authors derive design features, such as automated acknowledgments that quickly hand off to a human. Notably, no participant was comfortable with LLM-generated automated responses. The work shows how eliciting stakeholder values can ground the design of tools for sensitive clinical communication (Slater et al., 2027).

Read together, these papers are complementary. Digital traces show the structure and rhythm of communication across thousands of care episodes, while value-elicitation interviews bring out the tacit judgment, relational work, and competing obligations that no log records.

2.3. *Whose Voice and Whose Performance Counts*

A central insight of sociotechnical research is that data and metrics reflect the settings in which they were produced and the choices of those who built them. Two papers apply this lens to the evidence and evaluation practices behind clinical AI.

Ma et al. examine how the setting in which text is produced shapes the evidence for adverse drug events (ADEs). They compare six text corpora: patient forum narratives, multilingual social media posts, medical case reports, clinical notes (2 sources), and biomedical literature. A shared topic analysis of 6,713 drug-masked, ADE-positive texts found that patient forum narratives brought out experiences such as suicidal or panic-related thoughts, night sweats, vivid dreams, and memory difficulties. The other corpora were dominated by a generic medication, sleep, and fatigue theme (51.7–64.6% of instances). Linking drug mentions to the curated SIDER knowledge base exposed large gaps: exact-match coverage was only 9.5% for patient forum narratives but 91.0% for biomedical literature. In detection experiments, a QLoRA fine-tuned Llama-3.2-3B scored the best micro-averaged F1 scores (0.80–0.91) on the three corpora with the most training data, and SIDER-grounded retrieval with Llama-3.1-405B performed best on Dutch clinical notes (F1=0.95) and literature (F1=0.91). The work shows that who produces the evidence, and why, shapes what pharmacovigilance can see, and that curated knowledge bases can underrepresent patients' own language (Ma et al., 2027).

Narayan et al. focus on how clinical AI is evaluated. They present EchoStature, a model for pediatric left ventricular ejection fraction that turns patient demographic attributes into a synthetic echocardiogram clip and combines it with the real scan through a learned gate. On the EchoNet-Pediatric benchmark, EchoStature reaches a mean absolute error (MAE) of 4.40, compared with 4.52 for a video-only reference model, and its predictions change by at most 0.06 percentage points when demographic inputs are zeroed or shuffled. The authors' main finding, however, is about evaluation rather than architecture. Error varies more than ten times as much across pediatric age groups as between the two models: infants aged 0–1 year reach 5.23 MAE, compared with 3.90 for children aged 2–5. A model reported with an overall MAE of 4.40 therefore, has an MAE of 5.23 for the youngest patients, for whom echocardiography is often the only available measure of heart function. The work shows how pooled benchmarks can hide the patients a system serves worst and argues for making subgroup results standard in reporting (Narayan et al., 2027).

Both papers make a point that echoes ethnographic findings: the way a technology's evidence is gathered and its performance measured decides whose experience counts. Patient-reported harms that knowledge bases cannot match and infants whose errors are averaged away are both largely invisible in pooled data.

3. Conclusion

The papers in this session show a field learning to study technology in practice at scale without losing the context that gives observation its meaning. Three themes stand out. First, multimodal and trace data can operationalize ethnographic concepts such as attention, rapport, reciprocity, and voice, making them measurable across many encounters and teams. Second, computational scale and qualitative depth complement each other rather than compete: logs and recordings show patterns, while interviews and observations explain them. Third, every data source and metric reflects the setting that produced it, so representing patients' experience requires deliberately auditing what our records, knowledge bases, and benchmarks leave out.

These themes motivate the Observer Consortium, which this session launches. Building on the Observer repository of multimodal primary care encounters (Johnson et al., 2026), the Consortium aims to create a federated network of scientists and multimodal clinical datasets from partner institutions, with shared standards for capture, de-identification, provenance, and governance. Several papers here already show what such data make possible and why careful documentation of where data comes from matters. We invite computer scientists, informaticians, social scientists, clinicians, and others to join in building

technologies that are technically robust, fairly implemented, and grounded in the realities of clinical practice.

References

- Alasaly, B. et al. (2027) “A multimodal framework for assessing patient-physician engagement in ambient AI scribe evaluations,” in Pacific Symposium on Biocomputing (PSB 2027).
- Ash, J.S., Berg, M. and Coiera, E. (2004) “Some unintended consequences of information technology in health care: the nature of patient care information system-related errors,” *Journal of the American Medical Informatics Association*, 11(2), pp. 104–112. Available at: <https://doi.org/10.1197/jamia.M1471>.
- Berg, M. (1999) “Patient care information systems and health care work: a sociotechnical approach,” *International Journal of Medical Informatics*, 55(2), pp. 87–101. Available at: [https://doi.org/10.1016/S1386-5056\(99\)00011-8](https://doi.org/10.1016/S1386-5056(99)00011-8).
- Holden, R.J. et al. (2013) “SEIPS 2.0: a human factors framework for studying and improving the work of healthcare professionals and patients,” *Ergonomics*, 56(11), pp. 1669–1686. Available at: <https://doi.org/10.1080/00140139.2013.838643>.
- Johnson, K.B. et al. (2026) “Observer: creation of a novel multimodal dataset for outpatient care research,” *Journal of the American Medical Informatics Association*, 33(2), pp. 424–433. Available at: <https://doi.org/10.1093/jamia/ocaf182>.
- Ma, Y. et al. (2027) “Adverse Drug Events Across Data-Production Contexts: Multilingual Detection, Alignment, and Cross-Genre Discourse Analysis,” in Pacific Symposium on Biocomputing (PSB 2027).
- Narayan, S. et al. (2027) “EchoStature: Spatial Demographic Conditioning for Pediatric Ejection Fraction Regression, with Subgroup-Stratified Evaluation,” in Pacific Symposium on Biocomputing (PSB 2027).
- Osweiler, B.W. et al. (2027) “Using Multimodal Digital Traces to Characterize Healthcare Team Communication Behaviors,” in Pacific Symposium on Biocomputing (PSB 2027).
- Raj, A. et al. (2027) “Do Clinical Notes Reflect Clinical Conversations? A Sentiment Analysis using the Observer Dataset,” in Pacific Symposium on Biocomputing (PSB 2027).
- Rule, A., Chiang, M.F. and Hribar, M.R. (2020) “Using electronic health record audit logs to study clinical activity: a systematic review of aims, measures, and methods,” *Journal of the American Medical Informatics Association*, 27(3), pp. 480–490. Available at: <https://doi.org/10.1093/jamia/ocz196>.
- Sittig, D.F. and Singh, H. (2010) “A new sociotechnical model for studying health information technology in complex adaptive healthcare systems,” *Quality and Safety in Health Care*, 19(Suppl 3), pp. i68–i74. Available at: <https://doi.org/10.1136/qshc.2010.042085>.
- Slater, H., Walsh, C.G. and Novak, L.L. (2027) “Value Sensitive Design of Technology Supporting Suicide Risk Assessment and Intervention,” in Pacific Symposium on Biocomputing (PSB 2027).
- Street, R.L., Jr. et al. (2009) “How does communication heal? Pathways linking clinician–patient communication to health outcomes,” *Patient Education and Counseling*, 74(3), pp. 295–301. Available at: <https://doi.org/10.1016/j.pec.2008.11.015>.
- Tierney, A.A. et al. (2024) “Ambient artificial intelligence scribes to alleviate the burden of clinical documentation,” *NEJM Catalyst Innovations in Care Delivery*, 5(3). Available at: <https://doi.org/10.1056/CAT.23.0404>.

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

Basam Alasaly, MS^{1*}, Matthew Hill¹, Sriharsha Mopidevi¹, Kevin B. Johnson, MD, MS^{1,3}, Anna Schoenbaum, DNP, MS, RN-BC⁵, Srinivas Sridhara, PhD, MHS⁵, 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.

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.

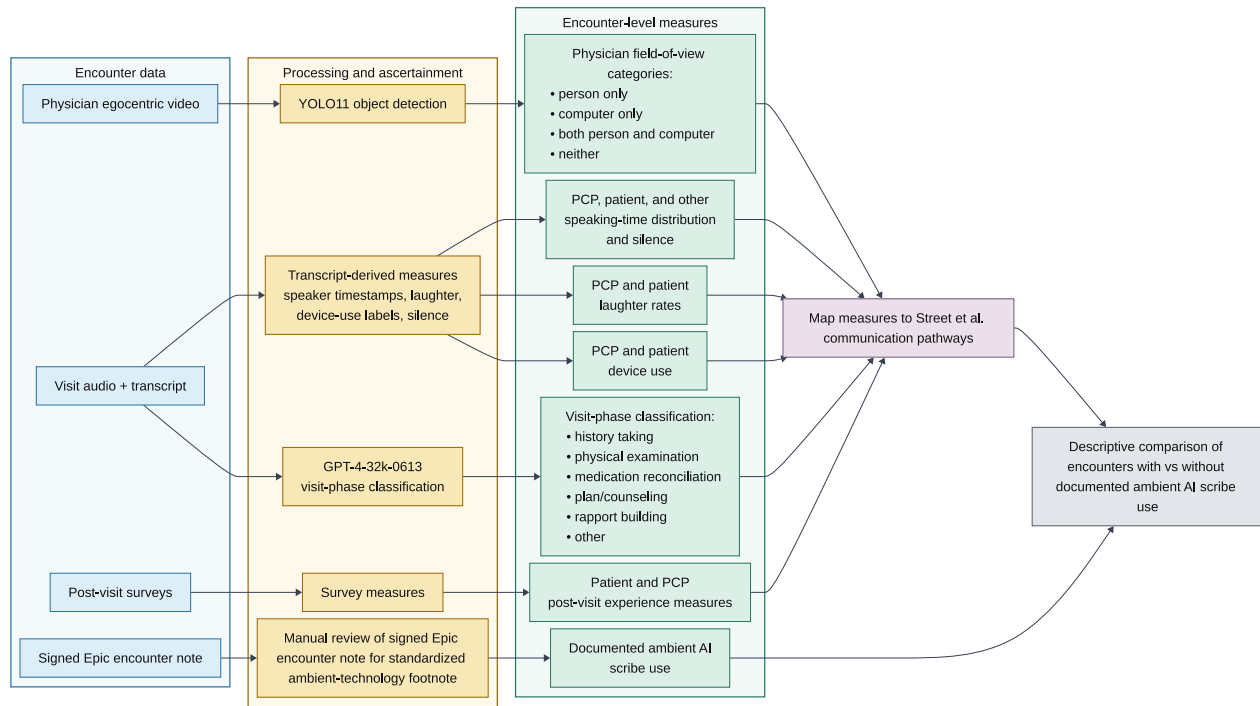


Figure. 1. Multimodal processing framework for encounter-level measurement of patient-physician communication and engagement.

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. Aloa2400659.
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. Aloa2500945.
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. Amofo, 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. Alpc2401260.
46. Allen, L.N., et al., *Artificial intelligence in primary care: frameworks, challenges, and guardrails*. The Lancet Primary Care, 2025.

Do Clinical Notes Reflect Clinical Conversations? A Sentiment Analysis using the Observer Dataset

Aishwarya Raj¹, Kenneth Yi¹, Rohith Palli MD, PhD^{1,3}, Pavankumar Suresh¹, Catherine Chia MS¹, Marvin Tran¹, Barbara D. Lam MD^{1,2}, Andrea L. Hartzler PhD^{1,3}, Patrick Wedgeworth MD, MS^{1,3}

¹*Department of Biomedical Informatics & Medical Education, University of Washington*

²*Division of Hematology & Oncology, Department of Medicine*

³*IT Services, UW Medicine*

University of Washington, Seattle, WA 98195, U.S.A.

Email: araj3609@uw.edu

Sentiment in clinical conversations may reflect aspects of patient experience relevant to care. Clinical notes compress these rich conversations into brief records, and prior work shows that important aspects of the provider-patient visit, such as patient life context, are often left out. The extent to which the affective content of the conversation is preserved in the note remains unknown. We investigated document fidelity by comparing sentiment in 89 clinical conversations from the Observer repository with sentiment in the corresponding notes using TRACE, a sentiment analysis pipeline built on Doc2Vec that scores text along five dimensions (sociability, morality, agency, ability, and status), each with positive and negative valence, for ten dimensions in total. TRACE also reports the lexicon terms contributing to each score, letting us inspect the vocabulary behind each measure. We computed Spearman correlations between transcript and note sentiment and compared which lexicon terms appeared in each source. Note sentiment correlated with provider speech on seven of the ten measures ($r_s \approx 0.25$ to 0.39 , adjusted $p < 0.05$), most strongly for morality, agency, and ability, but showed no significant correlation with patient speech on any measure. Lexicon coverage was higher in provider segmented transcripts than in notes across all sentiment dimensions, with 17.4% of sentiment lexicon terms appearing in transcripts versus 11.6% in notes. Negative ability was the most distinctive sentiment dimension across both sources, characterized by terms such as “slow,” “confused,” and “unable.” These results suggest that clinical conversations hold affective richness, especially sentiment expressed in provider speech, that is only partly preserved in the note and largely absent for patient speech. As ambient AI scribe systems increasingly mediate documentation, understanding which interpersonal and affective elements are retained or omitted may help guide the development of documentation systems that better preserve clinically meaningful context.

Keywords: Sentiment Analysis; Word Embeddings; Clinical Communication; EHR

1. Introduction

The clinical note is the enduring record of a clinical encounter (i.e., “visit”). While the conversation between provider and patient is transient, the note persists as the primary reference for downstream clinical and administrative systems. Because notes are selective summaries, their authors, whether human providers or, increasingly, artificial intelligence (AI) systems with a human in the loop,

© 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.

determine which elements of the encounter are preserved and which are excluded. These judgments occur under substantial time pressure: a time and motion study in ambulatory care found providers spent nearly half of their work day on electronic health record (EHR) and desk work, compared with only about a quarter in direct patient care.¹ Ambient AI scribes can now perform much of this work automatically, generating draft notes directly from the visit conversation^{2,3} while reducing provider documentation time.⁴ Whether notes drafted by AI preserve more, less, or simply different encounter content than notes written by providers remains an open empirical question. What is certain is that every note, regardless of its author, is the product of selection. This raises a question that grows more urgent as AI scribes scale: what does the note keep from the conversation, and what does it leave behind? Among the content at risk is sentiment, the affective meaning conveyed in language, which is the focus of this study.

Prior work establishes three findings that motivate this study. First, much of the clinical conversation never reaches the note,⁵ and the loss is not evenly distributed between speakers: a study of 49 visits across 11 providers found that psychosocial issues voiced by patients reached the note less than half the time, while issues voiced by providers were preserved at much higher rates.⁶ Documentation loss therefore carries a speaker asymmetry that falls disproportionately on the patient voice. Second, affective content is clinically consequential, so its loss would matter. Sentiment, the affective meaning conveyed in language about people, conditions, and events,⁷ shapes care beyond the clinical facts a note records. Affective information conveyed during the visit can improve downstream care and patient engagement, promoting better health outcomes.⁸ Conversely, negative descriptors in notes, such as expressions of disbelief, disapproval, or uncooperativeness,^{9,10} have been shown to shift subsequent providers' attitudes and management decisions.¹¹ Sentiment measured in notes also varies with patient outcomes and demographic characteristics.¹² Third, sentiment is measurable on both sides of the transformation: prior research has quantified sentiment within clinical notes,¹² and other work applies language models to patient and provider speech to capture the emotional context of the interaction.¹³ Because sentiment is consequential and can be measured in both transcripts and notes, the open question is whether the sentiment expressed in the conversation is reflected in the resulting note, particularly for the patient voice that prior work shows is most often lost.

Recent analyses of how providers edit notes drafted by AI begin to characterize the conversation to note transformation, including explicit edits that filter content related to empathy and rapport.¹⁴ Little research, however, has examined the affective dimension of this transformation or the extent to which sentiment is preserved, altered, or lost. This study is an exploratory empirical analysis of how affective and interpersonal language expressed during clinical encounters is preserved in clinical documentation. Our first aim measures the association of five sentiments (i.e., sociability, morality/trust, ability, agency, and status), each with positive and negative valences, between the conversational transcript and the note for a sample of outpatient visits. Because prior work shows that omission falls unevenly across speakers, we analyze provider speech and patient speech as separate segments of the conversation, allowing us to test whether note sentiment tracks each voice differently. These five sentiments derive from the well-established dimensions of warmth and competence in social perception.^{15,16,17} Warmth is associated with sociability and morality (akin to trust), while competence is associated with ability, agency, and status.^{17,18} We measure the five

individual sentiments rather than the two broad dimensions because they are more specific, and each has a validated dictionary suitable for automated text analysis.¹⁸ Our second aim identifies the specific terms that contribute to sentiment scores in notes and transcripts by speaker role. Together, the two aims reveal what sentiment relationships exist between note and transcript, whether those relationships differ for the patient and provider voice, and how the underlying terms drive the observed scores. This study lays the groundwork for tracing documentation fidelity with respect to affective communication in clinical encounters and associated notes.

2. Methods

2.1. Study design

We conducted a descriptive characterization of expressed sentiment in clinical transcripts and notes in a sample of real-world clinical visits from the Observer Repository.²⁰ For each visit, we compared the sentiment expressed in conversation with the sentiment documented in the note. To measure sentiment, we applied TRACE, a Doc2Vec pipeline¹⁹ that scores clinical text along validated sentiment dimensions and links each score to the terms that drive it. Across visits, we evaluate the association of sentiments between transcript and notes using correlation and then inspected terms driving those correlations. Figure 1 summarizes the study design.

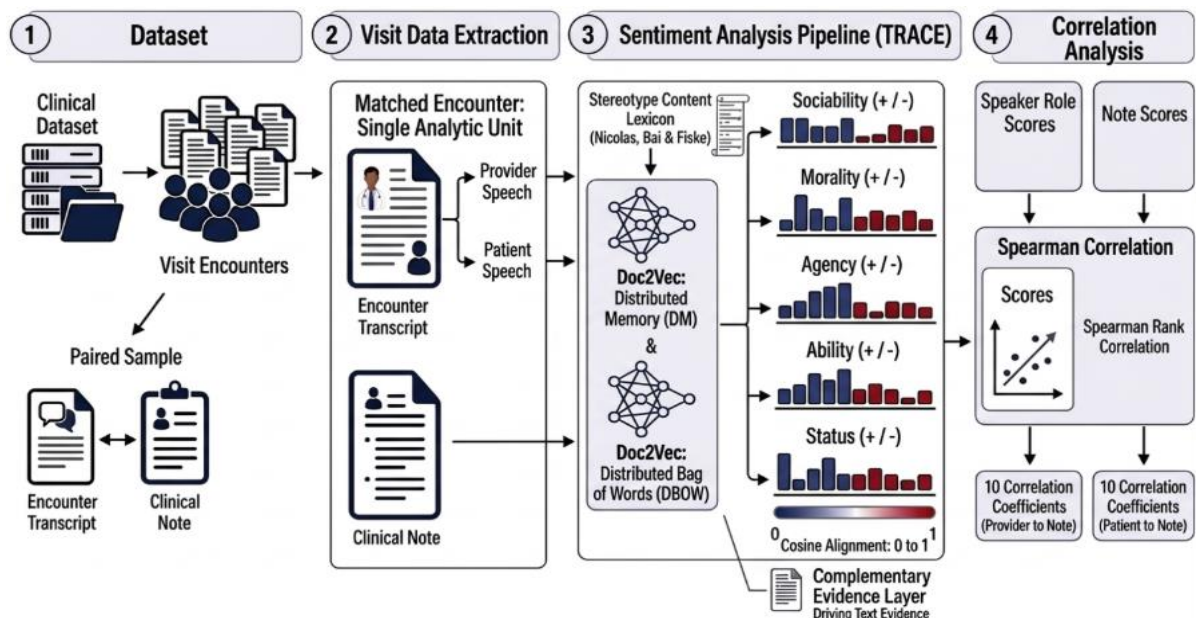


Fig 1. Schematic of study design. Clinical visit transcripts and associated notes were sampled from the Observer Repository. Notes and transcripts were extracted, provider and patient speech were separated, and then TRACE was applied to each of the three sources of clinical text to quantify scores across positive and negative valences of five sentiments: sociability, morality, agency, ability, and status. Spearman correlations were then computed between sentiment scores of speaker-segmented transcripts and clinical notes.

2.2. *Dataset and visit data extraction*

The sample was drawn from the Observer Repository²⁰ and comprised two sources for each clinical visit: a transcript of the recorded encounter and the corresponding primary clinical progress note. Transcripts were generated using WhisperX and de-identified with a local instance of PHilter known as MedVidDeID.²¹ We restricted the sample to visits with a primary progress note and excluded ancillary encounter documentation, including patient instructions, medical assistant notes, and notes limited to vitals or vaccine administration. Annual wellness visits and pre-operative examinations were also excluded because of their limited non-templated content. Visits were matched to transcripts using visit IDs, yielding 89 visits with paired conversation transcripts and notes. Transcripts were then segmented by speaker role, with provider and patient speech analyzed separately, resulting in three data sources per visit: provider speech, patient speech, and the associated clinical note. Speech attributed to other provider-associated roles captured in transcripts (e.g., medical residents or other care team members) was excluded from the primary analysis. Restricting analyses to a single speaker role was intended to characterize expressed sentiment from each speaker. However, because clinical notes may be collaboratively authored or partially drafted by team members, note authorship may not perfectly align with the speaker-role segmentation used in the transcript analysis. Before sentiment scoring, phrases appearing repetitively across a large fraction of documents (e.g., “Pt Name Verified”) were identified as template text and removed. This preprocessing step was included because language shared broadly across documents, such as templated notes or recurring conversational structure, can artificially inflate apparent document similarity.

2.3. *TRACE sentiment analysis*

To analyze sentiment, we applied TRACE, which uses the stereotype content dictionaries of Nicolas, Bai, and Fiske¹⁸ for five dimensions: sociability, morality, ability, agency, and status. Each dictionary marks terms in a positive and negative valence, yielding 10 sentiment scores in total. TRACE, our Doc2Vec-based word embedding model,¹⁹ uses this previously validated social perception lexicon as a sentiment measurement framework for clinical text (e.g. transcripts, notes). TRACE embeds each document and the sentiment lexicon word lists in a shared vector space, then scores every document by its cosine similarity to each sentiment’s vocabulary. Because each of the five sentiments (Table 1) have separate positive and negative lists, each document receives ten sentiment scores, ranging from -1 to 1 , where higher values indicate stronger semantic alignment. To obtain stable and reproducible scores, each sentiment dimension vector was computed as the average of 20 independent inference runs, a fixed random seed was used for model training, and training was restricted to a single thread so that results were reproducible across runs.^{22, 23}

Table 1. TRACE Sentiment, associated attributes, and representative language.

Sentiment	Attributes	Examples of Positive Language	Examples of Negative Language
Sociability	Captures perceptions of friendliness	Outgoing, Friendly	Frosty, Withdrawn
Morality	Captures perceptions of trust	Sincere, Fair	Abuse, Unkind
Agency	Captures perceptions of competence, and assertiveness	Determined, Autonomous	Cautious, Inactive
Ability	Captures perceptions of skill and capability	Educated, Alert	Slow, Clumsy
Status	Captures perceptions of social ranking, power	Privilege, affluent	Hardship, Exclude

2.4. Sentiment correlation between notes and transcripts

The three data sources (i.e., patient speech, provider speech, and clinical notes) were analyzed using the same Doc2Vec modeling framework but embedded in separate vector spaces; therefore, raw scores were not directly comparable across sources. Consistent with this, score distributions occupied different ranges: median scores ranged from 0.30 to 0.54 in notes, 0.36 to 0.44 in patient speech, and 0.13 to 0.23 in provider speech. Within each source, however, distributions were approximately symmetric across all dimensions (absolute skewness ≤ 0.53). To facilitate descriptive comparison across sources, each TRACE dimension was standardized within source to a mean of 0 and standard deviation of 1 (z-score transformation); these standardized values are reported in Table 2. Because z-score standardization is a monotonic linear transformation within a data source, it preserves the rank ordering of visits and therefore does not affect the Spearman correlations reported below. Accordingly, we evaluated concordance between speaker role in transcript and notes, using Spearman rank correlations, which assess agreement in relative ordering rather than raw magnitude and do not require scores to share a common scale.²⁴ For each TRACE dimension, we computed a Spearman correlation between note-based and conversation transcript-based sentiment scores across the 89 visits, yielding 10 correlation coefficients. Because these analyses were exploratory and the TRACE dimensions are conceptually related, p-values were adjusted using the Benjamini–Hochberg false discovery rate procedure rather than the more conservative Bonferroni correction, which assumes independent tests.^{25,26}

2.5. Inspection of terms underlying sentiment scores

TRACE is built on Doc2Vec embeddings, which allow sentiment dimensions to capture semantically related terms beyond those explicitly included in the lexicon. Directly examining the learned embedding space would most faithfully characterize what the model captures. However, for this combination of model and corpus size, the learned dimension vectors were dominated by document topics and generalized language rather than sentiment-discernable content. We therefore focused inspection on how the pre-defined sentiment lexicons appeared within the transcript and note corpora.

To characterize sentiment expression across conversation transcripts and visit notes, lexicon terms were normalized using the same preprocessing pipeline applied to both document types. Because highly frequent terms provide limited discriminative value and may reflect habitual or structural language rather than clinical evaluative meaning, we excluded lexicon terms appearing in more than 60% of documents, applying the filter separately to transcripts and notes.²⁷ In conversational speech, habitual language included common discourse fillers or interactional terms (e.g., “right,” “good,” “know,” “well”) that occur frequently independent of sentiment. In clinical notes, structural language referred to recurring documentation or template-related terms (e.g., “visit,” “date,” “name”) that inflate dictionary matches without conveying interpersonal evaluation. Thresholds from 40% to 100% document frequency were examined, and 60% was selected because it removed predominantly structural or habitual language while preserving clinically and contextually meaningful sentiment terms.²⁸ After filtering, we quantified sentiment lexicon representation for each dimension by calculating (1) the proportion of lexicon terms observed in the corpus (provider-speech segmented transcripts and notes), (2) document frequency, defined as the number of documents containing each term at least once, and (3) collection frequency, defined as the total number of occurrences across all documents.

3. Results

3.1. Visit characteristics and sentiment score distributions

The final dataset comprised 89 visits contributed by 88 unique patients and 9 providers, with one patient contributing two visits. Visit length averaged 30.5 minutes (SD=14.1; range =8.68-79.75 min). Providers had a mean age of 45.4 years (SD =10.7; range =34-62 years) and were predominantly men (77.8%) and White (55.6%), with no provider identifying as Hispanic or Latino. Patients were substantially older, with a mean age of 58.5 years (SD=16.8; range= 24-89 years), and were close to evenly divided by gender (46 women, 52.3%). Patient race and ethnicity differed substantially from the provider group: 43 patients identified as Black or African American (48.9%) and 34 as White (38.6%), and 5 identified as Hispanic or Latino (5.7%). Race was not reported for 8 patients (9.1%) and 1 provider (11.1%).

After within-dataset standardization using Z-score normalization, TRACE sentiment score distributions were approximately symmetric across all 10 dimensions (absolute skewness ≤ 0.53), with no evidence of extreme outliers (Table 2). Standardized score ranges were broadly similar across notes, patient speech, and provider speech (median standardized score range -0.18 to 0.34), supporting the use of Spearman rank correlations to evaluate transcript and note concordance.

Table 2. Standardized (z-score) sentiment score distributions by TRACE dimension and source across 89 visits (values expressed in standard deviation units).

Sentiment Dimension	Dataset	Median	Min	Max	Range
Positive Sociability	notes	0.337	-2.209	1.507	3.716
	patient	0.019	-2.655	2.299	4.955
	provider	0.125	-2.169	1.928	4.097

Negative Sociability	notes	0.028	-2.209	1.712	3.921
	patient	-0.072	-2.327	2.285	4.612
	provider	0.155	-2.298	2.474	4.772
Positive Morality	notes	0.063	-2.25	2.053	4.303
	patient	-0.037	-2.452	2.265	4.717
	provider	-0.136	-2.050	2.604	4.654
Negative Morality	notes	-0.176	-1.887	2.423	4.309
	patient	-0.034	-2.262	1.806	4.069
	provider	0.053	-2.734	2.296	5.030
Positive Agency	notes	0.235	-2.242	1.753	3.994
	patient	0.102	-2.446	1.734	4.180
	provider	0.128	-2.536	1.877	4.412
Negative Agency	notes	0.255	-2.185	1.715	3.900
	patient	-0.023	-2.016	2.434	4.451
	provider	0.041	-2.147	2.720	4.866
Positive Ability	notes	0.134	-1.988	1.945	3.933
	patient	0.068	-2.017	2.438	4.454
	provider	-0.017	-1.820	2.852	4.672
Negative Ability	notes	0.152	-2.234	2.539	4.773
	patient	-0.031	-2.739	2.320	5.059
	provider	0.021	-3.090	2.211	5.301
Positive Status	notes	0.005	-1.880	2.232	4.112
	patient	0.045	-2.566	2.069	4.635
	provider	0.140	-2.473	1.964	4.437
Negative Status	notes	0.210	-2.106	1.651	3.757
	patient	-0.085	-2.273	2.605	4.878
	provider	-0.011	-2.556	3.041	5.597

3.2. *Sentiment concordance between note and conversation*

Table 3 shows correlations between sentiment scores in visit notes and sentiment scores computed separately for provider speech and patient speech from transcripts for the same visits. Provider speech showed weak to moderate positive associations with the corresponding note in 7 of 10 sentiment dimensions. Strength of associations among significant sentiment dimensions were modest ($r_s = 0.25$ to 0.39), with the strongest associations evident for positive and negative ability, morality, and agency. Patient speech showed no measurable association with note sentiment on any dimension. Therefore, subsequent lexicon coverage and term-inspection analyses were limited to provider speech and visit notes.

Table 3. Spearman correlations (r_s) between note and transcript sentiment scores by speaker role ($N = 89$ visits). *p.adj (adjusted p value) < 0.05 after Benjamini–Hochberg correction.

Dimension	Provider speech vs. note		Patient speech vs. note	
	r_s	p.adj.	r_s	p.adj.
Positive sociability	0.148	0.185	0.064	0.791
Negative sociability	0.304	0.010*	0.026	0.898
Positive morality	0.387	0.001*	−0.130	0.622
Negative morality	0.250	0.026*	−0.124	0.622
Positive agency	0.358	0.001*	0.189	0.622
Negative agency	0.270	0.018*	0.155	0.622
Positive ability	0.387	0.001*	0.037	0.898
Negative ability	0.393	0.001*	0.007	0.948
Positive status	0.198	0.079	0.068	0.791
Negative status	0.056	0.605	−0.068	0.791

3.3. Lexicon coverage and term level inspection

Table 4 shows the number of sentiment lexicon coverage in provider-speech segmented transcripts and notes. Only a minority of each lexicon's terms occurred in transcripts and notes, however lexicon coverage was higher in transcripts than in visit notes for all ten dimensions. In total, 17.4% of lexicon terms appeared in transcripts versus 11.6% in notes, and per-sentiment dimension coverage ranged from 11.9% to 28.6 in transcripts and from 6.0% to 21.2% in notes. Among the seven sentiment dimensions with significant provider-speech associations, the transcript–note coverage gap was largest for positive ability (+7.4 pp) and smallest for negative agency (+1.1 pp).

Table 4. Sentiment lexicon coverage by document type for sentiment dimensions. Number of lexicon terms appearing at least once in the corpus ($N = 89$ patient visits). Note. pp = percentage points. Coverage is higher in transcripts than in visit notes for all dimensions.

Dimension	Lexicon terms	Transcripts (Provider-Speech) n (%)	Visit notes n (%)	Difference (pp)
Pos. Sociability	497	105 (21.1%)	64 (12.9%)	+8.2
Neg. Sociability	429	55 (12.8%)	33 (7.7%)	+5.1
Pos. Morality	470	75 (16.0%)	50 (10.6%)	+5.4
Neg. Morality	1841	219 (11.9%)	123 (6.7%)	+5.2
Pos. Agency	447	108 (24.2%)	85 (19.0%)	+5.2
Neg. Agency	270	41 (15.2%)	38 (14.1%)	+1.1
Pos. Ability	604	173 (28.6%)	128 (21.2%)	+7.4
Neg. Ability	300	37 (12.3%)	18 (6.0%)	+6.3
Pos. Status	289	86 (29.8%)	56 (19.4%)	+10.4
Neg. Status	190	31 (16.3%)	24 (12.6%)	+3.7
All dimensions	5337	930 (17.4%)	619 (11.6%)	+5.8

Table 5 shows the top 5 terms by sentiment type in provider-speech segmented transcripts and notes. The most frequent terms were generic (e.g., "set", "sense", "back") but differed in character between the two sources: conversational vocabulary dominated the transcripts (e.g., "wow", "best", "bad"), whereas clinical vocabulary dominated the notes (e.g., "care", "blood", "risk"). Negative ability was the clear exception: the top terms in both sources were specific to this sentiment dimension (e.g., "slow", "confused", "inability"). However, it was also the sparsest sentiment dimension, with no term exceeding 20% document frequency. On the other hand, our simple filter for high-frequency terms excluded words likely to be attributed to speech mannerisms for visit transcripts (e.g., "right", "know", "see") and templated language for clinical notes (e.g., "date", "name", "visit").

Table 5. Top 5 terms per significant sentiment lexicon and document type. Terms are ranked by document frequency within each dimension (N = 89). *Note.* DF = document frequency, the number of visits whose text contains the term at least once. CF = collection frequency, the total number of occurrences across all visits.

Dimension	Transcripts (Provider-Speech)				Visit notes		
	Rank	Term	DF n (%)	CF	Term	DF n (%)	CF
Negative sociability	1	bad	49 (55.1)	98	negative	38 (42.7)	141
	2	turn	32 (36.0)	60	rash	19 (21.3)	25
	3	feet	21 (23.6)	36	shortness	18 (20.2)	26
	4	love	18 (20.2)	32	feet	18 (20.2)	20
	5	alone	15 (16.9)	20	acid	14 (15.7)	16
Positive morality	1	best	46 (51.7)	84	blood	43 (48.3)	109
	2	sense	43 (48.3)	90	care	40 (44.9)	96
	3	helpful	39 (43.8)	88	right	36 (40.4)	85
	4	real	38 (42.7)	66	good	27 (30.3)	57
	5	sit	36 (40.4)	50	present	25 (28.1)	28
Negative morality	1	set	51 (57.3)	84	weight	45 (50.6)	83
	2	bad	49 (55.1)	98	drug	45 (50.6)	50
	3	using	47 (52.8)	98	blood	43 (48.3)	109
	4	four	47 (52.8)	135	mass	42 (47.2)	51
	5	worse	42 (47.2)	84	male	41 (46.1)	64
Positive agency	1	set	51 (57.3)	84	risk	48 (53.9)	106
	2	hard	49 (55.1)	117	taking	46 (51.7)	82
	3	working	47 (52.8)	93	active	38 (42.7)	41
	4	risk	38 (42.7)	92	work	30 (33.7)	53
	5	life	37 (41.6)	77	low	25 (28.1)	55
Negative agency	1	follow	44 (49.4)	89	follow	41 (46.1)	72
	2	patient	37 (41.6)	116	needs	34 (38.2)	65
	3	needs	26 (29.2)	33	soft	25 (28.1)	28
	4	safe	23 (25.8)	38	anxiety	20 (22.5)	57
	5	feet	21 (23.6)	36	help	18 (20.2)	24

Positive ability	1	set	51 (57.3)	84	see	36 (40.4)	72
	2	order	50 (56.2)	116	start	31 (34.8)	43
	3	thought	46 (51.7)	88	able	31 (34.8)	98
	4	reason	43 (48.3)	83	good	27 (30.3)	57
	5	sense	43 (48.3)	90	hearing	26 (29.2)	55
Negative ability	1	slow	15 (16.9)	17	mental	18 (20.2)	25
	2	mental	11 (12.4)	21	unable	12 (13.5)	22
	3	green	9 (10.1)	14	rigidity	7 (7.9)	8
	4	ignore	8 (9.0)	8	slow	5 (5.6)	5
	5	confused	7 (7.9)	9	inability	4 (4.5)	7

4. Discussion

This study examined how sentiment expressed during clinical conversations was reflected in the corresponding clinical note. Using TRACE to score sentiment in clinical notes and transcripts across visits, we found that provider speech showed modest but significant associations with note sentiment across seven of ten dimensions, particularly morality, agency, and ability, whereas patient speech showed no measurable association with note sentiment on any dimension. Lexicon coverage was also consistently higher in transcripts than in visit notes, indicating that clinical conversations contain a broader range of sentiment-related language than is retained in notes. The observed asymmetry between provider and patient sentiment may partly reflect communication-related factors not captured by correlation, including differences in talk-time distribution and shorter length patient speech that may reduce sensitivity to sentiment associations. The asymmetry may also reflect the function and authorship of the clinical note itself. Visit notes are provider-authored documents designed to support clinical reasoning, care coordination, billing, and regulatory requirements²⁹, rather than to reproduce the encounter. As a result, documentation may preferentially preserve provider framing, assessment, and decision-making language while compressing or omitting aspects of patient emotional expression and interpersonal exchange.¹⁴ This interpretation is consistent with prior work demonstrating that clinically relevant contextual and socioemotional information is often excluded from documentation, contributing to contextual errors in care.^{5,30} Under this framing, patterns of concordance and divergence between conversation and note may help characterize which affective elements, such as sentiment, and clinical elements are carried forward into documentation and which remain primarily embedded within the spoken encounter. More broadly, sentiment-based measures may offer a scalable way to study how loss of socioemotional context in documentation relates to downstream outcomes such as diagnostic accuracy, care quality, or patient safety.

The absence of detectable patient-sentiment transfer extends prior work^{5,6} showing that patient concerns are incompletely represented in clinical documentation. Our findings suggest that this selective loss may involve not only patient-reported experience, but also affective content. At the same time, the modest effect sizes observed for provider-note concordance (r_s approximately 0.25-0.39) indicate that even provider sentiment is only partially preserved in documentation. These findings support the view of the clinical note as a condensed⁵ and task-oriented representation³¹ of a substantially richer interaction.

TRACE also addresses a comparability problem: because sentiment scores depend on the representation and scoring approach used,^{5,7} sentiment measured in conversations and sentiment measured in notes are not directly comparable unless the same instrument is applied to both. By evaluating conversations and notes using the same social-perception sentiment dimensions and linking scores back to the underlying lexicon terms, TRACE provides a shared and interpretable framework for comparing affective language across modalities. Inspection of high-frequency terms suggested that some TRACE sentiment dimensions captured not only affective meaning but also broader clinical topicality, including biomedical and documentation-related language. At the same time, affective meaning in clinical communication is often embedded within biomedical framing rather than expressed through overt emotional vocabulary alone. Consistent with this, term-level inspection revealed substantial variability across dimensions. Some dimensions, particularly negative ability, were characterized by interpretable evaluative terms such as ‘slow,’ ‘confused,’ and ‘unable,’ whereas others were dominated by more general clinical vocabulary including ‘blood,’ ‘weight,’ and ‘risk.’ Accordingly, TRACE-derived scores should be interpreted as exploratory measures of affective-semantic concordance rather than definitive sentiment estimates, highlighting the need for future methods that better distinguish interpersonal sentiment from biomedical context.⁷

Interpretation of these findings should consider the study population and dataset characteristics. This study was conducted in a single academic clinic with a small sample of nine providers, and because providers were aware that visits were being recorded, communication and documentation behaviors may have differed from those in other clinical settings, limiting generalizability. As communication, affective expression, and documentation practices may vary across demographic and sociocultural contexts, the observed patterns may reflect this specific sample and clinical environment. Future work should expand this study to larger and more diverse datasets, including analyses across patient-provider demographic combinations, talk-time distribution, and institutional settings to better understand sentiment manifestation in conversation and documentation. Additional methodological work could assess large language and contextual embedding approaches, validate TRACE-derived sentiment dimension terms against human ratings or communication frameworks such as RIAS,³² and examine how ambient documentation systems preserve, alter, or omit affective and interpersonal content during automated summarization.

5. Conclusion

We present a descriptive analysis of how sentiment is preserved from the clinical conversation to the clinical note documented in a sample of visits, using TRACE to score both sources along validated social perception sentiment dimensions and to link each score back to its underlying language. Provider-expressed sentiment corresponded modestly but reliably with the note on the morality, agency, and ability sentiments, while patient-expressed sentiment showed no measurable relationship to the note, supporting that the written record functions as a provider-framed rendering of the visit rather than a balanced account of both speakers. These findings lay the groundwork for future research to explore representation of patient narratives in context of care outcomes, and on how ambient scribe systems may leverage meaningful conversational participation representation from both the patient and provider.

6. Acknowledgments

This study was supported by the Observer Pilot Award from University of Pennsylvania and the Institute for Medical Data Science (IMDS) Pilot Award from the University of Washington. We wish to thank our advisory board, Drs. Trevor Cohen, Janice Sabin, Wanda Pratt, and Brian Wood, Priscilla Lui, and the invaluable efforts of all team members who contributed to the richness of our analysis.

References

1. Sinsky CA, Colligan L, Li L, et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Ann Intern Med*. 2016;165(11):753-60.
2. Haberle T, Cleveland C, Snow GL, Barber C, Stookey N, Thornock C, et al. The impact of nuance DAX ambient listening AI documentation: a cohort study. *J Am Med Inform Assoc*. 2024;31(4):975-9. doi:10.1093/jamia/ocae022.
3. Duggan MJ, Gervase J, Schoenbaum A, Hanson W, Howell JT 3rd, Sheinberg M, et al. Clinician experiences with ambient scribe technology to assist with documentation burden and efficiency. *JAMA Netw Open*. 2025;8(2):e2460637. doi:10.1001/jamanetworkopen.2024.60637.
4. Rotenstein LS, Holmgren AJ, Thombley R, Sriram A, Dbouk RH, Jost M, et al. Changes in clinician time expenditure and visit quantity with adoption of artificial intelligence-powered scribes: a multisite study. *JAMA*. 2026;335(16):1408-17. doi:10.1001/jama.2026.2253.
5. Weiner SJ, Wang S, Kelly B, Sharma G, Schwartz A. How accurate is the medical record? A comparison of the physician's note with a concealed audio recording in unannounced standardized patient encounters. *J Am Med Inform Assoc*. 2020;27(5):770-5. doi:10.1093/jamia/ocaa027.
6. Weiner M, Flanagan ME, Ernst K, Cottingham AH, Rattray NA, Franks Z, et al. Accuracy, thoroughness, and quality of outpatient primary care documentation in the U.S. Department of Veterans Affairs. *BMC Prim Care*. 2024;25(1):262. doi:10.1186/s12875-024-02501-6.
7. Denecke K, Deng Y. Sentiment analysis in medical settings: new opportunities and challenges. *Artif Intell Med*. 2015;64(1):17-27. doi:10.1016/j.artmed.2015.03.006.
8. Street RL Jr, Makoul G, Arora NK, Epstein RM. How does communication heal? Pathways linking clinician-patient communication to health outcomes. *Patient Educ Couns*. 2009;74(3):295-301.
9. Park J, Saha S, Chee B, Taylor J, Beach MC. Physician use of stigmatizing language in patient medical records. *JAMA Netw Open*. 2021;4(7):e2117052. doi:10.1001/jamanetworkopen.2021.17052.
10. Beach MC, Saha S, Park J, Taylor J, Drew P, Plank E, et al. Testimonial injustice: linguistic bias in the medical records of Black patients and women. *J Gen Intern Med*. 2021;36(6):1708-14. doi:10.1007/s11606-021-06682-z.
11. Goddu AP, O'Connor KJ, Lanzkron S, Saheed MO, Saha S, Peek ME, et al. Do words matter? Stigmatizing language and the transmission of bias in the medical record. *J Gen Intern Med*. 2018;33(5):685-91. doi:10.1007/s11606-017-4289-2.

12. Ghassemi MM, Mark RG, Nemati S. A visualization of evolving clinical sentiment using vector representations of clinical notes. *Comput Cardiol* (2010). 2015;2015:629-32. doi:10.1109/CIC.2015.7410989.
13. Chatzimina ME, Papadaki HA, Pontikoglou C, Tsiknakis M. A comparative sentiment analysis of Greek clinical conversations using BERT, RoBERTa, GPT-2, and XLNet. *Bioengineering* (Basel). 2024;11(6):521. doi:10.3390/bioengineering11060521.
14. Guo Y, Hu D, Yang Z, Kim S, Tran B, Lee J, et al. What do clinicians edit in ambient AI-drafted clinical documentation? A qualitative content analysis. *J Am Med Inform Assoc*. 2026;ocag073. doi:10.1093/jamia/ocag073. Epub ahead of print.
15. Fiske ST, Cuddy AJC, Glick P, Xu J. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *J Pers Soc Psychol*. 2002;82(6):878-902.
16. Fiske ST. Stereotype content: warmth and competence endure. *Curr Dir Psychol Sci*. 2018;27(2):67-73.
17. Koch A, Imhoff R, Unkelbach C, Nicolas G, Fiske S, Terache J, et al. Groups' warmth is a personal matter: understanding consensus on stereotype dimensions reconciles adversarial models of social evaluation. *J Exp Soc Psychol*. 2020;89:103995. doi:10.1016/j.jesp.2020.103995.
18. Nicolas G, Bai X, Fiske ST. Comprehensive stereotype content dictionaries using a semi-automated method. *Eur J Soc Psychol*. 2021;51(1):178-96.
19. Le Q, Mikolov T. Distributed representations of sentences and documents. In: *Proceedings of the 31st International Conference on Machine Learning (ICML)*; 2014 Jun 21-26; Beijing, China. PMLR; 2014. p. 1188-96.
20. Johnson KB, Alasaly B, Jang KJ, Eaton E, Mopidevi S, Koppel R. Observer: creation of a novel multimodal dataset for outpatient care research. *J Am Med Inform Assoc*. 2026;33(2):424-33. doi:10.1093/jamia/ocaf182.
21. Mopidevi S, Jang KJ, Alasaly B, Pugh S, Park J, Batugo A, et al. MedVidDeID: protecting privacy in clinical encounter video recordings. *J Biomed Inform*. 2025;170:104901. doi:10.1016/j.jbi.2025.104901.
22. Antoniak M, Mimno D. Evaluating the stability of embedding-based word similarities. *Trans Assoc Comput Linguist*. 2018;6:107-19. doi:10.1162/tacl_a_00008.
23. Lau JH, Baldwin T. An empirical evaluation of doc2vec with practical insights into document embedding generation. In: *Proceedings of the 1st Workshop on Representation Learning for NLP*; 2016 Aug 11; Berlin, Germany. Stroudsburg (PA): Association for Computational Linguistics; 2016. p. 78-86.
24. Schober P, Boer C, Schwarte LA. Correlation coefficients: appropriate use and interpretation. *Anesth Analg*. 2018;126(5):1763-8. doi:10.1213/ANE.0000000000002864.
25. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*. 1995;57(1):289-300.
26. Curtin F, Schulz P. Multiple correlations and Bonferroni's correction. *Biol Psychiatry*. 1998;44(8):775-7. doi:10.1016/S0006-3223(98)00043-2.

27. Luhn HP. The automatic creation of literature abstracts. *IBM J Res Dev.* 1958;2(2):159-65. doi:10.1147/rd.22.0159.
28. Manning CD, Raghavan P, Schütze H. *Introduction to information retrieval.* Cambridge: Cambridge University Press; 2008.
29. Colicchio TK, Cimino JJ. Clinicians' reasoning as reflected in electronic clinical note entry and reading/retrieval: a systematic review and qualitative synthesis. *JAMIA.* 2019;26(2):172–184.
30. Weiner SJ, Schwartz A, Weaver F, Goldberg J, Yudkowsky R, Sharma G, et al. Contextual errors and failures in individualizing patient care: a multicenter study. *Ann Intern Med.* 2010 Jul 20;153(2):69–75. doi: 10.7326/0003-4819-153-2-201007200-00002. PMID: 20643988.
31. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *J Am Med Inform Assoc.* 2011;18(2):181-6. doi:10.1136/jamia.2010.007237
32. Roter D, Larson S. The Roter interaction analysis system (RIAS): utility and flexibility for analysis of medical interactions. *Patient Educ Couns.* 2002;46(4):243-51. doi:10.1016/s0738-3991(02)00012-5

Using Multimodal Digital Traces to Characterize Healthcare Team Communication Behaviors

Bailey W. Osweiler[†]

Division of Computational & Data Sciences, Washington University in St Louis, Saint Louis, USA

Email: b.osweiler@wustl.edu

Laura R. Baratta

Division of Biology and Biomedical Sciences, Washington University School of Medicine in St. Louis, Saint Louis, USA

Email: barattalaurar@wustl.edu

Daphne Lew

Institute for Informatics, Data Science and Biostatistics (I2DB), Washington University School of Medicine in St. Louis, Saint Louis, USA

Email: daphne.lew@wustl.edu

Sunny S. Lou and Thomas Kannampallil

Department of Anesthesiology, Washington University School of Medicine in St. Louis, Saint Louis, USA

Email: slou@wustl.edu, thomas.k@wustl.edu

Jasmine Wu

Department of Communication Studies, Moody College of Communication, University of Texas at Austin, Austin, USA

Email: jasminewu@austin.utexas.edu

Background: Secure messaging and the electronic health record (EHR) are two digital tools that are regularly used in healthcare delivery. **Objective:** This study used multimodal behavioral data to investigate patterns of secure chat messaging use, and to determine whether those patterns are associated with patient or team outcomes. **Methods:** Relational event modeling (REM) was used to identify patterns of secure messaging use for each patient. REM results were investigated for associations with patient length of stay (LOS) in inpatient medicine and aggregate audit log-derived EHR time. **Results:** Secure message responses (“reciprocity”) were common and quick. Reciprocity was associated with shorter LOS, suggesting that teams with higher message response rates may have more efficient patient throughput. In contrast, teams more likely to message a third healthcare professional after receiving a message (i.e., “turn receiving”, AB-BY pattern) were associated with longer LOS, although results were not significant in sensitivity analysis. Message repetition was associated with more EHR time per day. **Conclusion:** While not causal, these results suggest that healthcare team secure messaging patterns are associated with patient and team outcomes. Combining multimodal data (EHR audit logs, secure messaging, clinical data) may reveal common team communication behaviors and their associations with patient and team outcomes.

Keywords: electronic health records; audit logs; chat messaging; length of stay; healthcare team processes; relational event modeling

[†] Work partially supported by grant T32DA015035 and F31DA066305

1. Introduction

1.1. Background

Healthcare professionals spend increasing amounts of time performing digital tasks¹ such as documentation, order entry, and chart review on the electronic health record (EHR), making their digital activities increasingly relevant to understanding healthcare delivery. Primary care physicians, for example, spend half of their overall clinical time on EHR-based activities², demonstrating that EHR tasks are an integral part of modern clinical work.

EHR audit logs, digital traces of EHR actions primarily recorded for security purposes, provide a powerful opportunity to track EHR activities.^{3,4} They can be aggregated to derive measures such as team collaboration^{5–7} and EHR workload^{8–10}. EHR audit log-derived workload measures have been consistently associated with other measures of healthcare professionals' workload.^{2,8,11} Although audit log-based measures provide a valuable approximation of digital activities, their ability to capture other foundational activities, such as team communication, is limited. Studying such behaviors requires complementary data and novel methods.

Healthcare professionals increasingly communicate through secure messaging platforms¹², allowing them to asynchronously share information, ask questions, or clarify details of patient care. Different types of secure messaging applications are now widely used across clinical settings¹³, demonstrating their growing adoption. Prior work has associated messaging volume at the individual healthcare professional-level to outcomes such as decreased patient length of stay (LOS)¹⁴, increased cognitive load,⁹ and increased clinician workload.⁹

By studying multiple digital behaviors together, researchers may be able to gather a more comprehensive understanding of clinical digital activities than by studying only one dataset (audit logs or secure messaging) alone. For example, while EHR audit log data provides an ethnographic view of individual-level EHR workload, secure messaging data may allow an investigation at the team-level, as communication is inherently team-based. Clinical team measures have been extensively researched for their associations with patient outcomes; for example, fragmented and discontinuous physician care teams are associated with longer LOS^{15,16}, while multidisciplinary collaboration is associated with reduced in-hospital mortality¹⁷ but inconclusive results with LOS^{18,19}. Team secure messaging behaviors may similarly be associated with patient outcomes. Since healthcare communication involves multiple parties, an individual's secure messaging habits may additionally be related to their teammates' secure messaging habits and other digital behaviors. Using multimodal digital datasets can begin to reveal the complex relationships between team secure messaging, EHR workload, and patient outcomes.

1.2. Objective

This study aimed to investigate how patterns of team secure messaging are associated with both patient- and clinical team-based outcomes for inpatient hospitalizations, by merging multiple observational digital datasets. Identifying patterns of messages associated with relevant outcomes

may direct future research towards identifying communication inefficiencies and ultimately, improving team collaboration. The research question was: what characteristics of secure messaging interactions among clinicians are associated with patient LOS and team-based EHR workload?

2. Methods

2.1. *Study population*

All inpatient hospitalizations between January 1, 2023 and April 30, 2023 (4 months) at 12 hospitals affiliated with Washington University School of Medicine and BJC Health System were included. Adult patients (18 years or older) admitted to an inpatient medicine clinical service with a hospital LOS 3 days or longer were included. If a patient had multiple encounters meeting these criteria during the study period, only their first encounter was considered. These criteria limited heterogeneity across clinical units and excluded patients who were quickly discharged, who would likely have less variability in secure messaging use, team structure, and EHR time. Medicare's 3-day rule also informed the 3-day cutoff²⁰. All hospitals in the system used the same Epic EHR (Verona, WI).

2.2. *Data*

This study incorporated multiple data sources to capture healthcare professional digital behaviors. The primary behavioral data sources were secure messaging data (model inputs) and EHR audit logs (as a secondary outcome). Additional datasets to identify inclusion criteria and covariates included admission/discharge/transfer (ADT) events, and patient characteristics such as demographics, diagnoses, and discharge disposition. All data were secondary and retrieved from Epic EHR's Clarity database.

Data analysis was conducted from November 2025 to August 2026. This study received approval under a waiver of informed consent from the Institutional Review Board of Washington University School of Medicine (IRB#202305074).

2.2.1. *Secure messaging data*

The primary data source was the secure messaging data, which included all chat conversations when the chat was linked to a patient identifier corresponding to the patients included in the study sample. Secure messaging in our health system is available through both mobile and desktop applications, and both messaging versions were included in this study. Secure messaging conversations were filtered to only those that were dyadic (i.e., excluding group messages that involved multiple healthcare professionals due to less clear response expectations to group messages). All messages associated with included conversations between the patient's admission and discharge times were included. We included messages from and to all healthcare professional roles because this study was exploratory and we did not have role-specific hypotheses; rather, we wanted to measure messaging patterns across the entire care team. Relevant data elements that were retrieved included the conversation ID, patient ID, sender ID, receiver ID, and timestamp. Sender and receiver IDs

were included regardless of the user's clinical role (e.g. nurse, physician, administrative support, technician, etc) or setting (inpatient or outpatient).

2.2.2. *EHR audit logs*

EHR audit logs were extracted from the access logs table (ACCESS_LOG) in the Epic Clarity database. Audit logs included the action description, timestamp, patient ID, encounter ID, healthcare professional ID, healthcare professional role, and action ID. Audit logs were extracted for all healthcare professionals that interacted with eligible patients' EHR between admit and discharge.

2.2.3. *Admission, discharge, and transfer events*

The ADT events were used to identify hospitalizations admitted to a medicine service. Patients could be admitted through emergency medicine, adult emergency medicine, or pediatric emergency medicine specialties so long as they were transferred to a hospitalist or medical specialty in any of the hospitals affiliated with the health system.

2.2.4. *Patient characteristics*

Demographic information for patients seen in eligible hospitalizations was captured, including documented patient sex, race, age, and insurance status. All billing ICD-10 diagnoses associated with the encounter were recorded. Diagnosis-related groups (DRGs) represented the primary billed reason for hospitalization.

2.3. *Measures*

2.3.1. *Length of stay in hours (outcome)*

The primary outcome was patient LOS, calculated as the time difference between admission and discharge times, in hours. LOS is reported as median and interquartile range [Q1, Q3].

2.3.2. *EHR time per day in minutes (outcome)*

Total time on the EHR was used as a proxy for EHR workload, similar to prior research^{21,22}. For each healthcare professional interacting with each patient's chart, the time delta between subsequent EHR audit log actions was calculated and summed across all actions to determine total EHR time. Any time deltas greater than 5 minutes were removed, representing a 5-minute timeout period of inactivity, as has been used in previous research^{10,23,24}. Total EHR time for all healthcare professionals for each patient were aggregated to represent the total EHR time across the entire encounter (in minutes). EHR time was divided by the patient's LOS (in hours) so that the variable is an estimate of relative EHR workload rather than primarily an artifact of the patient's LOS.

2.3.3. *Covariates*

Additional covariates included patient demographics of age, sex (male, female), race (White, Black or African American, or other), and insurance (Medicare, Medicaid, Managed Care, Commercial,

Worker's Comp, Multiple insurance types, or missing). Clinical characteristics controlled for treatment setting and patient clinical complexity, which were hypothesized to be related with clinical workload; treatment setting variables included the hospital name and a binary indicator of whether they spent any time in the intensive care unit (ICU). Clinical complexity was operationalized as the DRG and unweighted Elixhauser comorbidity score. All billed ICD-10 codes were aggregated into an unweighted Elixhauser comorbidity score²⁵, which counts the chronic co-occurring conditions billed out of a list of 38 conditions, and represents overall sickness.

2.4. Data pre-processing

Age, Elixhauser score, messaging volume, and care team size were coded as continuous variables. Log transformation was used for LOS, EHR time per day, care team size, and messaging volume, which were all right-skewed. All other variables were categorical with the majority group used as the reference group (ref; White race, Medicare insurance, female sex, no ICU).

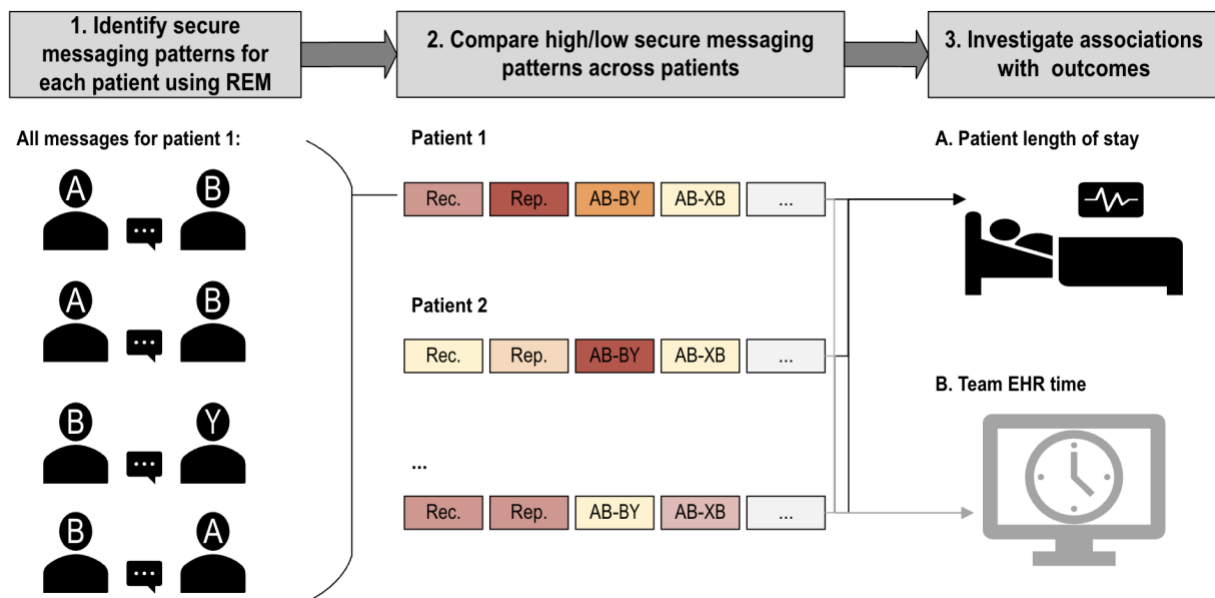


Fig. 1. Methodology to identify secure messaging patterns associations with patient and team outcomes. For each patient, we measured secure messaging patterns of: reciprocity (Rec), repetition (Rep), and four turn-switching behaviors (AB-BY, AB-XB, AB-AY, AB-XA). We investigated associations of each standardized term with outcomes of LOS and EHR time.

2.5. Statistical Approach

The analysis followed three steps (**Figure 1**). First, relational event modeling (REM) was used to identify patterns in secure messaging use at the patient-level. Second, REM estimates were standardized across all patients to derive how messaging patterns compared to the average. Finally, these standardized terms were investigated for associations with the outcomes using mixed effects linear regression. Details on each step are below and in **Figure 1**.

2.5.1. Relational event modeling

Our analytical approach applied REM to identify patterns of secure messaging for each patient (**Figure 1- Step 1**). A growing body of literature in social science has used relational events or similar frameworks to analyze timestamped networked communicative events in various settings^{26–29}. REMs identify patterns of communication by user attribute, dyadic attributes, team attributes, or participation switching. REMs fit a set of survival models to identify the rate of certain communication patterns occurring. They consistently reveal well-understood social network patterns such as reciprocity, triadic closure, and more in varied communication settings^{26–29}.

Table 1. relational event model participation shift patterns

Term	Definition
REM intercept	baseline “hazard” to message, beyond all specified messaging patterns
Reciprocity	likelihood for any person “A” to message person “B”, given the recency that “B” has messaged person “A”.
Repetition	likelihood for any person “A” to message person “B”, given the recency that “A” has messaged person “B”.
Turn continuing (AB-AY)	likelihood that the sender, “A”, of a given message will be the sender to a third person, “Y”, in the very next event.
Turn receiving (AB-BY)	likelihood that the receiver, “B”, of a given message will be the sender to a third person, “Y”, in the very next event.
Turn usurping (AB-XA)	likelihood that the sender, “A”, will be the receiver of the next message by someone who they did not just message.
Turn usurping (AB-XB)	likelihood that the receiver, “B”, will be the receiver of two subsequent messages from two different senders.

An REM was fit for each patient separately to identify their team’s messaging patterns. The REM effects included an intercept for baseline messaging hazard, repetition, reciprocity, and participation shifts for turn switching³⁰. Definitions of each pattern are summarized in **Table 1**.

We exponentiated effect estimates to calculate the hazard ratio (HR), which represents the rate of each pattern occurring relative to all other patterns, as has been done in other research²⁹. For instance, a higher reciprocity HR means the patient’s healthcare team responded at a higher rate (more frequently and/or more rapidly). HRs are interpreted relative to a null of 1: a pattern with HR greater than 1 occurred at a greater rate, while a pattern with HR less than 1 occurred at a lower rate. Estimates were capped at the largest value the computer could store – $\exp(709)$ – where predictive power was near-perfect and exponentiated HR became too large. Results are reported as median, quartile 1 (Q1) and quartile 3 (Q3).

2.5.2. Standardization

The resulting (un-exponentiated) REM effect estimates (intercept, repetition, reciprocity, four participation shifts) were standardized across all patients in the cohort. This transformed each estimate to have a mean of zero and a standard deviation of one (“z-score”), allowing us to assess how a patient’s team’s secure messaging habits differ from the rest (**Figure 1-Step 2**). Patients with a positive z-score have “high” reciprocity, for example, if their team’s secure messaging patterns revealed higher-than-average rate of response. They have “low” repetition, for example, if their

team's secure messaging patterns revealed a negative standardized repetition term, indicating a lower-than-average rate of messaging the same recipient repeatedly.

2.5.3. *Mixed effects linear regression*

Finally, two linear regressions investigated the association of messaging patterns with dependent variables of LOS (hours) and EHR time per day (minutes; **Figure 1-Step 3**). The independent variables were the standardized REM estimates (repetition, reciprocity, participation shifts). Covariate fixed effects included demographics (race, sex, age, insurance) and clinical variables (ICU and Elixhauser comorbidity score). Random effects included the hospital at which the patient was treated and their DRG. Results are reported as beta coefficients and 95% confidence intervals (CI).

2.5.4. *Sensitivity analysis*

Several secondary analyses investigated robustness of the results to model specification. Results are presented as stepwise analyses (models 1-4) to demonstrate how effect estimates change with the inclusion of subsequent terms. First, models included only reciprocity, repetition, and REM intercept (model 1), then added the four participation shifts (model 2). Then, models were modified to include fixed effects for the messaging volume (number of messages, model 3) and care team size (number of unique secure messaging users, model 4). This analysis was intended to investigate whether associations between REM patterns and outcomes were explainable by messaging volume and care team size, or if the micro-patterns of messaging, independent of volume and care team size, explain the relationships. Second, the patient cohort was filtered to only those with at least three unique secure messaging users, since teams with only two users will not demonstrate any participation shifting behaviors from which the model can learn. Although REMs have been proven to produce reliable explanatory estimates even with small risk sets³¹, this sensitivity analysis reduced the uncertainty of REM estimates by filtering to patients with more data on which to train.

2.5.5. *Software*

Audit log, secure messaging, and ADT data preprocessing was done using Python version 3.10.6. Statistical analysis was done using R version 4.4.0 and packages relevant³² for REM, lme4³³ and lmerTest³⁴ for mixed-effects linear models, and sjPlot³⁵ for summary tables.

3. Results

3.1. *Summary statistics*

6,737 hospitalizations met inclusion criteria and had at least one dyadic secure message during their encounter, which constituted our final analytic sample. Patients had a median LOS of 136 [Q1 96.6, Q3 202] hours and median EHR time per day of 327 [272, 400] minutes. The median secure messages per patient during their hospitalization was 29 [14, 60] messages across 8 [4, 14] conversations and a care team of size 10 [6, 16] healthcare professionals involved in messaging. 3,643 (54.1%) patients were female, 4,276 (63.5%) were White, 4,468 (66.3%) had Medicare insurance, and the median age was 67 [56, 78] years. 856 patients (12.7%) had any amount of time

in the ICU during their hospitalization and median Elixhauser score was 5 [4, 7]. The most common DRGs were “septicemia or severe sepsis without MV >96 hours with major complication and comorbidity (MCC)” (n=509, 7.6%), “heart failure and shock with MCC” (n=365, 5.4%), and “respiratory infections and inflammations with MCC” (n=243, 3.6%).

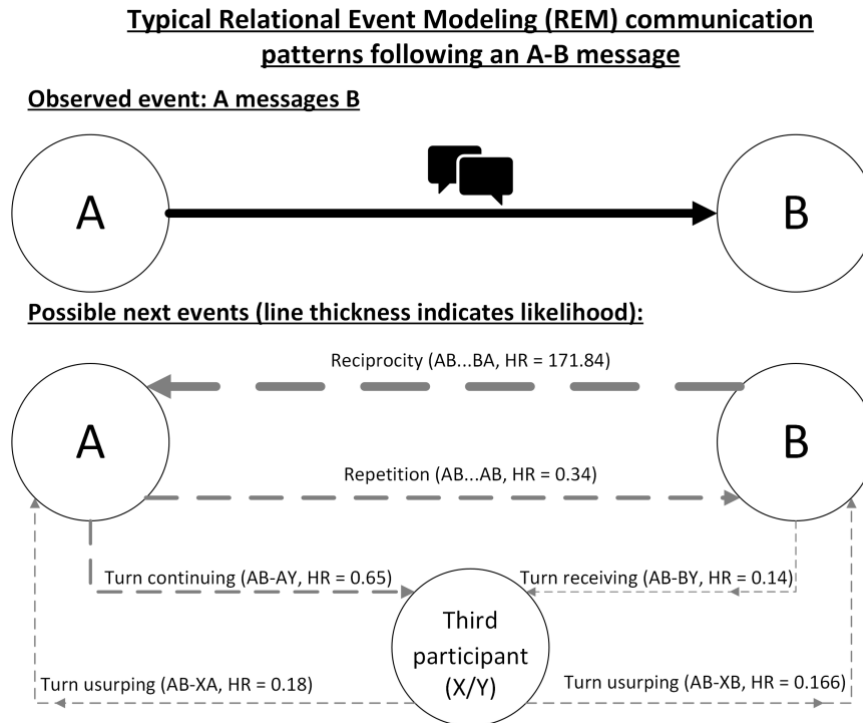


Fig. 2. Possible participation shifts after an observed A-B message (solid black line), where rate of unobserved events (dashed grey lines) is indicated by line thickness and median hazard ratios (HR).

3.2. REM model descriptive statistics

On average, patients’ care teams were highly likely to message those who had messaged them (**Figure 2**; reciprocity median HR=171.84, Q1=21.95, Q3=3664.93). A secure messaging user was 171 times more likely to respond to a person who had recently messaged them than they were to message a different person. Care teams were less likely to message those they had already messaged (repetition, median HR=0.34, Q1=0.011, Q3=1.53). Of participation shift patterns, turn continuing (AB-AY) was most likely (median HR = 0.65, Q1=<0.00001, Q3=2.04). Both turn usurping patterns were similarly likely (AB-XB median HR=0.166, Q1=<0.00001, Q3=0.98; AB-XA median HR=0.18, Q1=<0.00001, Q3=0.92). Turn receiving (AB-BY) was least likely (median HR=0.14, Q1=<0.00001, Q3=1.69).

3.3. Associations of REM model estimates with LOS

Reciprocity was robustly associated with shorter LOS and turn receiving (AB-BY) was robustly associated with longer LOS. In other words, a one-standard-deviation increase in a patient’s care team’s reciprocity was associated with a 3% shorter LOS ($\exp(-0.02894) - 1$; **Table 2, column 2**:

estimate = -0.02894, CI=-0.04805, -0.00983). A one-standard-deviation increase in turn receiving was associated with a 5% longer LOS ($\exp(0.04262)-1$; **Table 2, column 2**: estimate = 0.04262, CI=0.02320, 0.06203). Repetition, turn continuing, and turn usurping (AB-XB) were associated with longer LOS (**Table 2, column 2**) but the association was no longer significant after controlling for care team size or messaging volume (**Table 2, columns 3-4**).

Table 2. Association between secure messaging patterns and log-LOS: results of multiple stepwise model specifications

Predictor	1. Reciprocity + repetition	2. + participation shifts	3. + messaging volume	4. + care team size
REM intercept	0.04836 (0.03620, 0.06051)	-0.02526 (-0.04376, -0.00675)	-0.04949 (-0.06634, -0.03263)	-0.02507 (-0.04111, -0.00903)
Reciprocity	-0.03931 (-0.06142, -0.01719)	-0.02894 (-0.04805, -0.00983)	-0.02054 (-0.03791, -0.00318)	-0.01835 (-0.03491, -0.00179)
Repetition	0.04338 (0.02105, 0.06572)	0.04530 (0.02564, 0.06497)	-0.00024 (-0.01826, 0.01777)	-0.01294 (-0.03015, 0.00426)
Turn continuing (AB-AY)		0.02256 (0.00324, 0.04188)	0.01390 (-0.00366, 0.03145)	0.00070 (-0.01606, 0.01746)
Turn receiving (AB-BY)		0.04262 (0.02320, 0.06203)	0.01944 (0.00178, 0.03711)	0.02077 (0.00394, 0.03760)
Turn usurping (AB-XA)		0.01323 (-0.00709, 0.03356)	-0.00175 (-0.02023, 0.01672)	0.00332 (-0.01429, 0.02093)
Turn usurping (AB-XB)		0.02080 (0.00095, 0.04064)	0.01099 (-0.00704, 0.02903)	0.00806 (-0.00915, 0.02526)
Messaging volume			0.23824 (0.22607, 0.25040)	
Care team size				0.41992 (0.40289, 0.43696)

Significant findings at a significance level of 0.05 are indicated in bold in the table. Covariates that remain significant across all models are indicated by “Sig” in this legend. All models additionally adjusted for sex, race, insurance, age, Elixhauser (Sig), and ICU (Sig). Random effects were included for hospital and DRG.

3.4. Associations of REM model estimates with team EHR time

Repetition was associated with less EHR time. In other words, a one-standard deviation increase in repetition was associated with a 1% increase in EHR time per day ($\exp(0.01099) - 1$, **Table 3, column 2**: estimate = 0.01099, CI=0.00106, 0.02091). This suggests that patient care teams who were more likely to repeatedly message the same people had higher EHR time, but the association is at least partially explained by higher overall messaging volume. No other participation shifts were consistently associated with EHR time (**Table 3, columns 2-4**).

3.5. Sensitivity analysis

When including fixed effects for messaging volume and care team size, some communication patterns remain significantly associated with LOS and EHR time. Reciprocity remains significantly associated with shorter LOS when controlling for messaging volume (**Table 2, column 3**), and care team size (**Table 2, column 4**). Similarly, turn receiving remains significantly associated with

longer LOS when controlling for messaging volume (**Table 2, column 3**) and care team size (**Table 2, column 4**). Repetition's association with EHR time per day is marginally significant ($p<0.1$) when controlling for care team size (**Table 3, column 4**), but is not significant when controlling for messaging volume (**Table 3, column 3**).

Table 3. Association between secure messaging patterns and log-EHR time per day: results of multiple stepwise model specifications

Predictor	1. Reciprocity + repetition	2. + participation shifts	3. + messaging volume	4. + care team size
REM intercept	0.01553 (0.00942, 0.02163)	0.00776 (-0.00159, 0.01711)	0.00558 (-0.00378, 0.01493)	0.00775 (-0.00160, 0.01710)
Reciprocity	-0.01116 (-0.02226, -0.00006)	-0.00834 (-0.01799, 0.00130)	-0.00763 (-0.01725, 0.00200)	-0.00815 (-0.01780, 0.00150)
Repetition	0.01218 (0.00096, 0.02340)	0.01099 (0.00106, 0.02091)	0.00696 (-0.00302, 0.01695)	0.00987 (-0.00015, 0.01989)
Turn continuing (AB-AY)		-0.00079 (-0.01054, 0.00895)	-0.00156 (-0.01129, 0.00817)	-0.00122 (-0.01098, 0.00854)
Turn receiving (AB-BY)		0.00748 (-0.00230, 0.01726)	0.00536 (-0.00442, 0.01513)	0.00704 (-0.00275, 0.01683)
Turn usurping (AB-XA)		0.00275 (-0.00750, 0.01301)	0.00135 (-0.00889, 0.01159)	0.00254 (-0.00772, 0.01280)
Turn usurping (AB-XB)		-0.00047 (-0.01049, 0.00955)	-0.00128 (-0.01128, 0.00872)	-0.00070 (-0.01072, 0.00932)
Messaging volume			0.02119 (0.01445, 0.02793)	
Care team size				0.00809 (-0.00181, 0.01799)

Significant findings at a significance level of 0.05 are indicated in bold in the table. Covariates that remain significant across all models are indicated by "Sig" in this legend. All models additionally adjusted for sex, race (Sig), insurance (Sig), age (Sig), Elixhauser (Sig), and ICU (Sig). Random effects were included for hospital and DRG.

When filtering the dataset to only patients with care team size greater than or equal to three, the results are in the same direction for both outcomes, but not consistently significant. N=6,359 (94%) patients had a messaging care team of at least three. Reciprocity's association with LOS is significant for two step-wise models (model 1: estimate = 0.04403, CI=0.02142, 0.06665); model 2: 0.0342, CI=0.01486, 0.05355); it is marginally significant ($p<0.1$) when including messaging volume (model 3: -0.01035, CI=-0.02793, 0.00723) or care team size (model 4: -0.01654, CI=-0.03316, 0.00009). Similarly, turn receiving is significantly associated with longer LOS for the base model (model 2: 0.04986, CI=0.03002, 0.06970) but is not significant after controlling for messaging volume (model 3: 0.00582, CI=-0.01219, 0.02384) nor when controlling for care team size (model 4: 0.00606, CI=-0.01094, 0.02307). Repetition's association with EHR time is significant and positive (model 1: 0.01157, CI=0.00025, 0.02289) but no longer significant when controlling for participation shifts (model 2: 0.008, CI=-0.00175, 0.01776), messaging volume (model 3: 0.00522, CI=-0.00460, 0.01504), or care team size (model 4: 0.00853, CI=-0.00129, 0.01835).

4. Discussion

This study used multimodal data consisting of secure messaging metadata, EHR audit logs, and clinical data to ascertain the association between hospital team communication and both patient and team outcomes. There were four main findings: 1) secure messages were consistently responded to in inpatient medicine (high reciprocity). 2) Reciprocity was associated with shorter patient LOS, suggesting that care teams that consistently responded to messages may have more efficient patient throughput. 3) In contrast, messaging repetition was associated with more patient care-related EHR time per day (for the team), which may represent complexity of patient care or needing to “nudge” the recipient for responses. 4) Meanwhile, care teams that were more likely than average to follow turn receiving patterns were associated with longer LOS. This may suggest that teams whose message recipient involves a third person in communication represent task dependencies, or that complex information is communicated over text-based messages. Patterns of repetition and all four turn switching behaviors vary widely across patients, as demonstrated by the CIs crossing 1. These results can inform future research on the directionality of efficient team communication: studying the dynamics of reciprocity, repetition, and turn receiving in medical settings may enable more efficient patient throughput and reduce unnecessary workload. Alternatively, it may reveal that certain messaging patterns are not the cause of high LOS and high workload, but they are effects of patient care. This could enable better resource allocation if hospital systems can anticipate that certain patients will require more intensive communication and EHR workload needs.

4.1. Implications

The finding that reciprocity was the most likely communication pattern is intuitive and consistent with previous research that shows secure messages have a short time-to response (median=2.4 minutes), with most conversations including at least one response³⁶. In many communication settings, it is both required and a social norm to respond to those who communicate with you. However, it is important to note that the reciprocity hazard may overrepresent the likelihood of a response³⁷; the results should be interpreted for their direction and not for their magnitude. Nonetheless, reciprocity is the most likely pattern of secure use messaging, representing a strong propensity for secure messaging users to respond to messages they receive. In contrast, healthcare professionals were less likely to message the same person again than to message another person. This represents that secure messaging communication networks were distributed across teams for the typical patient, rather than confined to dyadic communication. Overall, team secure messaging use patterns appear to matter more for efficiency (as measured by patient LOS) than for workload (as measured by EHR time), given that multiple patterns are associated with LOS.

Further, the importance of reciprocity to LOS has potential implications for communication efficiency in clinical coordination. Higher secure messaging reciprocity was associated with shorter LOS. This suggests that teams who more consistently and quickly respond to secure messages treat patients who are discharged quicker. One possible explanation is that high reciprocity represents better visibility into action progress (through a response or confirmation of completed actions), which in turn may enable better coordination. Interestingly, this relationship was still significant in the base model (model 2) when filtering to teams of at least three users, but only marginally

significant after both filtering and controlling for team size or messaging volume. This suggests that the relationship with LOS was partially driven by patients with small care teams (only two people messaging), whose reciprocity was smaller and who had short LOS. When excluding these patients, the remaining relationship with LOS is primarily explained by messaging volume and/or by care team size, although reciprocity remains nearly significant. Reciprocal communication may contribute to more efficient clinical workflows and positively impact patient outcomes, but more work is needed on ways to differentiate messaging volume and care team size from reciprocity.

On the other hand, repetition of secure messages was associated with higher entire-team EHR time, which is a proxy for EHR workload. This may represent healthcare professionals needing to follow-up or “nudge” message recipients to gather information. Rather than being a causal relationship, it may represent that more clinically complex patients require both additional clarification through repeated messaging and require more EHR workload to read prior notes, manage comorbidities, or document care. The relationship between higher patient complexity and greater EHR workload has been demonstrated in prior studies⁸. Alternatively, if the recipient is extremely busy managing the patient’s care (with high EHR workload), a second secure message may be necessary to remind them to respond during periods of high cognitive burden. Multiple messages sent sequentially to communicate the same idea would contribute to the repetition terms in these models. Without message content, it is difficult to say with certainty whether two messages are a meaningful repetition, or if they are due to sender messaging preference. Nonetheless, the results suggest that this pattern is associated with team workload, thus the pattern may be interruptive even if repeated messages communicate the same thought.

Beyond dyadic patterns, turn receiving (AB-BY) was associated with LOS. This pattern represents a specific kind of coordination where someone who is asked a question or requested to perform an action must first get clarification from another person. This pattern is associated with longer LOS. It is possible that this is due to delayed dependencies: for example, if a nurse messages a physician to ask about a symptom a patient is experiencing, and the physician must message another physician (perhaps the patient’s primary care doctor) before they prescribe a medication, there may be delays between each interaction. These delays could prolong the patient’s overall LOS if severity of condition worsens or if required care is delayed while waiting. Alternatively, these relationships could represent secure messaging requests which require clarification, and act as a proxy for more complex patient care. However, turn receiving is no longer associated with LOS in the filtered sensitivity analysis when controlling for team size or messaging volume. Turn receiving cannot be observed in teams of two; nonetheless, these patients had higher average AB-BY values representing higher likelihood of this pattern occurring (potentially due to less training data). Patients with small care teams additionally had a lower average LOS, which contributed to the overall relationship. Although exploratory, these results may suggest team coordination in two-person care teams is substantially different from coordination in larger care teams. More work is needed to understand the nature of these differences.

Methodologically, this analysis makes possible an ethnographic study of healthcare professional communication patterns and how they relate to important patient and health system outcomes. Other studies could apply these methods to specific healthcare specialties or patient populations to understand conversational patterns in those settings.

4.2. *Limitations*

These results should be interpreted in the context of a few important limitations. First, we did not have data on the content of these secure messages. While reciprocity may indicate a “response” to a message, it cannot conclusively be determined that the reciprocal message pattern indicated a response to the same topic. Message length or topic may contribute differentially to the association with LOS or EHR workload, and natural language processing of message content would help with understanding differences in the complexity or urgency of messages. Second, this was a single site, single specialty study, limiting generalizability of results. Third, there may be residual confounding; for example, specific inpatient medicine units who work more closely face-to-face may use secure messaging less, and also treat patients who have different clinical presentations, leading to inherent differences in LOS. We attempted to control for this by limiting the setting to one specialty only, controlling for Elixhauser comorbidity, and using random effects for hospitals and DRGs; however, some unit-level, team-level, or severity-level variation may remain. Finally, some results were no longer significant in sensitivity analysis after filtering to patients with larger care teams, and results changed slightly when controlling for care team size or messaging volume. Additional analysis is needed to better understand these patients with smaller care teams and how they may contribute to the overall relationships between reciprocity, turn receiving, repetition, and both outcomes.

4.3. *Future work*

Future work could include only healthcare professionals determined to be in the patient’s core care team, as defined by a previously validated algorithm³⁸, to investigate whether the messaging patterns differed within the primary care team versus the wider team of collaborating healthcare professionals. It could alternatively assess EHR workload changes for only these individuals, or for the individuals involved in messaging. Replicating these results in other healthcare systems, settings, and specialties would reveal whether secure messaging patterns are local or shared across many clinical settings. As mentioned above, additional analysis into REM accuracy for patients who have few messages or smaller care teams would help answer an open question in REM research as to the minimum required risk set for reliable REM estimates. Finally, future work could explore interaction terms between healthcare roles and patterns of secure chat messages to investigate whether communication patterns, and their associations with outcomes, differ by professional role (e.g. nurse, physician).

4.4. *Conclusion*

This study used multimodal observational datasets to derive patterns of team digital communication. Secure messaging logs and EHR audit logs capture communication and clinical workload attributes that otherwise could not be measured at-scale. Moreover, measuring them together in a multimodal analysis enables researchers to understand how two common digital activities are related and reflect clinical behaviors. These messaging behaviors were associated with both patient (LOS) and team (EHR workload) outcomes, demonstrating the importance of measuring secure messaging use patterns at the team-level and providing insight into care coordination efficiency.

5. References

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(8):827-836. doi:10.1016/j.healthpol.2018.05.014
2. Arndt BG, Beasley JW, Watkinson MD, et al. Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. *Ann Fam Med*. 2017;15(5):419-426. doi:10.1370/afm.2121
3. Rule A, Chiang MF, Hribar MR. Using electronic health record audit logs to study clinical activity: a systematic review of aims, measures, and methods. *J Am Med Inform Assoc*. 2020;27(3):480-490. doi:10.1093/jamia/ocz196
4. Adler-Milstein J, Adelman JS, Tai-Seale M, Patel VL, Dymek C. EHR audit logs: A new goldmine for health services research? *J Biomed Inform*. 2020;101:103343. doi:10.1016/j.jbi.2019.103343
5. Kim C, Lehmann CU, Hatch D, Schildcrout JS, France DJ, Chen Y. Provider Networks in the Neonatal Intensive Care Unit Associate with Length of Stay. *IEEE Conf Collab Internet Comput IEEE Conf Collab Internet Comput*. 2019;2019:127-134. doi:10.1109/CIC48465.2019.00024
6. Rose C, Thombley R, Noshad M, et al. Team is brain: leveraging EHR audit log data for new insights into acute care processes. *J Am Med Inform Assoc*. 2023;30(1):8-15. doi:10.1093/jamia/ocac201
7. Li P, Chen B, Rhodes E, et al. Measuring Collaboration Through Concurrent Electronic Health Record Usage: Network Analysis Study. *JMIR Med Inform*. 2021;9(9):e28998. doi:10.2196/28998
8. Lou SS, Baratta LR, Lew D, Harford D, Avidan MS, Kannampallil T. Anesthesia Clinical Workload Estimated From Electronic Health Record Documentation vs Billed Relative Value Units. *JAMA Netw Open*. 2023;6(8):e2328514. doi:10.1001/jamanetworkopen.2023.28514
9. Lew D, Baratta LR, Xia L, et al. Association of EHR-Integrated Secure Messaging Use with Clinician Workload and Attention Switching. *J Gen Intern Med*. 2025;40(10):2240-2247. doi:10.1007/s11606-025-09466-x
10. Lou SS, Lew D, Harford DR, et al. Temporal Associations Between EHR-Derived Workload, Burnout, and Errors: a Prospective Cohort Study. *J Gen Intern Med*. 2022;37(9):2165-2172. doi:10.1007/s11606-022-07620-3
11. Sinha A, Stevens LA, Su F, Pageler NM, Tawfik DS. Measuring Electronic Health Record Use in the Pediatric ICU Using Audit-Logs and Screen Recordings. *Appl Clin Inform*. 2021;12(4):737-744. doi:10.1055/s-0041-1733851
12. Byrd TF, Fancher KG, Liebovitz DM, O'Leary KJ. Trends in secure mobile communication technology use among hospitalists in North America, 2016–2021. *Health Policy Technol*. 2022;11(4):100689. doi:10.1016/j.hlpt.2022.100689
13. Baratta LR, Lew D, Kannampallil T, Lou SS. Contributors to Electronic Health Record-Integrated Secure Messaging Use: A Study of Over 33,000 Health Care Professionals. *Appl Clin Inform*. 2024;15(3):612-619. doi:10.1055/s-0044-1787756
14. Gulacti U, Lok U. Comparison of secure messaging application (WhatsApp) and standard telephone usage for consultations on Length of Stay in the ED. *Appl Clin Inform*. 2017;08(03):742-753. doi:10.4338/ACI-2017-04-RA-0064

15. Epstein K, Juarez E, Epstein A, Loya K, Singer A. The impact of fragmentation of hospitalist care on length of stay. *J Hosp Med*. 2010;5(6):335-338. doi:10.1002/jhm.675
16. Turner J, Hansen L, Hinami K, et al. The impact of hospitalist discontinuity on hospital cost, readmissions, and patient satisfaction. *J Gen Intern Med*. 2014;29(7):1004-1008. doi:10.1007/s11606-013-2754-0
17. Kim MM, Barnato AE, Angus DC, Fleisher LF, Kahn JM. The Effect of Multidisciplinary Care Teams on Intensive Care Unit Mortality. *Arch Intern Med*. 2010;170(4):369-376. doi:10.1001/archinternmed.2009.521
18. Zwarenstein M, Goldman J, Reeves S. Interprofessional collaboration: effects of practice-based interventions on professional practice and healthcare outcomes. *Cochrane Database Syst Rev*. 2009;(3):CD000072. doi:10.1002/14651858.CD000072.pub2
19. Matthys E, Remmen R, Van Bogaert P. An overview of systematic reviews on the collaboration between physicians and nurses and the impact on patient outcomes: what can we learn in primary care? *BMC Fam Pract*. 2017;18:110. doi:10.1186/s12875-017-0698-x
20. The Medicare Learning Network. Skilled Nursing Facility 3-Day Rule Billing. May 2025. Accessed January 30, 2026. <https://www.cms.gov/files/document/skilled-nursing-facility-3-day-rule-billing.pdf>
21. Holmgren AJ, Apathy NC, Sinsky CA, Adler-Milstein J, Bates DW, Rotenstein L. Trends in Physician Electronic Health Record Time and Message Volume. *JAMA Intern Med*. 2025;185(4):461-463. doi:10.1001/jamainternmed.2024.8138
22. Rotenstein LS, Holmgren AJ, Healey MJ, et al. Association Between Electronic Health Record Time and Quality of Care Metrics in Primary Care. *JAMA Netw Open*. 2022;5(10):e2237086. doi:10.1001/jamanetworkopen.2022.37086
23. Lou SS, Kim S, Harford D, et al. Effect of clinician attention switching on workload and wrong-patient errors. *Br J Anaesth*. 2022;129(1):e22-e24. doi:10.1016/j.bja.2022.04.012
24. Ouyang D, Chen JH, Hom J, Chi J. Internal Medicine Resident Computer Usage: An Electronic Audit of an Inpatient Service. *JAMA Intern Med*. 2016;176(2):252-254. doi:10.1001/jamainternmed.2015.6831
25. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. *Med Care*. 1998;36(1):8-27. doi:10.1097/00005650-199801000-00004
26. Butts CT. 4. A Relational Event Framework for Social Action. *Sociol Methodol*. 2008;38(1):155-200. doi:10.1111/j.1467-9531.2008.00203.x
27. Bianchi F, Filippi-Mazzola E, Lomi A, Wit EC. Relational Event Modeling. *Annu Rev Stat Its Appl*. 2024;11(Volume 11, 2024):297-319. doi:10.1146/annurev-statistics-040722-060248
28. Leenders RThAJ, Contractor NS, DeChurch LA. Once upon a time: Understanding team processes as relational event networks. *Organ Psychol Rev*. 2016;6(1):92-115. doi:10.1177/2041386615578312
29. Pilny A, Schecter A, Poole MS, Contractor N. An illustration of the relational event model to analyze group interaction processes. *Group Dyn Theory Res Pract*. 2016;20(3):181-195. doi:10.1037/gdn0000042
30. Gibson CB, Buchler N, Hoffman B, Fleur CGL. Participation shifts explain degree distributions in a human communications network. *PLOS ONE*. 2019;14(5):e0217240. doi:10.1371/journal.pone.0217240
31. Schecter A, Quintane E. The Power, Accuracy, and Precision of the Relational Event Model. *Organ Res Methods*. 2021;24(4):802-829. doi:10.1177/1094428120963830

32. Carter T. Butts <buttsc@uci.edu>. relevent: Relational Event Models. Published online November 10, 2010;1.2-1. doi:10.32614/CRAN.package.relevent
33. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw.* 2015;67:1-48. doi:10.18637/jss.v067.i01
34. Kuznetsova A, Brockhoff PB, Christensen RHB, Jensen SP. lmerTest: Tests in Linear Mixed Effects Models. Published online March 5, 2026. Accessed July 21, 2026. <https://cran.r-project.org/web/packages/lmerTest/index.html>
35. Lüdtke D, Bartel A, Schwemmer C, Powell C, Djalovski A, Titz J. sjPlot: Data Visualization for Statistics in Social Science. Published online July 10, 2025. Accessed July 21, 2026. <https://cloud.r-project.org/web/packages/sjPlot/index.html>
36. Baratta LR, Harford D, Sinsky CA, Kannampallil T, Lou SS. Characterizing the Patterns of Electronic Health Record–Integrated Secure Messaging Use: Cross-Sectional Study. *J Med Internet Res.* 2023;25(1):e48583. doi:10.2196/48583
37. Carter T. Butts. Modeling Relational Event Dynamics with statnet. In: 2022. https://statnet.org/workshop-relevent/sunbelt_2022_relevent_tutorial.html
38. Baratta LR, Wang J, Osweiler BW, et al. Algorithmic Identification of the Daily Primary Frontline Clinician during Inpatient Medicine Encounters Using EHR Audit Log Data. *Appl Clin Inform.* 2026;17(3):540-548. doi:10.1055/a-2903-0478

Value Sensitive Design of Technology Supporting Suicide Risk Assessment and Intervention: Perspectives of Healthcare Professionals

Hannah Slater, MS; Colin G. Walsh, MD, MA; and Laurie L. Novak, PhD, MHSA
*Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Avenue
Nashville, TN 37203, USA
Email: hannah.slater@vanderbilt.edu*

Despite significant amounts of research on the widespread adoption of patient portals, research has been limited with respect to how healthcare professionals (HCPs) respond to portal messages communicating suicidality. This project interviews HCPs who respond to patient messages to learn current practice and concerns in handling suicide risk asynchronously. There are many values endorsed by HCPs in the assessment and intervention of suicidality in an asynchronous setting.

Keywords: Suicide; Patient Portals; Value Sensitive Design; Ethnography

1. Introduction

Suicide is one of the leading causes of death in the United States with rates increasing over time¹ and more than 720,000 people die worldwide from suicide². Suicide is preventable³, but prevention requires identification of risk in time to intervene.

Currently, identifying individuals at high risk for suicide has limitations with respect to individual self-reporting. Confidential resources, such as the 988 suicide and crisis line⁴, can be helpful short-term, but it makes long-term monitoring of health outcomes and progress difficult. In contrast, individuals in clinical settings do not have the perceived anonymity afforded by resources like 988 and thus individuals may be more hesitant to self-report due to self-stigma⁵⁻⁷ and/or fear of stigmatizing responses, unwanted treatment, and/or loss of privacy and autonomy^{8,9}. Newer research shows that suicidal thoughts or behaviors (STB) are dynamic phenomenon and individuals who are not currently feeling suicidal might not disclose STB at a routine visit^{8,10}, suggesting that suicide interventions may benefit from “just-in-time adaptive interventions”¹⁰.

We have reported that people who have experienced suicidal thoughts and behaviors (STB) send messages in the patient portal (PPM)¹¹, however research has been limited with respect to how people use PPMs to communicate STB¹². Clinics have a duty to identify and intervene upon messages communicating STB. To our knowledge, there is no comprehensive system that is automatically tracking and intervening upon PPM messages communicating STB¹³. Without such a system, there are people who may be reaching out for help and not receiving it in a timely manner. Little work has been conducted to know how to identify STB in asynchronous formats such as portal messages¹³. An exception is a study that suggests those who have moderate to severe symptoms are more likely to respond asynchronously when given the option of responding synchronously vs asynchronously¹⁴. These results underscore that patients may be likely to self-report STB via PPM.

The subsequent challenge to identifying these concerning messages is acting on them. Assuming correct identification, there must be a clear procedure for acting on the messages, both with the patient and within the healthcare team. Previous research has focused on alerting the provider, however outcomes were not reported¹⁴. We have only been able to identify one study reporting the perspectives of individuals who respond to messages related to suicidality. The study participants consisted of volunteer counselors in a large online support platform and found that many of them wanted more support and preparation for conversations with people experiencing STB¹⁵. It is unknown how healthcare professionals (HCPs) assess and intervene on messages communicating STB, how to best support HCPs in this assessment and intervention, and what values HCPs hold regarding the assessment and intervention of patients experiencing STB. This project reports the perspectives of HCPs who respond to patient messages. We sought to learn current practice and their concerns in handling suicide risk asynchronously. These HCPs are direct stakeholders¹⁶ in the design of any technological solution. Our goal was to understand what HCPs find most useful in assessing and intervening in an asynchronous communication environment. Understanding HCP narrative accounts can help demonstrate patterns and values invisible to metrics alone¹⁷.

2. Methods

2.1 Sampling strategy and recruitment

Interviews were conducted at an academic medical center in the mid-south of the United States. This study was approved by the Vanderbilt University Medical Center Institutional Review Board. Participants were recruited via purposive and snowball sampling to identify HCPs primarily from three outpatient clinics: primary care, neurology, and psychiatry. For recruitment we combined visiting the clinics in person and emailing flyers. The goal was to have approximately five interviews from each clinic conducted as saturation is not necessarily expected due to possible heterogeneity.

2.2 Semi-structured Interviews

The interview guide was developed through iterative process focused on what helps and hinders HCPs in responding to messages, and HCPs' desiderata for the design of a system to help with assessment and intervention. Embedded in these questions, we hoped to uncover themes and values to help us understand what is most important when HCPs are using the PPM for care related to suicidality using some known heuristics for interviewing stakeholders¹⁸. We also hope to identify the features of what might be considered a "message of concern". Two practice interviews were conducted prior to recruitment to refine questions and were not included in the analysis. Interviews were conducted virtually via Microsoft Teams¹⁹ and in-person. Participants were not compensated.

2.3 Analysis

The participants were characterized using descriptive statistics on their roles, age, training backgrounds, and clinic they are representing. Interviews were transcribed using automated Microsoft Teams transcript¹⁹ and Apple voice memo transcript²⁰ and manually validated.

We used Dedoose software²¹ employing an open coding approach to help identify themes, including barriers and facilitators²², using reflexive thematic analysis procedures²³. Transcripts were also coded using the Systems Engineering Initiative for Patient Safety (SEIPS) 2.0²⁴ and value sensitive design (VSD) frameworks¹⁸. For SEIPS, focus was primarily focused on the work system, processes, and outcome codes. For VSD, a hybrid coding approach was adopted combining codes from previous VSD literature¹⁸ and building upon their existing definitions of values and open coding for new values identified. Consensus through discussion was used for agreement on codes, but coding was primarily conducted by a single coder. After coding, analysis focused on describing the current system, facilitators and barriers in the current system, values highlighted by stakeholders, and identifying design features of tools and workflows that incorporate stakeholder values.

3. Findings

3.1 *Recruitment and participants*

Recruitment was November 2025-May 2026 with all interviews conducted January-May 2026. We interviewed 16 HCPs from five different clinics. We interviewed five HCPs from primary care, four from neurology, four from psychiatry, two from the discharge care center, and one from the trauma, burn, and surgical specialties. Most HCPs interviewed were nurses (7 RNs, 2 LPNs, 2 clinical nurse managers, 1 psychiatric-mental health nurse practitioner). There were two psychiatrists, one outpatient psychologist, and one medical social worker represented in the interviews. Median age was 36.5 years with 23-45 years old age range. Participants had been in their current roles for a median of 4 years with a 1-16-year range. Four HCPs had just an associate's degree, four had just a bachelor's degree, and eight with formal training beyond a bachelor's degree (ex. MD, PhD, MSN, MS, residency, fellowship, internship, postdoctoral training). Additionally, some of the participants had research expertise and some were co-authors of the organizational policy for response to STB.

3.2 *Current system*

3.2.1 *Actors and environment*

Participants described a complex sociotechnical system outlined in Figure 1. Patients are the primary actors in the system, initiating contact and communicating STB. Sometimes they initiate contact and then do not wish to share or answer questions. Patient avoidance of questions often elevates HCP's concern, resulting in increased intervention and sometimes diminished patient autonomy.

Nurses are the primary actors responding to the patient, via both telephone and PPM. Messages can be forwarded to the primary care physician (PCP) and/or psychiatrist for follow-up or referral.

When a patient calls the clinic via telephone, a centralized call center (access center) receives the call. The access center is primarily focused on scheduling tasks. Clinical concerns are assessed and triaged to nurses via the message basket, which can contain both access center-initiated call messages and direct PPM messages from patients. The access center marks calls involving STB or other distressing symptoms with a red exclamation mark in the message basket. Nurses reported prioritizing calls over messages. All patient-initiated messages are unmarked for acuity.

The discharge care center is a completely virtual clinic consisting of nurses, social workers, and other clinical staff. Their focus is to reduce hospital readmissions 30 days after hospital discharge. Due to rising calls related to STB since the COVID-19 pandemic, they have been tasked with responding to calls communicating STB outside of clinic hours.

Family and friends also play an important role in helping patients experiencing STB. Parents can help communicate concern for their child. Spouses can help communicate concern and remove lethal means (even if the patient does not directly communicate). If HCP reaches out and a patient does not respond, there is a potentially complex classification of family and friends to navigate including emergency contacts, people designated as “trusted adults”, and a “communication for family and others” contact, depending on what kind of information release the patient has approved.

In calls with patients exhibiting STB, the presence of children can escalate the situation through limiting what the patient will discuss and introducing the risk of imminent danger to the children. When a language translator is needed, HCPs reported a preference for professional translators over family and friends to ensure the integrity of the translation.

Mobile crisis is a behavioral safety net throughout the state. Individuals can call or text and receive telephone services or an in-person or telehealth assessment. Their counselors can visit patients where they are, but the nurses lamented the delays in this service, resulting in a resource intensive experience for the nurse, who must stay on the line with the patient.

Police are also very important actors as they can detain and transport patients. They can transport for voluntary hospitalization if the patient has no other mode of transport but are often used for involuntary hospitalization under state law. All other options are usually exhausted before this.

3.2.2 *Assessment and triage*

Messages of concern take many forms. Explicit messages, suggesting self-awareness from the patient and a desire for help, are less common. For example, *“I want to kill myself”* (P4) or *“I’m on a medicine that makes me feel this way. This has happened before. I don’t have a plan. I’ve got everything locked up.”* (P14). When people are no longer scared of the situation, it becomes particularly concerning to the HCPs. The most common type of concerning message is cryptic, conveying some sort of hopelessness, finality, or unsettling vagueness. Some examples of these are *“I’m feeling more sad than usual. I just feel like a failure”* (P4), *“I wish I could go to sleep and not wake up”* (P3), or *“I feel kind of blah. I feel unmotivated. I’ve been taking the same mood medicine for a long time. It used to seem like it worked better”* (P3). Participants reported that with cryptic messages, the HCP needs to get information from the patient to understand what is going on. The message might not even be something they say but actions they take, like hysterically crying on phone, cancelling appointments after positive biopsy results, calling HCPs with flat affect and then withholding information when HCP is inquiring (e.g., *“there’s nothing else to talk about”*). Other concerning messages are multifactorial, such as sending *“goodbye and thank you”* (P1) to multiple HCPs, sending pictures of lethal means, or subtle shifts in tone from previous messages that are difficult to detect. Patients cannot flag messages as urgent but can use subjects such as “depression”.

Somewhat related to messages of concern are threats. Sometimes they are for attention, and sometimes they are because a patient is frustrated or feels hopeless due to treatment-resistant chronic

illness. Some examples that might be threats for attention are “*if I don't get X, I'm going to hurt myself*” (P3) or “*if I don't get some relief soon, I'm going to have to do something*” (P7). To the HCPs, this is one of the reasons listening and validating the patient is so important.

Messages of concern warrant an immediate phone call from the HCP (multiple if unanswered). If the patient does not respond, the HCP must assess the risk and determine the next step. Depending on their estimation of the risk, they may message the patient back with resources, call the patient's family/friends, or call the police for a welfare check. If the patient responds, HCPs reported being more reluctant to call the police. They reported asking the patient's location and screening using the Columbia-suicide severity rating scale (C-SSRS²⁵). Depending on the score, the patient will be assigned yellow (elevated risk), orange (moderate risk), or red (high risk) which will help determine next steps for intervention. HCPs noted that while the C-SSRS is a tool that can help, it does not replace their clinical judgement. Participants agreed that high and lower risk patients are easier to identify. Patients in the moderate category can sometimes be more difficult. Above all, the HCPs make sure to document everything. This protects both themselves and the organization. Figure 1 below outlines some of the workflow for assessment and intervention described thus far.

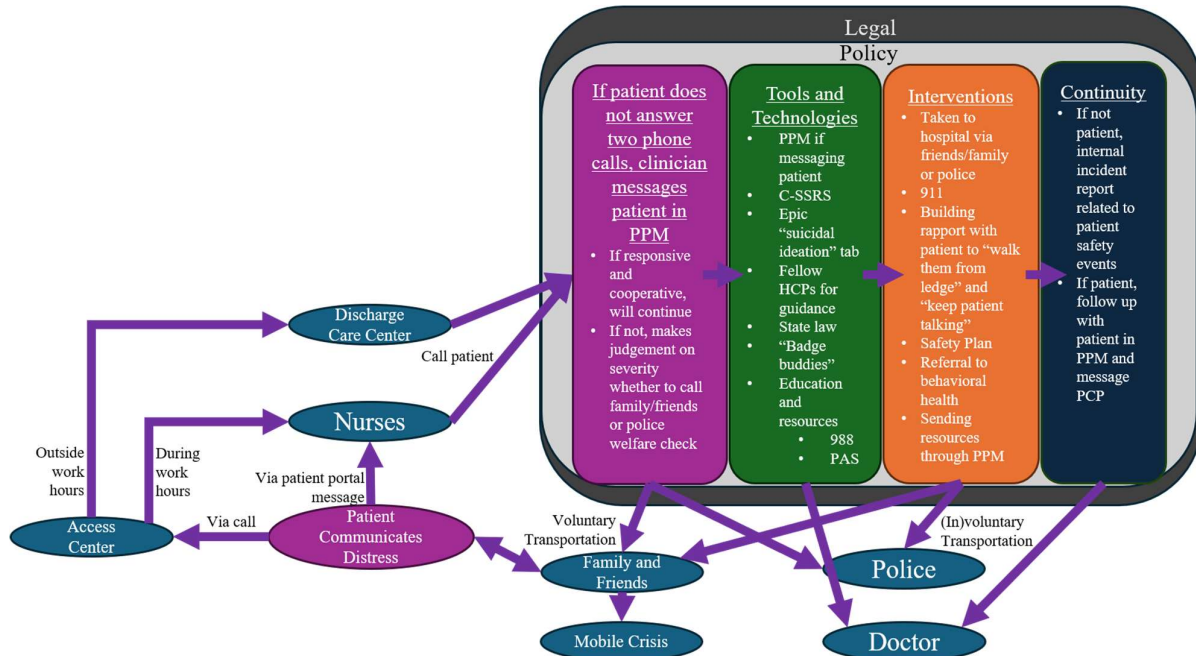


Figure 1. The actors, environment, tools, and interventions involved in management of a message or call from patient communicating distress compose a complex sociotechnical system

3.2.3 Interventions

The main purpose of interventions is patient safety. Can the patient safety plan and maintain safety if they encounter a suicidal crisis? Do they know what resources are available to them, how to access them, and feel comfortable accessing them? Figure 1 shows how interventions fit in with the system.

For yellow (elevated risk) patients, it seems the most common intervention is to let the patient know about the resources available to them such as local mental health resources, like the psychiatric

assessment service (PAS), and crisis lines (like 988). For specific issues such as postpartum depression or alcoholism, they can suggest specific resources for them such as rehabilitation programs. HCPs make a safety plan with the patient, with the most common usually being a Stanley-Brown Safety plan²⁶. They can also connect the patient with clinic resources such as their PCP or psychiatrist or a referral to behavioral health. Primary care usually becomes the catch-all if they never go to behavioral health (which can sometimes be hard to get in). If patients are already on medication, they may need to change their dose and/or they can be recommended to seek psychotherapy, although this is sometimes difficult due to availability.

For orange (moderate risk), HCPs are advised to stay with the patient. They should request help from their fellow HCPs to connect the patient to an immediate assessment by an HCP (preferably specializing in mental health). For example, they may write on a note “call 911” so they can stay on the phone with the patient. They stay on the phone with the patient until mobile crisis arrives.

For red (high risk) patients, it is a similar process, except it is recommended the patient go to the emergency room. If there is concern of imminent danger, another HCP should call local law enforcement to the patient’s physical location for an urgent welfare check and communicate that the patient may be an imminent danger to themselves. HCPs can notify their emergency contacts, and this is often when state law is utilized to detain the patient.

Another theme highlighted with each of the interventions is that of continuity and timeliness/urgency. The orange and red workflows attempted to avoid patients being left alone to figure out what to do next and with limited people knowing what was going on.

The workflows above represented clinics outside of psychiatry. For those in psychiatry, there was a very strong sense of **therapeutic alliance**²⁷, trust and rapport building, honesty, and avoiding stigma that must happen for patients to be willing to disclose STB. Making sure the patient feels heard and understood is paramount. Participants described it variously as, “*Just being near and sitting with a patient and acknowledging like that you’re on the same team and that you want to continue to help them and work with them,*” (P11) and “*Sitting with them in the fear and in their sadness*” (P8). One participant stated, “*Effectiveness means maintaining safety. I mean, that’s the most effective thing. The second most effective thing I think is candor. It’s just trust, like therapeutic alliance.*” (P1) They use the Stanley-Brown²⁶ to “*game plan suicidal crises*” (P8). Participants also focus on removing lethal means because as one stated, “*people have a preferred method of attempting suicide and if you can help them avoid that, they’re not going to be a sort of readily accessing another method*” (P8). Overall, the participants from psychiatry backgrounds communicated more comfort with uncertainty and patient autonomy than the other clinics. They recognized the limitations of their role, with one participant summarizing it this way: “*Even the best, the best clinicians have lost patients to suicide, and I think it’s prideful to think that anything I could do as a clinician would prevent all my patients from ever attempting suicide. So, I think there’s a real humility about this.*” (P8).

3.2.4 Facilitators

A frequently cited facilitator was the recently implemented organizational policy for suicidal ideation (SI) response and accompanying annual online training. For many of the HCPs (particularly

those outside of psychiatry), this policy helped them feel more comfortable helping patients who endorse STB because it gave them a clear workflow and script that they can read word for word. Some wished that the associated training was more in depth and/or hands-on, but overall, there was positive reception. Additionally, this policy appears to have been a resource for standardization, as the assessment and interventions across all the clinics were very similar. However, the policy was not mentioned by any of the participants from the psychiatry department, suggesting it might be less relevant to them because they rely more heavily on their clinical expertise and training. Overall, the workflows described by the participants were closely aligned with the published policy.

In addition to the policy, there were many tools and technologies that participants found helpful for assessment and intervention, including the C-SSRS²⁵ for assessment and the Stanley-Brown safety plan²⁶ for intervention. Both are available in the electronic health record (EHR), although many of the HCPs reported that they like to keep physical copies in a folder for ready access. Many participants used smart phrases to document SI in the note, with some customized for specific locations. During a nurse's call with a patient, a nurse triage function is available that produces the C-SSRS and other educational resources to help assist the patient. Participants described the need for a smart phrase to send patient educational resources through the PPM, and some HCPs maintain their own list of resources. Another physical resource available to the HCPs are "badge buddies," or cards that are attached to their ID badges that have the assessment and intervention workflows outlined. In the discharge care center, they have physical SI crisis packets with hard copies of educational tools and C-SSRS. Additionally, there is an intranet website that has "tip sheets" with quick and easy explanations of the policy (and other resources) as some of the HCPs complained that the policy can sometimes be too bulky and complicated to make sense of in real time. Another less mentioned tool is the ability to send the patient an automated message to check on them to see how they are doing. Nurses cannot send these, but the physicians can. Many of the HCPs highlighted how tools are helpful, but they do not replace their clinical judgement. They still need to pay attention to clues beyond what the patient is saying like their tone or withholding of information.

In addition to organizational resources, participants described facilitators for assessment learned through their experience and training. Most commonly cited was the idea of personalizing the assessment of risk by really knowing and understanding the patient. Examples included knowing what is going on in the patient's life, e.g. a recent death of family member or pet. Knowing the patient's usual disposition helped them identify changes. HCPs in psychiatry were often concerned with comorbid psychiatric diagnoses and understanding how they may interplay with patients' disclosure of STB, and if it is expected symptom of another diagnosis. If someone has chronic suicidality, and/or this is a marked change from their baseline, understanding what or why that has happened and personalizing the assessment to the patient is very important. The psychiatric HCPs highlighted the distinctions between passive vs active, chronic vs acute, and static vs variable risk factors, all of which rely heavily on personalizing the assessment of risk and understanding the nuance based on their previous training. To them, this assessment is not black and white; there are many shades of gray. Their extensive previous training and experience also seems to make them more comfortable with helping patients endorsing STB. Additionally, because the patient is not physically present, non-verbal clues such as background noise were highlighted. One HCP described how *"one of [the patient's] plans was to jump in front of a bus. Well, then I hear that he is, there's*

a car is behind him and he tells me that he's, you know, somewhere on the street and I'm like, okay, well now that's elevated the situation" (P15). For most of the participants, the goal was to focus on building rapport and trust with the patient to strengthen the therapeutic relationship. They desired to remain calm (and keep the patient calm) through deep breaths, and they sought to keep the patient talking. The HCPs also worked hard to avoid stigma by normalizing and validating the patient for being vulnerable with them and making sure the patient felt understood. For those who help train other HCPs in this, they found it is helpful for HCPs less experienced with this to think of this as any other medical diagnosis. *"It's treating it like you would any other medical condition that you're making it to where they can speak up and share and feel like they're going to be treated just the same way they would if they were experiencing heart attack or stroke symptoms because there's always been this stigma with depression and suicidality... I think it helps us provide more compassionate care when we put it in a lens of I'm asking you these questions because this is how we treat this condition."* (P14). Another HCP mentioned that some patients prefer messaging over calling due to the perceived anonymity of it. They felt like they received more open and honest answers when messaging with patients.

There were fewer facilitators mentioned for intervention than assessment, but a lot of this seems to be due to the strong focus on establishing the therapeutic alliance as the primary focus in the initial assessment. The HCPs described how they work together with the patient to figure out what they are going to do next, *"not being paternalistic"* (P12) and gaining agreement of the patient. This largely helps with patient buy-in and follow-through. For example, one of the other facilitators is helping the patient know the resources available to them, especially local resources. It is important to make sure that the patient feels comfortable accessing those resources and if not, helping them access them. For example, one patient's suicidal ideation involved using her car, so they made sure she had transportation to the next steps of her treatment plan.

3.2.5 Barriers

There are many barriers to responding to PPM messages that are not specific to suicidality. For example, there is nothing for a patient to do to establish the acuity of a message. The patient decides the subject line, but many of the HCPs go through the messages in chronological order. Another HCP or calls from the access center can mark messages as "urgent" but the patient cannot. This is one of the biggest concerns expressed by the HCPs, made worse by the fact that patients can send messages outside of work hours and concerning messages may be missed.

Overall, the non-psychiatry HCPs desired more training and/or knowledge of response to STB and the tools and technologies available to them. They also expressed a need for access to more mental health professionals to refer patients to, particularly outside the emergency room or the psychiatric hospital, as these settings carry stigma or can be far away for some patients.

There were many barriers to assessment addressed through the interviews. One of the most commonly expressed barriers is the discomfort of the subject due to the gravity of the topic. They want to respond appropriately and in a personalized way. HCPs highlighted how frameworks and tools can be helpful, but they are never going to be applicable to every situation, and one must still use their clinical judgment. Another commonly expressed barrier was the idea that the "onus is on

the patient” for disclosure. The patient takes the lead on how much they wish to disclose/not disclose. One participant put it this way: *“that person is help seeking because if they weren’t help seeking, they probably could attempt at home and never tell anyone, right?”* (P1). For most clinics, many of the standardized assessments are only conducted at patient intake so if the patient does not bring it up on their own or it is a known ongoing issue, it may not be a topic of discussion. Because of this, it is important to make patients feel safe enough to disclose, further highlighting the importance of the therapeutic alliance, building rapport and trust, and avoiding stigma with the patient. Several HCPs mentioned how often a patient will reach out and express distress, but when directly asked what is going on, they may withhold information or stop responding/hang up which raises concern by the HCPs and they feel they must “drag it out of the patient”. One participant described how all she knew about a patient was his occupation and the city he worked in. The patient got disconnected and they were able to call the police with the little information she had, and the police found the patient. The HCPs care deeply about their patients, and it is frustrating to them when they do not want to communicate what is going on so they can receive help. To this end, it is also frustrating to many of the HCPs that they do not have a lot of the usual non-verbal communication that they have in person, so there is a lot of uncertainty about the patient’s situation and they have to ask more basic questions like where the patient is and who they are with.

Stigma remains a major challenge to assessment. Overall, the HCPs expressed this seems to be getting better and they are actively working to help avoid stigma by normalizing and validating the patient, but this stigma still exists. Along with this stigma, both patients and HCPs worry about the laws related to their external environment. Some patients seem worried about disclosure due to fear of losing autonomy due to laws that allow people to be detained for involuntarily hospitalization. Additionally, HCPs are very concerned about not doing enough for a patient for fear something bad will happen to the patient and they will be held liable, so they focus a lot on “covering themselves” by documenting everything thoroughly, including when the patient refuses to answer or denies help.

In addition to barriers to assessment, there are many barriers to intervention. First and foremost, HCPs are nervous and unclear on what the risk level is of a message that has been sitting for a long time. For this reason, they call the patient as soon as possible. If the patient does not respond, it is also difficult to know how to respond. This is coupled with the fact that most HCPs outside of psychiatry do not know what the outcome was of their previous interventions, as most of the interventions are very immediate as opposed to longer-term, making it more difficult to learn.

3.3 Value sensitive design

3.3.1 Values highlighted

Value Sensitive Design (VSD)-based interviews are useful for eliciting HCP values that would be difficult to ascertain by using EHR data, video, or observations. Many of the values were elicited through dialogue. Table 1 below outlines some of the values expressed in the interviews with definition and examples. Overall, the most endorsed values were safety, collaboration (both with the patient and other parties such as HCPs, police, and family), therapeutic alliance, personalized assessment of risk, timeliness/urgency, human welfare, calmness, trust, and the desire to do better.

Table 1. Values and definition with examples of value

Value	Definition with examples
Therapeutic alliance	Building rapport and trust with the patient so they feel heard and understood; this relationship helps the patient desire to be more honest and forthcoming; establishing that the patient and the HCP be on the "same team" and will attack this together; sitting with the hurt with the patient and not being scared
Clinician care	HCPs care deeply about their patients; they wonder if they did the best they could do even once they leave work or if they followed all the company protocols
Personalized assessment of risk	Understanding how circumstances past and present affect patient's current state; understanding the patient on an individual level
Safety	Concern for the patient's immediate safety and ability to keep themselves safe
Human welfare	"Refers to people's physical, material, and psychological wellbeing" ¹⁸ ; this is more specifically related to their current physical well-being than safety
Timeliness/urgency	Doing something while it is still relevant or useful; oftentimes there is need for quick action by HCPs
Avoiding stigma	Normalizing and validating the patient; thanking them for sharing and being vulnerable
Collaboration	Not being paternalistic with the patient and working together; working together with nurses, family/friends, police, and doctors to help the patient
Trust	"Refers to expectations that exist between people who can experience good will, extend good will toward others, feel vulnerable, and experience betrayal" ¹⁸ ; this is related to establishing therapeutic alliance, but focuses more on the initial process of building rapport and maintaining relationship with genuine candor and honesty so the patient is willing to disclose STB
Informed consent	"Refers to garnering people's agreement, encompassing criteria of disclosure and comprehension (for "informed") and voluntariness, competence, and agreement (for "consent")" ¹⁸ ; HCPs will often outline explicitly what they expect the patient to do next (sometimes with a timeline) and confirm with the patient if that is acceptable to them
Accountability	"Refers to the properties that ensures that the actions of a person, people, or institution may be traced uniquely to the person, people, or institution" ¹⁸ ; each role has their own responsibilities and outcomes they are responsible for; patients are held accountable even if they stop responding or hang up
Privacy	"Refers to a claim, an entitlement, or a right of an individual to determine what information about himself or herself can be communicated to others" ¹⁸ ; patients do not have to disclose STB; limited information can be shared with family/friends
Autonomy	"Refers to people's ability to decide, plan, and act in ways that they believe will help them to achieve their goals" ¹⁸ ; patients have the right to decide how to disclose STB and if cooperative, they can collaborate to design their treatment plan
Necessary loss of autonomy	If patient is a true danger to themselves or others, they lose the agency to deny help and can be legally detained for involuntary hospitalization
Continuity	Making sure the patient has clear next steps and is not forgotten
Calmness	"Refers to a peaceful and composed psychological state" ¹⁸ ; oftentimes HCPs focus on maintaining calm both for themselves and the patient
Identity	"Refers to people's understanding of who they are over time, embracing both continuity and discontinuity over time" ¹⁸ ; HCPs often have strong identity as someone who helps people
Freedom from bias	"Refers to systematic unfairness perpetrated on individuals or groups, including pre-existing social bias, technical bias, and emergent social bias" ¹⁸ ; regardless of any other factors, all patients will be assessed and treated as needed
Universal usability	"Refers to making all people successful users of information technology" ¹⁸ ; making sure the assessment and intervention fits the particular patient
Legal protection	Making sure to document everything to protect yourself and the institution
Desire to do better	Recognizing the current system is not perfect; desiring to do best with what resources are available

3.3.2 Toward a value sensitive design technical framework

Based on the values endorsed by the HCPs, Table 2 outlines some of the design features of a system that helps with assessment and intervention of STB in PPM. Participants were skeptical that the personalized nature of risk assessment would make the identification of messages of concern difficult for an automated system. What is considered risky for one individual might not be risky for another. Additionally, there was some concern that patients might react negatively to automated responses, but overall, HCPs were more concerned about safety than the patient's immediate feelings. They recognize the patient is likely in a heightened state so treating this delicately is important. Because of this heightened state, it is also interesting to note that none of the HCPs were comfortable with a large language model (LLM) automated response. Automation was discussed in terms of a pre-scripted response, not a chatbot. Usually, these pre-scripted responses began with

establishing the therapeutic relationship by thanking them for sharing and trying to connect them to a human as soon as possible.

Table 2. Design features, their practical advantages, and the values they support

Design feature	Practical advantage(s)	Value(s) endorsed
Automated message response based on automated identification	Responded to right away	Timeliness/urgency; continuity
Automated message that must be "humanized" right away	Can receive needed care immediately by human. Allows patient to be heard and understood	Therapeutic alliance; timeliness/urgency; collaboration; continuity; personalized assessment of risk; accountability; legal protection
Having dedicated team to respond to patient messages outside of normal clinic hours	Patient can immediately connect with a person to feel cared for	Therapeutic alliance; clinician care; timeliness/urgency; collaboration; accountability; continuity; universal usability
Having automated feature that allows concerning messages to be marked as urgent for HCPs	Reduces cognitive burden on HCPs and improves timeliness	Timeliness
Content of message not saying outright that it is a message related to STB. Calling it a "concerning" message	Does not scare the patient if they were just saying it as a threat but also doesn't further escalate or heighten situation if being serious	Human welfare; calmness; avoiding stigma
Not immediately calling police or mobile crisis if STB message is received. Gives patient option if they would like to themselves	Does not scare the patient if they were just saying it as a threat but also does not diminish patient's autonomy	Personalized assessment of risk; avoiding stigma; autonomy; calmness; privacy
Content of message explicitly stating that it is an automated message, and someone might not respond to their message right now	Informs patient of situation so they can decide how to proceed	Informed consent; trust; legal protection
Content of message thanking patient for their message and being vulnerable with them. We are not scared by the content shared	Acknowledges the patient for being vulnerable with them	Therapeutic alliance; trust; avoiding stigma; calmness
Content of message giving patient easy to access resources they can access right now	Patient may be in heightened state with limited capacity to think rationally or complete complicated tasks	Personalized assessment of risk; human welfare; timeliness/urgency; collaboration; informed consent; autonomy; safety
Content of message explicitly asking patient if they are suicidal and making response simple "yes"/"no"	Sometimes messages communicating STB are threats for attention. This will distinguish if serious and the simple usability will help with likely heightened state if being serious	Freedom from bias; universal usability
Content of message making options for them to contact their PCP later that day	Makes patient feel heard and understood and that they will be taken care of	Therapeutic alliance; clinician care; timeliness/urgency; continuity; autonomy
Making it clear that response is mandatory, but giving them multiple easy options to select from	Help HCPs know what is going on and what steps the patient chose to take while still acknowledging the likely heightened state of the patient	Clinician care; autonomy; universal usability
Urgent care psychiatric evaluations	Can receive needed care immediately	Therapeutic alliance; timeliness/urgency; continuity
Information-gathering tool to help HCPs better understand the patient	Reduce cognitive burden and improve response time	Timeliness/urgency; personalized assessment of risk

4 Discussion

This study explored assessment and intervention of STB in patients over the phone and through PPM by examining perspectives on this work of HCPs at a large academic medical center. We found that there are multiple types of "messages of concern", but most common was a vague or cryptic message that demands a personalized and nuanced assessment of the patient to clarify the risk. This corroborates other studies finding that effective natural language processing tools will require more nuanced and context-specific approaches to ascertain suicide risk²⁸. HCPs often must engage the

patient, asking difficult questions, to understand the situation. This is a similar finding to another study looking at primary care experiences with PPM messages²⁹. Based on the patient's responses, there are actors in a variety of roles facilitating assistance for a patient communicating distress. Through interviews, we elicited the values held by HCPs carrying out assessment and intervention. Using these values, we developed design features of a possible framework to help HCPs in the future. Many of the values and design features described in our study are similar to those recommended by the Coalition for Health AI for crisis and severe distress handling³⁰. Another study of informatics solutions in suicide prevention, specifically a postvention tool facilitating follow-up on patients at risk for suicide, found similar design solutions³¹.

Some limitations of this work include this was a single-center study with a limited number of participants. Participants represented only a few clinics with various prevalences of STB at each clinic. Patient and caregiver's perspectives were not explored. Future studies could explore if these results validate in another external clinic system or different setting (such as volunteer crisis counselors) and explore patient and caregiver perspectives.

Future research should also explore building a value conscious model³², or specifying an ethical theory to be deliberately designed with, for assessing suicide risk from portal messages. Using classifications prior work on values in technology design, specifically in healthcare³³, we primarily identified values relevant to “professional or individual practice” (trust, autonomy, privacy) and “societal values that span many sociotechnical systems” (safety and well-being). There was minimal focus on “organizational values” (stewardship, employee well-being, fair competition). This echoes what has been reported in the scant value sensitive design literature in healthcare informatics³³. We acknowledge that values can inherently conflict with one another. For example, many of the interviewees noted how they limit their work in the messages to clinic hours for their own mental health/wellbeing, and this can conflict with their desire for the well-being and safety of a patient. Another analysis of mental health across the care continuum found that safety, privacy, and equity were important, but in cases of suicide “the clinician retains diagnostic and treatment authority at all times”³⁴. Another analysis of values important in mental health focused on safety, equity, evidence, and clinical accountability³⁵. Another analysis focused on “equity, replicability, privacy, efficacy, and engagement” and highlights these have remained relatively unchanged over the years, but with the advent of LLMs, the sense of urgency has become heightened³⁶. A study focused on youth mental health and digital technology using VSD found that “personalization, empowerment, and autonomy” to be most important³⁷. Currently, there does not seem to be clear consensus on what values should be consciously included in a mental health model based on literature, so further exploring ethical frameworks for this conscious design is necessary, especially as found through our study, there are many values endorsed by HCPs in this assessment and intervention of patients in crisis, and sometimes these can be dynamic, conflicting, and emergent. It is doubtful that current technologies can, without human oversight and involvement, participate effectively in this very human task of responding to patients in crisis while supporting the values endorsed by HCPs.

5 Funding and Acknowledgements

Funded by the National Library of Medicine 5T15LM007450. Special thank you to my family.

References

1. Suicide - National Institute of Mental Health (NIMH). n.d. <https://www.nimh.nih.gov/health/statistics/suicide> (29 July 2026, date last accessed).
2. Suicide. n.d. <https://www.who.int/news-room/fact-sheets/detail/suicide> (29 July 2026, date last accessed).
3. Schwartz-Lifshitz M, Zalsman G, Giner L *et al.* Can We Really Prevent Suicide? *Curr Psychiatry Rep* 2012;**14**(6):624–33. <https://doi.org/10.1007/s11920-012-0318-3>.
4. 988 Lifeline. 988 Lifeline. n.d. <https://988lifeline.org/> (29 July 2026, date last accessed).
5. Ammerman BA, Piccirillo ML, O’Loughlin CM *et al.* The role of suicide stigma in self-disclosure among civilian and veteran populations. *Psychiatry Res* 2022;**309**:114408. <https://doi.org/10.1016/j.psychres.2022.114408>.
6. Corrigan PW, Rao D. On the Self-Stigma of Mental Illness: Stages, Disclosure, and Strategies for Change. *Can J Psychiatry* 2012;**57**(8):464–9. <https://doi.org/10.1177/070674371205700804>.
7. Roškar S, Kralj D, Andriessen K *et al.* Anticipated Self and Public Stigma in Suicide Prevention Professionals. *Front Psychiatry* 2022;**13**. <https://doi.org/10.3389/fpsyt.2022.931245>.
8. Richards JE, Whiteside U, Ludman EJ *et al.* Understanding Why Patients May Not Report Suicidal Ideation at a Health Care Visit Prior to a Suicide Attempt: A Qualitative Study. *Psychiatr Serv* (Washington, D.C.) 2019;**70**(1):40–5. <https://doi.org/10.1176/appi.ps.201800342>.
9. Sheehan L, Oexle N, Armas SA *et al.* Benefits and risks of suicide disclosure. *Soc Sci Med* (1982) 2019;**223**:16–23. <https://doi.org/10.1016/j.socscimed.2019.01.023>.
10. Coppersmith DDL, Dempsey W, Kleiman EM *et al.* Just-in-Time Adaptive Interventions for Suicide Prevention: Promise, Challenges, and Future Directions. *Psychiatry* 2022;**85**(4):317–33. <https://doi.org/10.1080/00332747.2022.2092828>.
11. Slater H, Yin Z, Walsh CG. Characterization of Adult Patients with Autism Spectrum Disorder that Use Patient Portal. *AMIA Annu Symp Proc* 2024;**2023**:1267–76.
12. Duvall MJ, Miller NE, North F *et al.* Portal Message Language Use Prior to Suicide, Suicide Attempts, and Hospitalization for Depression. *Telemedicine and E-Health* 2022;**28**(8):1143–50. <https://doi.org/10.1089/tmj.2021.0318>.
13. Slater H, Walsh CG. Rapid Review of Models Assessing Suicide Risk from Patient Portal and Crisis Text Line Messages [Unpublished manuscript]. *AMIA Annu Symp Proc* n.d.;**2026**.
14. Sattler A, Dunn J, Albarran M *et al.* Asynchronous Versus Synchronous Screening for Depression and Suicidality in a Primary Health Care System: Quality Improvement Study. *JMIR Ment Health* 2024;**11**:e50192. <https://doi.org/10.2196/50192>.
15. Stapleton L, Liu S, Liu C *et al.* ‘If This Person is Suicidal, What Do I Do?’: Designing Computational Approaches to Help Online Volunteers Respond to Suicidality. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA) CHI ’24, 11 May 2024:1–21. <https://doi.org/10.1145/3613904.3641922>.

16. Friedman B, Kahn PHJr, Borning A. *Value Sensitive Design: Theory and Methods*, Technical Report. University of Washington: Dept. Of Computer Science and Engineering, 2002. <https://dada.cs.washington.edu/research/tr/2002/12/UW-CSE-02-12-01.pdf>.
17. Jabin MSR. Why health information technology safety problems remain invisible. *Front Digit Health* 2026;**8**. <https://doi.org/10.3389/fdgth.2026.1785141>.
18. Friedman B, Kahn PHJr, Borning. *Value Sensitive Design: Theory and Methods*. June 2003. <https://research.cs.vt.edu/ns/cs5724papers/6.theoriesofuse.cwaandvds.friedman.vsd.pdf>.
19. Video Conferencing, Meetings, Calling | Microsoft Teams. n.d. <https://www.microsoft.com/en-us/microsoft-teams/group-chat-software> (29 July 2026, date last accessed).
20. Voice Memos App - App Store. n.d. <https://apps.apple.com/us/app/voice-memos/id1069512134> (29 July 2026, date last accessed).
21. Dedoose. *Dedoose*, version 10.1.5. n.d. <https://www.dedoose.com/>.
22. Holden RJ, Carayon P. SEIPS 101 and seven simple SEIPS tools. *BMJ Qual Saf* 2021;**30**(11):901–10. <https://doi.org/10.1136/bmjqs-2020-012538>.
23. Byrne D. A worked example of Braun and Clarke’s approach to reflexive thematic analysis. *Qual Quant* 2022;**56**(3):1391–412. <https://doi.org/10.1007/s11135-021-01182-y>.
24. Holden RJ, Carayon P, Gurses AP *et al*. SEIPS 2.0: a human factors framework for studying and improving the work of healthcare professionals and patients. *Ergonomics* 2013;**56**(11):1669–86. <https://doi.org/10.1080/00140139.2013.838643>.
25. Columbia-Suicide Severity Rating Scale. n.d. <https://www.cms.gov/files/document/cssrs-screen-version-instrument.pdf>.
26. Forms. *Stanley-Brown Safety Planning Intervention* n.d. <https://suicidesafetyplan.com/forms/> (1 Aug. 2026, date last accessed).
27. Stubbe DE. The Therapeutic Alliance: The Fundamental Element of Psychotherapy. *Focus (Am Psychiatr Publ)* 2018;**16**(4):402–3. <https://doi.org/10.1176/appi.focus.20180022>.
28. Bloom PA, Treves IN, Pagliaccio D *et al*. Identifying suicide-related language in smartphone keyboard entries among high-risk adolescents. *Npj Digit Med* published online 25 June 2026. <https://doi.org/10.1038/s41746-026-02921-x>.
29. Rule A, Vang P, Micek MA *et al*. Primary care staff members’ experiences with managing electronic health record inbox messages. *J Am Med Inform Assoc* 2025;**32**(6):1040–9. <https://doi.org/10.1093/jamia/ocaf067>.
30. Best Practice Guide: Generative AI-enabled Wellness Applications. Coalition for Health AI, n.d. https://assets.ctfassets.net/7s4afyr9pmov/3VKrnpv1N8xJi7zlRaOvk0/a9b2bbe25dfd3a06dce5295e7820e1d9/Best_Practice_Guide_-_Wellness_Applications__v1.0_.pdf.
31. Burkhardt HA, Laine M, Kerbrat A *et al*. Identifying opportunities for informatics-supported suicide prevention: the case of Caring Contacts. *AMIA Annu Symp Proc* 2023;**2022**:309–18.
32. Manders-Huits N. What Values in Design? The Challenge of Incorporating Moral Values into Design. *Sci Eng Ethics* 2011;**17**(2):271–87. <https://doi.org/10.1007/s11948-010-9198-2>.
33. Long Y, Novak L, Walsh CG. Searching for Value Sensitive Design in Applied Health AI: A Narrative Review. *Yearb Med Inform* 2025;**33**(1):75–82. <https://doi.org/10.1055/s-0044-1800723>.

34. Vowels M, Capps H, Erickson-Ridout K. *Mental Health Across the Care Continuum: A Review and Technology-Powered Reference Standard*. (<https://osf.io/>) n.d.
https://www.researchgate.net/publication/405414993_Mental_Health_Across_the_Care_Continuum_A_Review_and_Technology-Powered_Reference_Standard (3 June 2026, date last accessed).
35. Linardon J, Firth J, Carvalho AF *et al*. Multidisciplinary research priorities for artificial intelligence in mental health: a call to action. *The Lancet Psychiatry* 2026;**0**(0).
[https://doi.org/10.1016/S2215-0366\(26\)00127-6](https://doi.org/10.1016/S2215-0366(26)00127-6).
36. Birk MV, Kochhar S, Myrick K *et al*. Digital Mental Health Research Priorities, Revisited for the AI and Large Language Model Era. *JMIR Mental Health* 2026;**13**(1):e104118.
<https://doi.org/10.2196/104118>.
37. Poulsen A, Hickie IB, Chong MK *et al*. Personalised care, youth mental health, and digital technology: A value sensitive design perspective and framework. *Ethics Inf Technol* 2025;**27**(4):61. <https://doi.org/10.1007/s10676-025-09866-x>.

Adverse Drug Events Across Data-Production Contexts: Multilingual Detection, Alignment, and Cross-Genre Discourse Analysis

Yujun Ma

*Department of Computational Biomedicine, Cedars-Sinai Health Sciences University,
Los Angeles, CA, USA
E-mail: Yujun.Ma@cshs.org*

Davy Weissenbacher

*Department of Computational Biomedicine, Cedars-Sinai Health Sciences University,
Los Angeles, CA, USA
E-mail: Davy.Weissenbacher@cshs.org*

Jake Patock

*Department of Computational Biomedicine, Cedars-Sinai Health Sciences University,
Los Angeles, CA, USA
E-mail: Jake.Patock@cshs.org*

Graciela Gonzalez Hernandez[†]

*Department of Computational Biomedicine, Cedars-Sinai Health Sciences University,
Los Angeles, CA, USA
E-mail: Graciela.GonzalezHernandez@csmc.edu*

Adverse drug event (ADE) evidence is produced across patient-generated, clinical, and scientific settings that differ in language, documentation purpose, terminology, and degree of standardization. These differences shape both which adverse experiences become visible to pharmacovigilance systems and how readily they can be linked to curated drug-safety knowledge. We examine these relationships across six corpora representing distinct data-production settings: ADE Corpus V2 (medical case reports), SMM4H-2026 Task 1 (multilingual user-generated health content), CADEC V2 (patient-forum narratives), the Dutch ADE Corpus and KronosRX (EHR clinical notes), TwiMed-PubMed (biomedical literature). A shared BERTopic analysis of ADE-positive texts concerning antidepressants and antihypertensives across the four English-language corpora identified nine interpretable topics. CADEC V2 contained a more differentiated distribution of symptom-specific themes, including sexual effects, suicidal or panic-related thoughts, vivid dreams, and memory difficulties, whereas SMM4H-2026, TwiMed-PubMed, and ADE Corpus V2 were dominated by a broader medication, sleep, tiredness, and pain theme. These patterns indicate that data-production context shapes what adverse experiences are expressed and standardized, with patient-generated narratives surfacing subjective, symptom-specific experience largely absent from clinical and scientific sources.

We further examine how data-production context affects alignment with curated pharmacovigilance knowledge using SIDER 4.1. Retrieval coverage varied sharply across sources, from 9.5% in CADEC V2 to 91.0% in TwiMed-PubMed, while KronosRX showed high task-specific lexical coverage (91.7% cardiotoxicity; 81.7% neurotoxicity).

We compare corpus-specific QLoRA fine-tuning of Llama-3.2-3B with SIDER-grounded

inference using Llama-3.1-70B/405B. Among the evaluated configurations, QLoRA obtained the highest observed macro-F1 on ADE Corpus V2 (0.91), CADEC V2 (0.88), and SMM4H-2026 (0.80), whereas Llama-3.1-405B RAG-SIDER obtained the highest observed macro-F1 on Dutch ADE (0.95) and TwiMed-PubMed (0.91). In KronosRX, toxicity-specific RAG improved macro-F1 from 0.54 to 0.75 for cardiotoxicity and from 0.58 to 0.64 for neurotoxicity. Together, these findings suggest that ADE representation, knowledge-base alignment, and detection performance vary substantially across data-production contexts, supporting the use of source-sensitive supervision and knowledge grounding.

Keywords: Adverse drug event detection; pharmacovigilance; large language models; retrieval-augmented generation.

1. Introduction

Adverse drug events (ADEs) are harmful or unintended outcomes associated with medication use, including adverse reactions, medication errors, and other drug-related injuries.^{1,2} They remain a major patient-safety concern and are associated with increased mortality, prolonged hospitalization, elevated healthcare costs, disability, and death.^{3,4} Although clinical trials and spontaneous reporting systems provide important safety evidence, many ADEs are underreported or documented only indirectly in electronic health records (EHRs), patient-generated posts, medical forums, case reports, and biomedical literature.^{5,6} Automated methods can support earlier identification of such signals, though their effectiveness depends on how well they account for the representational differences across these sources.^{7–9}

These sources are not interchangeable representations of the same pharmacovigilance evidence. They are produced for different audiences and purposes and vary in terminology, language, documentation practices, and degree of standardization. Clinical and scientific sources may explicitly describe drug–outcome relations using specialized terminology, whereas social-media posts and patient-forum narratives often contain misspellings, slang, multilingual expressions, implicit symptom descriptions, and extended accounts of lived experience.¹⁰ In this study, we use the term *data-production context* to capture these differences. Through cross-source topic modeling, we examine how such contexts shape which adverse experiences are expressed and how they are framed, and how readily the resulting drug mentions can be linked to standardized pharmacovigilance knowledge. We complement this analysis with ADE detection experiments, using them as a lens to further probe how source-specific supervision, pretrained model knowledge, and external evidence interact with these representational differences, rather than as an end in themselves.

Prior discourse-oriented work has largely examined patient narratives within a single platform or population,^{11,12} with relatively little attention to how ADE expression and framing compare across patient-generated, clinical, and scientific sources within a shared analytic frame. A related open question is the form of knowledge on which downstream detection systems rely.¹³ Fine-tuning learns from corpus-specific labeled examples, whereas retrieval-

augmented generation (RAG) introduces external evidence at inference time. In this study, retrieval is grounded in SIDER, a curated resource containing standardized drug–adverse-effect associations derived primarily from drug labels and clinical-trial information.¹⁴ Although such knowledge can support ADE detection, SIDER’s predominantly English, standardized vocabulary may align unevenly with brand names, misspellings, colloquial forms, and language-specific drug mentions found in real-world text. Prior work has shown that SIDER-grounded RAG can improve drug–side-effect question answering on a single benchmark;¹⁵ however, its coverage across heterogeneous data-production contexts remains unclear. Recent LLM-based ADE detection work using zero- and few-shot prompting, parameter-efficient fine-tuning, and structured knowledge construction^{16–18} has likewise been evaluated largely on a single corpus or a small number of closely related sources,^{19,20} with limited attention to how data-production context shapes both ADE expression and model performance.

Taken together, these gaps motivate three research questions: (1) how ADE discourse differs across patient-generated, clinical, and scientific sources; (2) how well drug mentions from these settings can be linked to SIDER; and (3) how corpus-specific fine-tuning, prompting, and SIDER-grounded inference perform across different data-production contexts, as a means of probing the downstream consequences of (1) and (2).

Figure 1 summarizes the study design. As shown in Panel 1, we analyze six corpora representing distinct data-production contexts: ADE Corpus V2 medical case reports,²¹ SMM4H-2026 Task 1 multilingual user-generated health content,²² CADEC V2 patient-forum narratives,²³ the KronosRX and Dutch ADE Corpus EHR clinical notes,²⁴ and TwiMed-PubMed biomedical literature.²⁵ Panel 2 presents the shared BERTopic analysis²⁶ of ADE-positive English-language texts concerning antidepressants and antihypertensives from four corpora, which we use to examine how adverse experiences differ across data-production contexts and how they relate to SIDER coverage. Panel 3 presents the ADE detection approaches, including corpus-specific QLoRA fine-tuning of Llama-3.2-3B²⁷ and SIDER-grounded inference with Llama-3.1-70B and Llama-3.1-405B, alongside zero- and few-shot prompting baselines, which together let us assess how readily real-world drug mentions from each corpus can be linked to the standardized external knowledge resource and how this alignment relates to detection outcomes.

Our results reveal corpus-dependent differences in ADE discourse, knowledge-base alignment, and detection performance. The topic analysis shows that patient-generated, clinical, and scientific sources differ in the adverse experiences they foreground and in the level of thematic specificity with which those experiences are described. The retrieval analysis reveals uneven alignment between real-world drug mentions and SIDER’s standardized vocabulary, with coverage ranging from 9.5% in patient-forum narratives to 91.0% in biomedical literature. Detection performance also varies by corpus, with no single configuration dominating across all six settings (Section 4.1). Together, these findings demonstrate that data-production context shapes what ADE evidence is expressed, how it is framed and standardized, and how readily it can be retrieved and computationally detected.

2. Background and Related Work

This section reviews two areas relevant to this study: variation in pharmacovigilance evidence across data-production contexts, and computational approaches—detection models and knowledge-grounded retrieval—that depend on corpus-specific supervision or structured pharmacovigilance knowledge.

ADE Evidence Across Data-Production Contexts. Pharmacovigilance increasingly draws on evidence from patient-generated content, electronic health records (EHRs), medical case reports, biomedical literature, and curated drug-safety resources. These sources are complementary but not interchangeable. Patient forums and social media platforms provide direct accounts of medication use and adverse experiences, often expressed in colloquial, subjective, and multilingual language. Comparative reviews suggest that patient-generated sources can supplement spontaneous reports, product labels, and clinical literature by contributing timely or unexpected adverse-event evidence and information about how adverse experiences affect patients’ daily lives.⁹ At the same time, such data exhibit substantial linguistic variability, incomplete context, and platform-specific biases. Clinical notes contain longitudinal observations and documentation of suspected ADEs, but relevant evidence is often expressed indirectly in unstructured text and varies across institutions and documentation practices.⁹ Medical case reports and biomedical publications generally use more formal terminology and more explicit drug–outcome relations, but they are also shaped by clinical selection and scientific reporting practices.^{21,25} Consequently, each data-production context makes different aspects of drug-related harm visible and available for computational analysis.

Computational content analysis has been used to study patient discussions of health experiences and medication use. Wexler *et al.* identified recurring concerns in online patient forums,¹¹ while Chen *et al.* showed that combining adverse-effect extraction, signal detection, and topic modeling can reveal contextual information beyond event extraction alone.¹² BERTopic further enables semantically related documents to be organized in a shared topic space.²⁶ However, prior work has largely focused on single platforms, populations, or medications, with limited cross-source analysis of ADE discourse and its relationship to pharmacovigilance knowledge alignment and detection performance.

Computational Approaches to Detection and Knowledge Grounding. ADE detection has progressed from conventional classifiers such as support vector machines^{28,29} to transformer-based models, including BERT,³⁰ Bio-ClinicalBERT,³¹ and RoBERTa,³² and more recently to LLMs adapted through prompting, fine-tuning, and data augmentation.^{10,17,33} Domain-adapted models often perform well when trained and evaluated within a single corpus,^{34,35} but Dai *et al.* showed through the MultiADE benchmark that models trained on one ADE domain did not generalize reliably to unseen text types,²³ underscoring the influence of source language, annotation practices, and task formulation. Most LLM-based studies, however, remain limited to one or two source types and rarely examine source-specific supervision and structured external knowledge within the same cross-source analysis—a gap we address by evaluating prompting, QLoRA fine-tuning, and SIDER-grounded inference across six heterogeneous corpora.

Two adaptation strategies dominate this space and rely on different sources of pharma-

covigilance evidence. Parameter-efficient fine-tuning methods such as LoRA and QLoRA freeze most pretrained parameters and update only low-rank weights, enabling corpus-specific adaptation with reduced memory requirements.³⁶ Prior work shows that QLoRA can approach full fine-tuning performance in clinical summarization and bilingual primary-care notes while using far fewer trainable parameters.^{27,36} RAG, by contrast, grounds inference in external evidence without updating model parameters, making performance dependent on the coverage and terminology of the retrieved resource. Examples include MALADE, which uses drug labels, biomedical literature, and FDA resources for drug–outcome extraction;³⁷ SIDER-based dense retrieval for drug–side-effect queries;¹⁵ and AskFDALabel, which retrieves FDA labeling for adverse-event annotation.¹³ These approaches demonstrate the value of external pharmacovigilance knowledge, but primarily target querying or structured association extraction rather than ADE detection across heterogeneous text sources. We therefore compare local supervision and external knowledge grounding across multiple data-production contexts, with particular attention to how SIDER coverage relates to detection performance.

3. Methods

We combine cross-source discourse analysis with ADE detection experiments to examine how data-production context shapes both the expression of adverse drug events and the performance of computational detection methods. Figure 1 summarizes the overall study design. Panel 1 introduces the six ADE corpora, which represent distinct languages and data-production settings, and shows their preprocessing into a common binary ADE detection task. Panel 2 presents a cross-source analysis of how adverse drug experiences are discussed in patient-generated text, medical case reports, clinical notes, and biomedical literature. Panel 3 evaluates corpus-specific QLoRA fine-tuning, zero- and few-shot prompting, and SIDER-grounded retrieval-augmented inference for ADE detection, together with instance-level SIDER retrieval coverage as a measure of how well real-world drug mentions align with standardized pharmacovigilance knowledge.

3.1. Data Sources and Task Construction

We use six expert-annotated ADE corpora spanning patient-generated, clinical, and scientific data-production contexts. Because the corpora differ in annotation schemes and units of analysis (sentence-, document-, and relation-level), we map their outputs to a shared binary label space (ADE, Not-ADE) while retaining corpus-specific operational criteria. Performance is therefore evaluated within each corpus.

ADE Corpus V2 comprises medical case reports with sentence-level ADE annotations (17,637 train, 5,879 test). Following reports of cross-split duplication,¹⁹ we normalize case and whitespace and remove exact train–test duplicates.

SMM4H-2026 Task 1 addresses document-level ADE detection in multilingual user-generated health content (German, French, Russian, English, Mandarin, Japanese), using predefined training, validation, and test splits with normalized formatting.

CADEC V2 extends the AskAPatient corpus with drug and adverse-reaction annotations. We convert *brat* annotations to sentence-level binary labels and construct document-level splits to prevent sentences from the same review appearing across partitions.

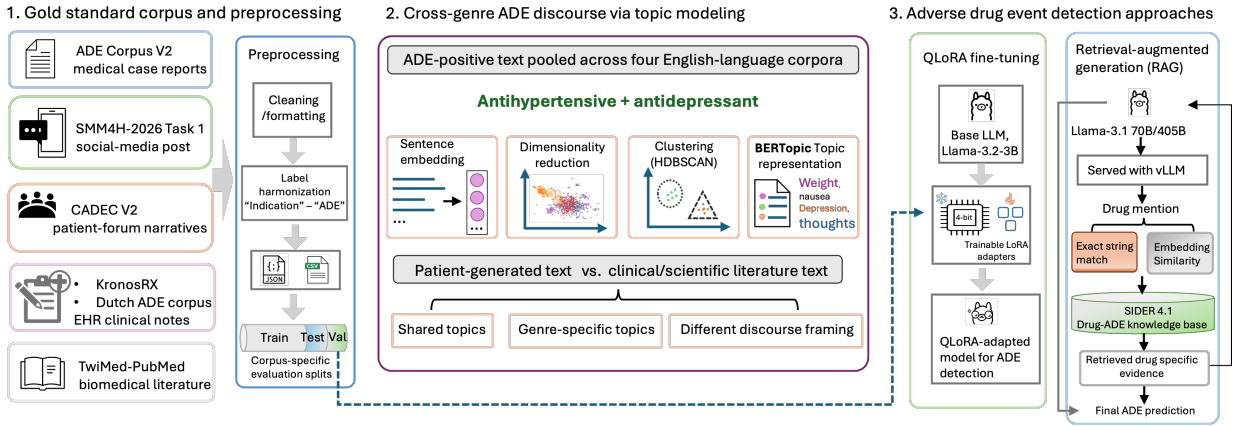


Fig. 1: Overview of the study design. Six ADE corpora spanning distinct data-production settings are standardized for binary ADE detection. We analyze cross-source discourse with BERTopic and compare source-specific fine-tuning with SIDER-grounded retrieval-augmented inference across documentation settings.

Dutch ADE Corpus consists of ICU clinical notes annotated with drug–disorder relations. We map adverse drug events to **ADE** and indications to **Not-ADE**, and split the 102 notes at the note level (70/15/15) to prevent leakage.

Twimed contains PubMed sentences and Twitter posts annotated for drugs and outcomes; we use the PubMed subset only. Sentences containing at least one outcome-negative relation are labeled **ADE**, and we construct a stratified 80/20 train–test split.

KronosRX is a non-public, physician-adjudicated EHR benchmark derived from Cedars-Sinai clinical data, with cohorts for medication-related cardiotoxicity and seizure-related neurotoxicity. Model predictions are made at the note level and aggregated to encounters using an any-positive rule. The cardiotoxicity cohort contains 583 encounters from 533 patients, split into 404/92/87 train/development/test encounters; the test set contains 22 ADE-positive and 65 Not-ADE encounters. The neurotoxicity cohort contains 88 encounters from 80 patients (12 ADE-positive, 76 Not-ADE) and is evaluated as a single external cohort.

3.2. Cross-Genre ADE Discourse via Topic Modelling

To examine how ADE discourse varies across data-production contexts, we focus on antihypertensives and antidepressants, two widely used drug classes associated with diverse adverse effects and represented across multiple corpora. The analysis includes ADE-positive English-language texts from ADE Corpus V2, CADEC V2, Twimed-PubMed, and the English subset of SMM4H-2026 Task 1.

Drug mentions are obtained from gold-standard entity annotations for CADEC V2 and Twimed-PubMed. Because ADE Corpus V2 and SMM4H-2026 do not provide complete drug-entity annotations for the classification instances, mentions are extracted using Llama-3.1-405B and verified against the original text. The Dutch ADE Corpus is excluded from this analysis because its Dutch-language texts are not directly comparable within the shared English-language topic space and provide limited representation of the selected drug classes; it remains included in the SIDER-coverage and ADE-detection analyses.

ADE-positive instances mentioning either drug class are pooled across the four corpora. To reduce clustering by medication identity, extracted drug mentions are replaced with a generic placeholder before embedding. A single BERTopic model²⁶ is fitted to the combined drug-masked texts, using sentence embeddings, UMAP, HDBSCAN, and class-based TF-IDF to derive topics in a shared space. Topics are labeled from their highest-weighted terms and representative sentences. Topic prevalence is calculated within each corpus as the proportion of qualifying instances assigned to each topic, and prevalence, terms, and representative examples are compared across sources to identify shared and source-specific themes. Because corpus sizes differ substantially, prevalence is interpreted comparatively rather than as a population-frequency estimate.

3.3. *Adverse Drug Event Detection Approaches*

We evaluate three ADE detection settings: zero- and few-shot prompting, corpus-specific QLoRA fine-tuning, and SIDER-grounded retrieval-augmented inference. All configurations use corpus-specific instructions reflecting the unit of analysis and operational ADE criteria.

Zero- and Few-Shot Inference. Llama-3.2-3B-Instruct, Llama-3.3-70B-Instruct, and Llama-3.1-405B-Instruct are evaluated under zero-shot prompting. Each prompt includes the corpus-specific task definition, the input (sentence, document, or drug-disorder context), and an instruction to return one of two labels: **ADE** or **Not-ADE**. Llama-3.3-70B-Instruct and Llama-3.1-405B-Instruct are additionally evaluated under few-shot prompting with labeled demonstrations. These baselines do not include retrieved SIDER evidence.

QLoRA Fine-Tuning. We fine-tune a separate Llama-3.2-3B-Instruct model for each corpus using Quantized Low-Rank Adaptation (QLoRA).³⁸ The base model is loaded with 4-bit NormalFloat quantization (double quantization), with computation in `bfloat16` precision. Following standard LoRA parameterization, low-rank adapters (rank $r = 16$, scaling $\alpha = 32$, dropout 0.05) are applied to all linear layers; base-model parameters remain frozen and only adapter weights are updated. Each training instance is formatted as a corpus-specific instruction-following exchange using the Llama chat template, where the system message defines the task, the user message contains the input, and the assistant message provides the gold label (**ADE** or **Not-ADE**). Models are trained with a causal language-modeling objective using a paged 8-bit AdamW optimizer, cosine learning-rate schedule, and gradient checkpointing.³⁸ At inference, the assistant response is omitted, greedy decoding is applied, and outputs are mapped to the binary label space.

SIDER-Grounded Retrieval-Augmented Inference . Our retrieval-augmented configuration introduces external pharmacovigilance evidence without task-specific parameter updates. Llama-3.3-70B-Instruct and Llama-3.1-405B-Instruct are served with vLLM and remain frozen during inference. Each input is augmented with evidence from SIDER 4.1, a curated resource containing drug-adverse-effect associations derived primarily from drug labels and clinical-trial information, with adverse effects standardized using MedDRA terminology.¹⁴

We constructed a knowledgebase mapping canonical drug names from SIDER and synonyms to lists of associated adverse effects (restricted to MedDRA Preferred Terms), retaining up to ten effects per entry to limit prompt length. Very short drug names are excluded from the lexical index to reduce ambiguity.

Retrieval combines exact lexical matching with embedding-based semantic retrieval and is adapted to the language and annotation structure of each corpus. For English-language inputs, lexical matching is applied to extracted drug mentions $M(x)$, while semantic retrieval uses the full input text as a query. The two evidence sources are merged, prioritizing exact matches when duplicates occur. Semantic retrieval uses BAAI/bge-large-en-v1.5 for English inputs and paraphrase-multilingual-MiniLM-L12-v2 for cross-lingual inputs, retaining the top $k = 3$ entries above corpus-specific similarity thresholds ($\tau_{\text{en}} = 0.80$, $\tau_{\text{nl}} = 0.62$).

For the Dutch ADE Corpus, retrieval follows a sequential strategy: the target drug entity is first normalized using rule-based transformations to approximate English forms (e.g., `-ine` $\rightarrow$ `-in`, `-ole` $\rightarrow$ `-ol`), and lexical matching is attempted before falling back to cross-lingual semantic retrieval if needed. The SMM4H corpus uses the English retrieval branch for English-language instances and the cross-lingual branch otherwise.

Retrieved evidence is inserted into a corpus-specific prompt containing the original input and any available drug or disorder. The model is instructed to use the retrieved information as supporting evidence and to generate **ADE** or **Not-ADE**. Inference uses deterministic decoding (temperature 0), and outputs are normalized and mapped to the binary label space. We evaluate both exact-match-only and hybrid lexical-semantic retrieval settings where applicable.

4. Results

SIDER Coverage Across Data-Production Contexts. Table 1 summarizes SIDER retrieval coverage across the corpora. For the original five datasets, exact coverage denotes lexical matching to a SIDER drug name or synonym, while semantic-only coverage denotes embedding-based retrieval when lexical matching fails. KronosRX uses task-specific lexical retrieval only, with coverage measured at the note level.

Coverage varied substantially across data-production settings. TwiMed-PubMed showed the highest coverage among the original corpora (91.0%), consistent with the standardized medication terminology common in biomedical literature. KronosRX also showed high lexical

Algorithm 1 SIDER-grounded retrieval-augmented inference for ADE detection

Require: Input instance x ; extracted drug mentions $M(x)$; target drug m and disorder a when explicitly annotated; corpus and language identifier g ; SIDER knowledge base $\mathcal{D}$; precomputed SIDER embeddings $\mathcal{V}_g$; retrieval size k ; similarity threshold τ_g

Ensure: Predicted label $\hat{y} \in \{\text{ADE}, \text{Not-ADE}\}$

```

1:  $\mathcal{E} \leftarrow \emptyset$ 
2: if  $g = \text{DUTCHADE}$  then
3:    $\mathcal{C} \leftarrow \text{NORMALIZEDUTCHDRUG}(m)$ 
4:    $\mathcal{E} \leftarrow \text{LEXICALMATCH}(\mathcal{C}, \mathcal{D})$ 
5:   if  $\mathcal{E} = \emptyset$  then
6:      $\mathbf{q} \leftarrow f_g(m)$ 
7:      $\mathcal{E} \leftarrow \text{SEMANTICRETRIEVE}(\mathbf{q}, \mathcal{D}, \mathcal{V}_g, k, \tau_g)$ 
8:   end if
9: else
10:   $\mathcal{E}_{\text{lex}} \leftarrow \text{EXACTMATCH}(M(x), \mathcal{D})$ 
11:   $\mathbf{q} \leftarrow f_g(x)$ 
12:   $\mathcal{E}_{\text{sem}} \leftarrow \text{SEMANTICRETRIEVE}(\mathbf{q}, \mathcal{D}, \mathcal{V}_g, k, \tau_g)$ 
13:   $\mathcal{E} \leftarrow \text{MERGEXACTFIRST}(\mathcal{E}_{\text{lex}}, \mathcal{E}_{\text{sem}})$ 
14: end if
15:  $p \leftarrow \text{BUILDCORPUSSPECIFICPROMPT}(x, M(x), m, a, \mathcal{E}, g)$ 
16:  $\hat{y}_{\text{raw}} \leftarrow \text{LLM}(p, \text{temperature} = 0)$ 
17:  $\hat{y} \leftarrow \text{PARSELABEL}(\hat{y}_{\text{raw}})$ 
18: return  $\hat{y}$ 

```

Fig. 2: Extracted drug mentions are matched lexically and semantically against a SIDER-based knowledge base, with a sequential lexical-first branch for Dutch ADE notes and a parallel lexical+semantic branch for English and multilingual corpora. Retrieved evidence is inserted into a corpus-specific prompt, and a frozen Llama model predicts a binary ADE/Not-ADE label under deterministic decoding.

Table 1: SIDER retrieval coverage across corpora. Semantic-only denotes embedding-based retrieval after lexical matching fails; KronosRX uses task-specific lexical retrieval only.

Corpus	Exact (%)	Semantic-only (%)	Total (%)
ADE Corpus V2	52.0	0.0	52.0
SMM4H-2026	4.7	58.3	63.0
Dutch ADE	28.2	2.5	30.6
CADEC V2	9.5	0.0	9.5
Twimed-PubMed	91.0	0.0	91.0
KronosRX-Cardio	91.7	N/A	91.7
KronosRX-Neuro	81.7	N/A	81.7

coverage, reaching 91.7% for cardiotoxicity and 81.7% for neurotoxicity. In contrast, SMM4H-2026 had very low exact coverage (4.7%), but semantic retrieval increased total coverage to 63.0%, highlighting the value of embedding-based retrieval for multilingual and linguistically variable user-generated content.

ADE Corpus V2 and Dutch ADE showed intermediate coverage (52.0% and 30.6%, respectively), whereas CADEC V2 remained lowest at 9.5%. The low coverage in CADEC reflects mismatches such as brand names, spelling variation, abbreviations, and colloquial drug references. Overall, these patterns indicate that the usefulness of curated pharmacovigilance resources depends strongly on the terminology and documentation practices of the source, which helps explain why retrieval-based configurations are more competitive in some corpora than others.

Cross-Genre Patterns in ADE Discourse. Figures 3 and 4 summarize the pooled BERTopic analysis of 6,713 drug-masked, ADE-positive instances from the four English-language corpora, comprising 6,001 instances from CADEC V2, 466 from SMM4H-2026, 181 from ADE Corpus V2, and 65 from Twimed-PubMed. The model identifies nine interpretable topics in addition to an outlier cluster. Sexual side effects, neurological symptoms, and swelling or localized pain appear across multiple corpora, whereas suicidal or panic-related thoughts, night sweats, vivid dreams, and memory difficulties are concentrated in CADEC V2. These patterns suggest that long-form patient narratives provide greater visibility into emotionally salient and highly specific experiences that are rarely represented in shorter or more formally structured sources.

Medication/Sleep/Tired is the largest topic in the pooled corpus and dominates SMM4H-2026, Twimed-PubMed, and ADE Corpus V2, accounting for 50.4%, 66.2%, and 66.9% of their qualifying instances, respectively. In CADEC V2, *Weight/Nausea/Loss* (18.7%) and *Medication/Sleep/Tired* (18.2%) are similarly prevalent, with the remaining instances distributed across a broader range of symptom-specific topics and outliers. Because CADEC V2 contributes most of the pooled instances, prevalence estimates for ADE Corpus V2 and Twimed-PubMed should be interpreted descriptively rather than as precise frequency estimates. Consistent with this pattern, Figure 4 shows that topics with more distinctive vocabularies form relatively compact clusters, whereas the two largest and more general topics occupy more diffuse regions of the embedding space.

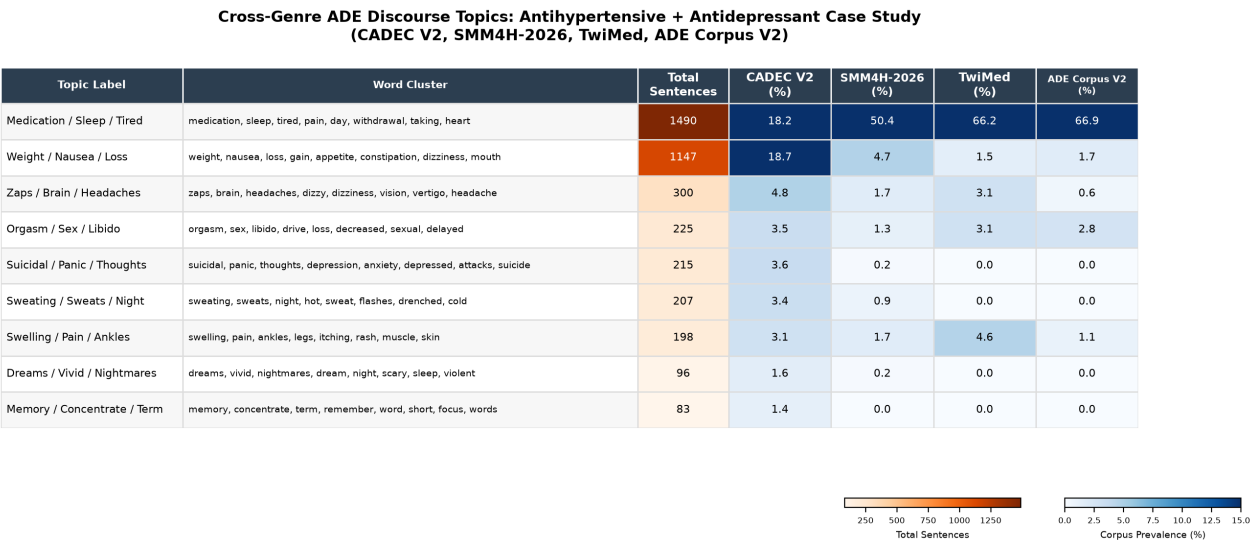


Fig. 3: Cross-genre ADE discourse topics identified by BERTopic (antihypertensive + antidepressant case study, CADEC V2, SMM4H-2026, TwiMed, ADE Corpus V2).

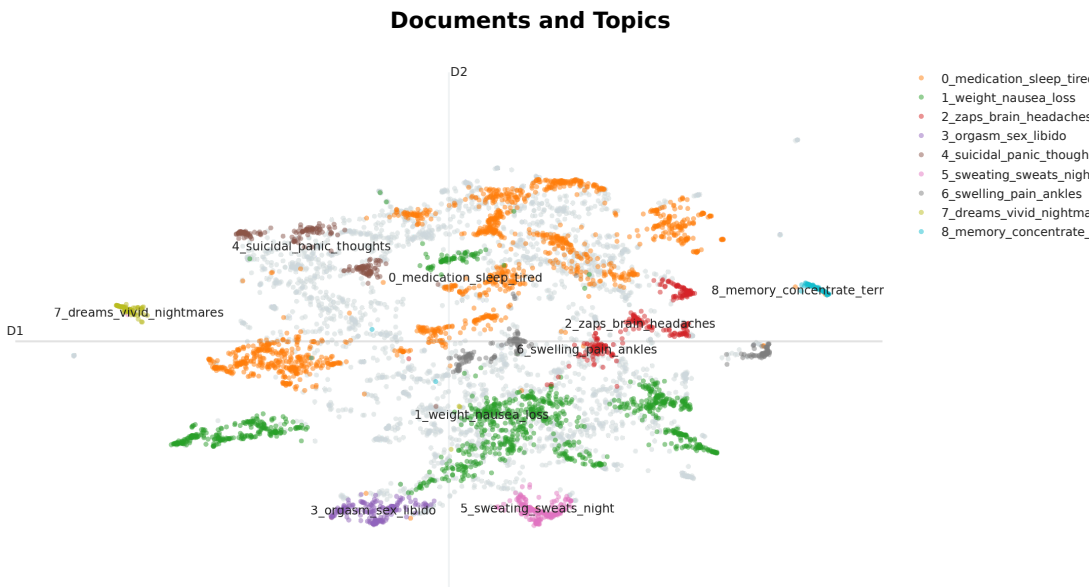


Fig. 4: UMAP projection of drug-masked ADE-positive instances colored by BERTopic topic assignment. Gray points denote outliers. Topics with distinctive vocabularies form tight clusters, whereas the two largest, more generic topics (Medication/Sleep/Tired and Weight/Nausea/Loss) occupy a more diffuse region of the embedding space.

Topic prevalence is broadly similar across antidepressant- and antihypertensive-related instances, supporting their representation within a shared topic space. The most notable differences are clinically plausible: swelling and ankle pain are more prevalent among antihypertensive-related instances, whereas sexual side effects are more prevalent among antidepressant-related instances. Overall, the findings indicate that patient forums provide non-redundant pharmacovigilance evidence, particularly for subjective, psychologically salient,

and highly specific adverse experiences that are less visible in social-media posts, medical case reports, and biomedical literature. From a sociotechnical perspective, these differences reflect how platform structure, documentation purpose, intended audience, and opportunities for extended narration shape which adverse experiences become visible for computational analysis.

Detection Performance Across Corpora. Table 2 compares zero-shot, few-shot, QLoRA, and RAG-SIDER configurations across the five corpora. Because the datasets are generally imbalanced toward **Not-ADE**, we report $F1_{\text{macro}}$ together with ADE-class precision and recall to distinguish overall classification performance from performance on the clinically relevant ADE class. Among the evaluated configurations, QLoRA-Llama-3.2-3B obtained the highest observed $F1_{\text{macro}}$ on ADE Corpus V2, SMM4H-2026, and CADEC V2 (0.91, 0.80, and 0.88, respectively). These three corpora provide more labeled training examples than the Dutch ADE Corpus and TwiMed-PubMed, suggesting that corpus-specific fine-tuning may be particularly effective when sufficient in-domain supervision is available. Under these evaluation settings, the compact QLoRA-adapted model performed comparably to or better than the substantially larger inference-only models evaluated in this study.

RAG-SIDER with Llama-3.1-405B attains the highest $F1_{\text{macro}}$ on the Dutch ADE Corpus and TwiMed-PubMed (0.95 and 0.91), and on the Dutch ADE Corpus achieves an ADE-class recall of 0.95, indicating high sensitivity to positive ADE instances. Few-shot prompting yields inconsistent changes relative to zero-shot prompting across corpora, with clear improvements in some settings and minimal or no benefit in others.

KronosRX external EHR evaluation. Table 3 shows a similar pattern in long-form English EHR data. On the cardiotoxicity test set, zero-shot Llama-3.1-405B achieved perfect ADE recall but produced substantial overprediction ($P_{\text{ADE}} = 0.35$). Few-shot prompting increased $F1_{\text{macro}}$ from 0.54 to 0.68, while the cardiac-specific RAG-SIDER configuration further increased it to 0.75 and improved P_{ADE} to 0.54, with recall remaining 0.86. For neurotoxicity, seizure-specific RAG-SIDER increased $F1_{\text{macro}}$ from 0.58 to 0.64 and precision from 0.27 to 0.31, while recall decreased from 1.00 to 0.92. In KronosRX, the toxicity-specific RAG configurations therefore reduced false-positive predictions while maintaining high sensitivity in encounter-level EHR evaluation.

Table 2: ADE detection performance across five corpora. Each cell reports $F1_{\text{macro}}/P_{\text{ADE}}/R_{\text{ADE}}$. Results are single-run descriptive estimates. Comparisons across model families and configurations are not statistically controlled because model size, supervision, and retrieval evidence differ.

Model	Configuration	ADE Corpus V2	SMM4H-2026	Dutch ADE	CADEC V2	TwiMed-PubMed
Llama-3.2-3B	Zero-shot	0.51/0.40/0.80	0.55/0.15/0.20	0.62/0.72/0.28	0.65/0.58/0.48	0.70/0.64/0.90
Llama-3.2-3B	QLoRA	0.91/0.85/0.82	0.80/0.76/0.53	0.87/0.94/0.75	0.88/0.87/0.83	0.85/0.75/0.77
Llama-3.3-70B	Zero-shot	0.67/0.40/0.95	0.73/0.39/0.79	0.83/0.95/0.61	0.70/0.66/0.82	0.88/0.76/0.90
Llama-3.3-70B	Few-shot	0.73/0.49/0.86	0.74/0.41/0.80	0.88/0.90/0.70	0.71/0.66/0.80	0.87/0.73/0.89
Llama-3.3-70B	RAG-SIDER	0.74/0.50/0.86	0.75/0.42/0.80	0.91/0.87/0.80	0.73/0.66/0.85	0.89/0.78/0.89
Llama-3.1-405B	Zero-shot	0.70/0.34/0.97	0.75/0.40/0.85	0.92/0.90/0.89	0.77/0.69/0.87	0.89/0.78/0.89
Llama-3.1-405B	Few-shot	0.76/0.49/0.85	0.77/0.45/0.76	0.92/0.89/0.85	0.77/0.69/0.80	0.89/0.78/0.90
Llama-3.1-405B	RAG-SIDER	0.78/0.52/0.93	0.78/0.48/0.80	0.95/0.89/0.95	0.80/0.68/0.74	0.91/0.82/0.89

Table 3: Encounter-level KronosRX results ($F1_{\text{macro}}/P_{\text{ADE}}/R_{\text{ADE}}$).

Cohort	Zero-shot	Few-shot	RAG-SIDER
Cardiotoxicity ($n = 87$)	0.54/0.35/1.00	0.68/0.45/0.95	0.75/0.54/0.86
Neurotoxicity ($n = 88$)	0.58/0.27/1.00	—	0.64/0.31/0.92

Across corpora, the strongest configuration is source-dependent: QLoRA performs best in several corpora with stronger task-specific supervision, whereas large-model RAG-SIDER is more competitive when external pharmacovigilance knowledge aligns well with the source and task. These patterns should not be interpreted as a controlled comparison of adaptation strategies, because the configurations differ in model scale, supervision, and retrieval evidence. Direct comparison with prior work is also limited by differences in task formulation and splits (for example, our ADE Corpus V2 experiments use a leakage-corrected split, and the Dutch ADE Corpus and CADEC V2 lack established binary baselines). We therefore emphasize comparisons among configurations evaluated on the same held-out test set within each corpus.

5. Discussion and Limitations

Our results show that data-production context strongly shapes how adverse drug experiences are expressed and how well they align with curated pharmacovigilance resources. Patient forums, social media, case reports, EHR notes, and biomedical literature are therefore not interchangeable sources of ADE evidence. SIDER coverage varies substantially across corpora, from 91.0% exact matches in TwiMed-PubMed to 9.5% in CADEC V2, reflecting its stronger alignment with standardized biomedical terminology than with patient-generated or multilingual language. Topic analysis similarly shows that CADEC V2 foregrounds subjective and psychologically salient experiences, whereas other sources emphasize broader medication-, sleep-, tiredness-, and pain-related themes. Model performance reflects these differences. QLoRA-Llama-3.2-3B performs well when sufficient in-domain supervision is available, whereas RAG-SIDER with Llama-3.1-405B is more competitive when external drug knowledge aligns with the source. The KronosRX EHR evaluation shows a related pattern: toxicity-specific RAG reduced false-positive encounter predictions while maintaining high recall. Together, these findings suggest that no single adaptation strategy is optimal across data sources; model design should account for both available supervision and knowledge-base alignment.

Several limitations remain. Comparisons confound model scale and training-set size, use binary ADE detection and single-run evaluations, and rely on LLM-assisted drug-mention extraction for some corpora. Repeated runs and confidence intervals were not available; therefore, the reported performance differences should be interpreted as descriptive observations rather than statistically significant differences. Topic estimates are also influenced by the larger CADEC V2 corpus, while KronosRX was evaluated only with 405B inference configurations rather than a matched 3B fine-tuned model. Future work should use matched model scales, repeated runs, joint extraction, and long-EHR training strategies such as evidence-preserving chunking or encounter-aware objectives. More broadly, combining source-sensitive models with knowledge-grounded retrieval and coverage auditing may help pharmacovigilance systems capture both documented harms and patient-expressed adverse experiences.

Ethics Statement

The KronosRX study was reviewed and approved by the Cedars-Sinai Institutional Review Board under protocol STUDY00004421 (KronosRx: Sending Patient Avatars to the Future to Transform Drug Safety Prediction). The study received expedited IRB approval effective August 24, 2025.

Acknowledgments

This work was supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM014731 and by the Advanced Research Projects Agency for Health (ARPA-H) CATALYST program through the KronosRx project, ‘KronosRx: Sending Patient Avatars to the Future to Transform Drug Safety Prediction’ (ARPA-H Award 3586 - <https://arpa-h.gov/explore-funding/awards/3586>) The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or ARPA-H.

References

1. M. S. Donaldson, J. M. Corrigan and L. T. Kohn, To err is human: building a safer health system (2000).
2. X. Qiu, S. Shao, H. Wang and X. Tan, Bio-k-transformer: A pre-trained transformer-based sequence-to-sequence model for adverse drug reactions prediction, *Computer Methods and Programs in Biomedicine* **260**, p. 108524 (2025).
3. D. C. Classen, S. L. Pestotnik, R. S. Evans, J. F. Lloyd and J. P. Burke, Adverse drug events in hospitalized patients: excess length of stay, extra costs, and attributable mortality, *Jama* **277**, 301 (1997).
4. M. De Rosa, G. Fenza, A. Gallo, M. Gallo and V. Loia, Pharmacovigilance in the era of social media: discovering adverse drug events cross-relating twitter and pubmed, *Future Generation Computer Systems* **114**, 394 (2021).
5. E. Kopacheva, A. Henriksson, H. Dalianis, T. Hammar and A. Lincke, Identifying adverse drug events in clinical text using fine-tuned clinical language models: Machine learning study, *JMIR Formative Research* **9**, p. e71949 (2025).
6. S. Golder, K. O’Connor, G. Lopez-Garcia, N. P. Tatonetti and G. Gonzalez-Hernandez, Leveraging unstructured data in electronic health records to detect adverse events from pediatric drug use: A scoping review, *Annual review of biomedical data science* **8**, 227 (2025).
7. L. Chen, Y. Gu, X. Ji, Z. Sun, H. Li, Y. Gao and Y. Huang, Extracting medications and associated adverse drug events using a natural language processing system combining knowledge base and deep learning, *Journal of the American Medical Informatics Association* **27**, 56 (2020).
8. I. Guellil, Y. Berrachedi, N. E. Chenni, M.-N. Abboud, J. Wu, H. Wu and B. Alex, Detecting adverse drug events in social media: A brief literature review, *SN Computer Science* **7**, p. 199 (2026).
9. S. Golder, K. O’Connor, Y. Wang, A. Klein and G. Gonzalez Hernandez, The value of social media analysis for adverse events detection and pharmacovigilance: scoping review, *JMIR public health and surveillance* **10**, p. e59167 (2024).
10. O. Elbiach, H. Grissette *et al.*, Benchmarking large language models for adverse drug reaction extraction in social media and clinical texts, *Results in Engineering* , p. 107362 (2025).
11. A. Wexler, A. Davoudi, D. Weissenbacher, R. Choi, K. O’Connor, H. Cummings and G. Gonzalez-

- Hernandez, Pregnancy and health in the age of the internet: A content analysis of online “birth club” forums, *PloS one* **15**, p. e0230947 (2020).
12. X. Chen, C. Faviez, S. Schuck, A. Lillo-Le-Louët, N. Texier, B. Dahamna, C. Huot, P. Foulquié, S. Pereira, V. Leroux *et al.*, Mining patients’ narratives in social media for pharmacovigilance: adverse effects and misuse of methylphenidate, *Frontiers in pharmacology* **9**, p. 541 (2018).
 13. L. Wu, H. Fang, Y. Qu, J. Xu and W. Tong, Leveraging fda labeling documents and large language model to enhance annotation, profiling, and classification of drug adverse events with askfdalabel: L. wu et al., *Drug Safety* **48**, 655 (2025).
 14. M. Kuhn, I. Letunic, L. J. Jensen and P. Bork, The sider database of drugs and side effects, *Nucleic acids research* **44**, D1075 (2016).
 15. S. Nygren, O. Erdogan, P. Avci, A. Daniels, R. Rassool, A. Beheshti and D. Galeano, Rag-based architectures for drug side effect retrieval using compact llms, *Scientific Reports* **16**, p. 12754 (2026).
 16. H. Prioleau, S. K. Aryal and J. Blackstone, Leveraging large language models for adverse drug event detection: A comparative study of token and span-based named entity recognition, 205 (2025).
 17. F. Zeidi, R. Christof, R. König and L. Childs, ‘pei at# smm4h-heard 2025: Enhancing adverse event detection via error-driven data augmentation, p. 56 (2025).
 18. Z. Duan, K. Wei, Z. Xue, J. Zhou, S. Yang, S. Ma, J. Jin *et al.*, Crowdsourcing-based knowledge graph construction for drug side effects using large language models with an application on semaglutide, *arXiv preprint arXiv:2504.04346* (2025).
 19. Z. Huang, J. Lu, Q. Sun and H. Lu, Multi-axis evaluation of adverse drug event classification: Clinical error cost, linguistic failure taxonomy, and leakage-corrected benchmarks on ade corpus v2, 223 (2026).
 20. F. Dong, W. Guo, J. Liu, T. A. Patterson and H. Hong, Bert-based language model for accurate drug adverse event extraction from social media: implementation, evaluation, and contributions to pharmacovigilance practices, *Frontiers in Public Health* **12**, p. 1392180 (2024).
 21. H. Gurulingappa, A. M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius and L. Toldo, Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports, *Journal of biomedical informatics* **45**, 885 (2012).
 22. G. Lopez-Garcia, J. Cortina, J. Berkowitz, J. Chan, S. K. Dey, I. F. Amaro, F. Gallego, L. Gryboski, A. Z. Klein, F. Z. Kolehparcheh *et al.*, Overview of the 11th social media mining for health (# smm4h) and health real-world data (heard) shared tasks at acl 2026, 353 (2026).
 23. X. Dai, S. Karimi, A. Sarker, B. Hachey and C. Paris, Multiade: A multi-domain benchmark for adverse drug event extraction, *Journal of Biomedical Informatics* **160**, p. 104744 (2024).
 24. R. M. Murphy, D. A. Dongelmans, N. F. de Keizer, R. J. Jongeneel, C. H. Koster, K. J. Jager, A. Abu-Hanna, I. Calixto and J. E. Klotowska, Creation of a gold standard dutch corpus of clinical notes for adverse drug event detection: the dutch ade corpus, *Language Resources and Evaluation* **59**, 2763 (2025).
 25. N. Alvaro, Y. Miyao and N. Collier, Twimed: Twitter and pubmed comparable corpus of drugs, diseases, symptoms, and their relations, *JMIR public health and surveillance* **3**, p. e24 (2017).
 26. M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, *arXiv preprint arXiv:2203.05794* (2022).
 27. P. R. Shakyia, A. Khaneja and K. B. Waghlikar, For clinical data extraction, qlora attains accuracy close to lora while requiring lower compute resources, *medRxiv*, 2025 (2025).
 28. J. Liu, G. Wang and G. Chen, Identifying adverse drug events from social media using an improved semisupervised method, *IEEE Intelligent Systems* **34**, 66 (2019).
 29. J. Wang, L. Zhao, Y. Ye and Y. Zhang, Adverse event detection by integrating twitter data and vaers, *Journal of biomedical semantics* **9**, p. 19 (2018).

30. J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, **1**, 4171 (2019).
31. E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann and M. McDermott, Publicly available clinical bert embeddings, 72 (2019).
32. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* (2019).
33. H. Li and H. Lin, Adverse drug reaction detection on social media based on large language models, *Information* **17**, p. 352 (2026).
34. C. Hiba, E. H. Nfaoui and C. Loqman, Fine-tuning transformer models for adverse drug event identification and extraction in biomedical corpora: A comparative study, 957 (2023).
35. K. Huang, J. Altosaar and R. Ranganath, Clinicalbert: Modeling clinical notes and predicting hospital readmission, *arXiv preprint arXiv:1904.05342* (2019).
36. C. Estupiñán-Ojeda, R. J. Sandomingo-Freire, L. Padró and J. Turmo, High-fidelity parameter-efficient fine-tuning for joint recognition and linking of diagnoses to icd-10 in non-standard primary care notes, *JAMIA open* **8**, p. ooaf120 (2025).
37. J. Choi, N. Palumbo, P. Chalasani, M. M. Engelhard, S. Jha, A. Kumar and D. Page, Malade: Orchestration of llm-powered agents with retrieval augmented generation for pharmacovigilance, *arXiv preprint arXiv:2408.01869* (2024).
38. P. N. Srinivasu, A. Samudrala, P. G. Devesh, S. A. Reddy and A. A. Nuaim, Quantized low-rank adaptation in large language models for clinical text simplification, *International Journal of Computational Intelligence Systems* **19**, p. 267 (2026).

EchoStature: Spatial Demographic Conditioning for Pediatric Ejection Fraction Regression, with Subgroup-Stratified Evaluation

Siddhi Narayan^{†1}, Khadeeja Hussain^{†1}, Sai Phani Parsa², Jahnavi Veeramachinani³, Mohammad Husain^{*1}, Sai Chandra Kosaraju^{*1}

¹*Department of Computer Science, California State Polytechnic University, Pomona, California, USA*

²*Department of Computer Science, University of Nevada Las Vegas, Las Vegas, Nevada, USA*

³*Department of Information Systems, University of Nevada Las Vegas, Las Vegas, Nevada, USA*

**Corresponding Authors: mihusain@cpp.edu, skosaraju@cpp.edu*

Left ventricular ejection fraction (EF) is the principal measure of systolic function in pediatric echocardiography, and deep video models now estimate it at accuracy comparable to expert readers. Pediatric cohorts nonetheless span an order of magnitude in body size, and reported accuracy is pooled across that range. Body size, rather than age or sex, governs pediatric cardiac dimensions, and it acts on echocardiograms through geometry rather than as a scalar offset, yet demographic conditioning in EF regression conventionally encodes it as a low-dimensional vector combined after spatial pooling.

We present EchoStature, which uses gated fusion to integrate real echocardiograms with demographically conditioned 3D reconstructions while limiting reconstruction artifacts. A hierarchical contrastive objective further organizes the latent space by EF and demographic subgroup. On EchoNet-Pediatric, ECHOStature matches the video-only reference model in overall MAE (4.49) and remains stable when demographic attributes are ablated at inference. However, subgroup analysis reveals substantially higher error in infants than in children aged 2-5 years (5.79 vs. 3.89 MAE), highlighting age-dependent performance disparities that are obscured by pooled metrics. Pooled accuracy therefore conceals clinically meaningful variation, particularly among the youngest patients, for whom echocardiography is often the only available assessment of ventricular function. Accordingly, we report subgroup-stratified error for every configuration and argue that such reporting should be standard in pediatric EF regression. Code availability: <https://github.com/KosarajuBioAI/EchoStature>

Keywords: Pediatric echocardiography; Ejection fraction regression; Subgroup performance; Demographic conditioning; Gated feature fusion; Conditional video generation; Counterfactual analysis; Explainable AI.

[†]Authors contributed equally to this work.

© 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. This work is partially supported by Mohammad Husain and Sai Chandra Kosaraju's Cal Poly Pomona Internal Grant

1. Introduction

Left ventricular ejection fraction (EF) anchors systolic assessment, guiding cardiomyopathy diagnosis, heart-failure staging, and therapy timing. Echocardiography, the frontline cardiac test in children, is the primary route to EF because it is fast, portable, and free of ionizing radiation,^{1,2} but manual estimates are imprecise and vary substantially between readers.¹ Pediatric cardiac structure changes rapidly with age, and EchoNet-Pediatric spans this range with 4,467 studies from 1,958 patients aged 0–18 years.²

EchoNet-Dynamic reached expert-level accuracy (mean absolute error [MAE] 4.1%) by regressing EF with factorized spatiotemporal convolutions,^{1,3} and later models refined frame selection,⁴ anatomical grounding,^{5,6} and temporal aggregation.⁷ Pediatric retraining outperformed an adult model ($P < .001$; MAE 3.66% with two views), yet pooled accuracy hides subgroup variation: 86% of EchoNet-Pediatric studies have EF $\geq 55\%$, and sex, age, and body habitus are unevenly represented.² Imbalanced training degrades performance on underrepresented groups in classification and cardiac segmentation,^{8,9} and chest radiograph classifiers reportedly underdiagnose under-served subgroups,¹⁰ though this was later contested. Pediatric EF regression therefore demands subgroup-stratified reporting.

Body size governs pediatric cardiac dimensions more strongly than age or sex,¹¹ and it acts spatially, altering chamber size, sector geometry, and the apparent scale of ventricular motion. Existing fusion does not preserve this: demographic vectors join video features only after pooling has discarded spatial structure,¹² and broadcast vectors stay spatially uniform. In semantic synthesis, such uniform conditioning washes out with depth, unlike spatially varying conditioning.¹³ Feature-wise modulation is a partial remedy but also acts uniformly across positions,¹⁴ and short metadata-to-label routes invite shortcut learning.¹⁵

Conditional generation can instead express attributes as image structure,¹⁶ in video through three-dimensional convolutions¹⁷ and in echocardiography through D’ARTAGNAN, which alters EF while preserving patient anatomy and imaging style.¹⁸ Generated data, however, mislead when realism substitutes for utility,¹⁹ and generator artifacts inject spurious correlations,²⁰ so their influence must be limited. Contrastive objectives structure imbalanced latent spaces,^{21,22} and Grad-CAM and counterfactual probes test whether predictions rely on cardiac evidence rather than metadata shortcuts.^{23–26}

These threads converge on design requirements: demographics should reach the encoder as spatial structure rather than a scalar offset, any generated stream should stay subordinate to the real acquisition, and the latent space should be organized by both EF and demographic subgroup. Evaluation should pair parameter-sensitive attribution with controlled demographic perturbation and report error by subgroup.

2. Methodology

2.1. Overview

ECHOStature estimates pediatric EF from an echocardiographic clip, its demographic attributes, and a demographically rendered synthetic clip. The real clip and the synthetic rendering pass through a shared frame encoder, and a learned gate combines their frame-level

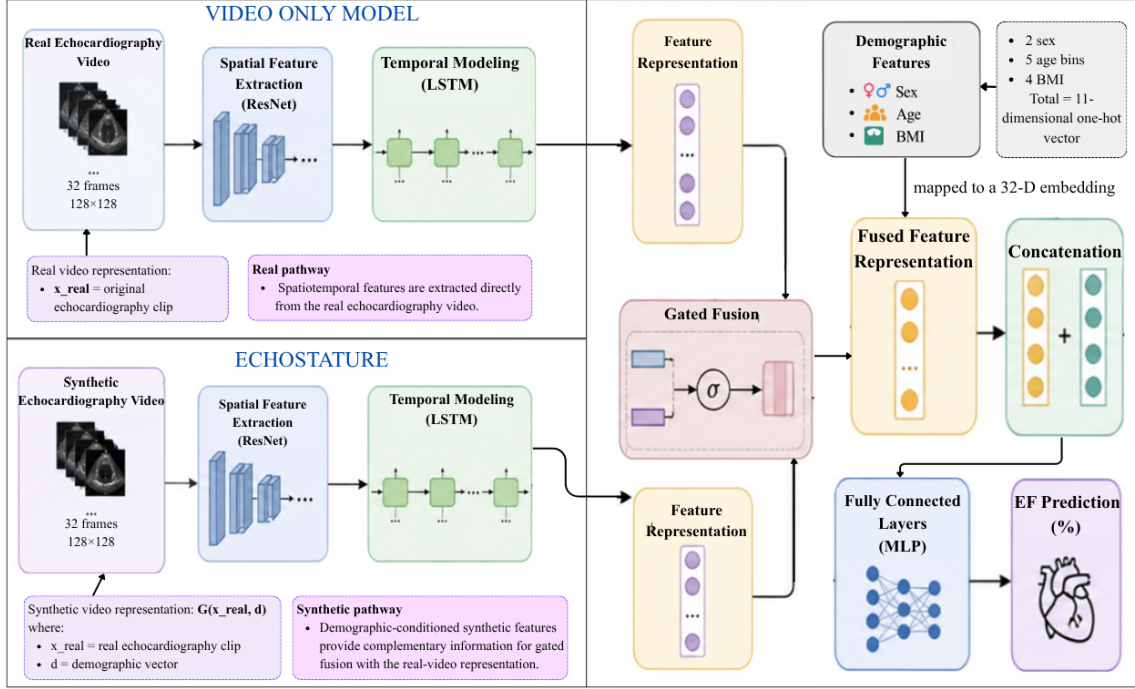


Fig. 1: Overview of the proposed ECHOSTATURE model. The benchmark model estimates ejection fraction (EF) from a real echocardiography video and demographic embedding, whereas ECHOSTATURE additionally incorporates a paired synthetic reconstruction through gated feature fusion to supply a rendered demographic stream.

embeddings before temporal aggregation. Demographic attributes additionally enter as a low-dimensional embedding concatenated with the pooled video representation, following standard practice in multimodal EF regression.¹² Weight sharing across the two streams keeps them in a common embedding space and confines the additional parameters to the gating network. Figure 1 summarizes the architecture.

2.2. Feature representation

Each echocardiographic clip is resized to 128×128 pixels, uniformly sampled or zero-padded to a fixed length of $T = 32$ frames, and normalized to $[0, 1]$. Demographic attributes supply contextual information that the clip does not directly reveal, since acquisition geometry and chamber scale depend on body size.¹¹ Sex, age bin, and BMI category are encoded as an 11-dimensional one-hot vector comprising two sex categories, five age bins, and four BMI categories,

$$d \in \{0, 1\}^{11}, \quad (1)$$

$$e = \psi(d) \in \mathbb{R}^{32}, \quad (2)$$

Here, $\mathbf{d}$ denotes the demographic attribute vector, $\psi(\cdot)$ denotes a lightweight multilayer perceptron, and $\mathbf{e}$ denotes the resulting 32-dimensional demographic embedding. The same de-

mographic vector conditions the C3DGAN generator, which produces a paired reconstruction of the real echocardiographic clip, as described below.

C3DGAN. The synthetic clip used by ECHOStature is produced by a conditional 3D reconstruction GAN (C3DGAN). The generator receives the real echocardiographic clip x_{real} together with its 11-dimensional demographic vector d , and produces a paired reconstruction

$$x_{\text{syn}} = G(x_{\text{real}}, d). \quad (3)$$

The generator therefore reconstructs the observed clip under its matched demographic condition rather than sampling a clip from latent noise. It operates on grayscale clips of 16 frames at 64×64 and preserves the temporal axis throughout. After generation, each stored reconstruction is resampled with its paired real clip to $T = 32$ frames at 128×128 for EF training. The checkpoint used here corresponds to epoch 183. ECHOStature does not call G during EF training: it loads the precomputed paired clip and inherits the EF label of x_{real} .

Generator. G is a 3D U-Net with four encoder levels of 64, 128, 256, and 512 channels. The demographic vector is mapped by a small MLP to a 128-dimensional embedding and fused into the video features by projection, broadcast, and $1 \times 1 \times 1$ convolution at the first three encoder levels and at the bottleneck. Spatial downsampling is used in the encoder while the temporal length is kept fixed. Residual blocks use $3 \times 3 \times 3$ convolutions with squeeze-and-excitation channel attention. The decoder uses transposed convolutions, U-Net skips, and a final tanh to emit a single-channel clip. The saved generator contains approximately 1.07×10^8 parameters.

Discriminator. The discriminator is a spectral-normalized 3D PatchGAN that scores real versus reconstructed video and does not take d as a direct input. Four spatial downsampling stages use 64, 128, 256, and 512 channels; a convolutional head produces patch-level real/fake scores, and an auxiliary head maps a pooled 512-dimensional representation to an 11-way demographic prediction. The saved discriminator contains approximately 2.90×10^6 parameters.

Training and use in EchoStature. Generator and discriminator were trained with Adam at learning rate 10^{-4} and $\beta = (0.5, 0.999)$. The checkpoint stores a single combined loss value and does not retain the individual loss terms or their weights, so those coefficients are not reported here. After training, one reconstruction was written for each real clip used by ECHOStature; the EF model therefore sees a fixed paired rendered stream rather than an on-the-fly sample from G . This stream is the spatiotemporal demographic pathway that the gate in Section 2.3 can admit or suppress; it is complementary to the post-pooling embedding $e = \psi(d)$, which remains a low-dimensional scalar pathway.

2.3. Architecture and training

Let $\phi(\cdot)$ denote the frame encoder, comprising a 2D convolutional backbone followed by a linear projection to dimension d_f which is the frame-embedding dimension. The same encoder is applied to corresponding frames of both streams,

$$r_t = \phi(x_t^{\text{real}}) \in \mathbb{R}^{d_f}, \quad (4)$$

$$\tilde{r}_t = \phi(x_t^{\text{syn}}) \in \mathbb{R}^{d_f}, \quad (5)$$

for $t = 1, \dots, T$. Frame-wise encoding admits a gate that operates at every time step, allowing the model to vary its reliance on the rendered stream across the cardiac cycle.

A gating network $\gamma(\cdot)$ estimates the per-dimension contribution of each stream at every frame,

$$c_t = [r_t; \tilde{r}_t] \in \mathbb{R}^{2d_f}, \quad (6)$$

$$g_t = \sigma(\gamma(c_t)) \in [0, 1]^{d_f}, \quad (7)$$

$$f_t = g_t \odot r_t + (1 - g_t) \odot \tilde{r}_t, \quad (8)$$

where $\odot$ denotes element-wise multiplication and $\sigma(\cdot)$ denotes the element-wise sigmoid. Here, $\mathbf{r}_t$ and $\tilde{\mathbf{r}}_t$ denote the frame-level representations of the real and synthetic streams, respectively, $\mathbf{c}_t$ is their concatenated representation, $\mathbf{g}_t$ is the learned gate, and $\mathbf{f}_t$ is the resulting fused representation at time step t . The convex form bounds the synthetic contribution and allows the network to suppress the rendered stream entirely as $g_t \rightarrow 1$, which limits the propagation of reconstruction artifacts.¹⁹

A recurrent temporal encoder $\tau(\cdot)$, instantiated as a bidirectional long short-term memory network with hidden dimension d_h per direction, models the dynamics of the fused sequence. Temporal attention then aggregates the resulting states into a single video representation,

$$z_t = \tau(f_t \mid \{f_k\}_{k=1}^T) \in \mathbb{R}^{2d_h}, \quad (9)$$

$$a_t = \frac{\exp(w^\top z_t)}{\sum_{k=1}^T \exp(w^\top z_k)}, \quad (10)$$

$$p = \sum_{t=1}^T a_t z_t, \quad (11)$$

where $w \in \mathbb{R}^{2d_h}$ is a learnable attention vector shared across time steps, $\mathbf{z}_t$ is the bidirectional temporal state at frame t , a_t is its attention weight, and $\mathbf{p}$ is the temporally pooled video representation. The pooled representation is concatenated with the demographic embedding, $u = [p; e]$, and passed through a regression head $\rho(\cdot)$ with sigmoid activation,

$$\hat{y} = \sigma(\rho(u)) \in [0, 1], \quad \widehat{\text{EF}} = 100 \hat{y}. \quad (12)$$

Here, y is the ground-truth EF divided by 100, so $y \in [0, 1]$, and $\hat{y}$ is the predicted value on that same scale. $\widehat{\text{EF}} = 100 \hat{y}$ is the prediction in percentage points. Training combines supervised regression with a hierarchical contrastive objective. The regression loss balances robustness against squared-error optimization,

$$\mathcal{L}_{\text{EF}} = (1 - w_r) \mathcal{L}_\delta(\hat{y}, y) + w_r \|\hat{y} - y\|_2^2, \quad (13)$$

where $\mathcal{L}_\delta$ denotes the Huber loss and $w_r \in [0, 1]$ sets the balance between the two terms. The contrastive objective operates on the projected video embedding and proceeds in two

Table 1: Summary statistics of the EchoNet-Pediatric cohort.

Characteristic	Train	Validation
Number of clips	6,227	1,558
EF (% , mean $\pm$ SD)	61.1 $\pm$ 10.5	61.4 $\pm$ 9.7
Sex (F/M/O)	2661 / 3561 / 5	653 / 905 / 0
View (A4C/PSAX)	2657 / 3570	613 / 945

F: female; M: male; O: other; A4C: apical four-chamber; PSAX: parasternal short-axis.

stages. The first stage, $\mathcal{L}_{H1}$, treats samples sharing an EF-derived task label as positives, following supervised contrastive learning.²² The second stage, $\mathcal{L}_{H2}$, refines the representation by treating samples sharing a demographic subgroup as positives within each EF group, which counteracts the concentration of pediatric labels in the normal range.² The combined auxiliary loss is

$$\mathcal{L}_{HCL} = \mathcal{L}_{H1} + \lambda_{H2}(\epsilon) \mathcal{L}_{H2}, \quad (14)$$

where $\lambda_{H2}(\epsilon)$ increases with epoch ϵ , so that EF structure forms before subgroup structure is imposed. The overall objective is

$$\mathcal{L} = \mathcal{L}_{EF} + \alpha \mathcal{L}_{HCL}, \quad (15)$$

where α controls the contribution of the hierarchical contrastive objective, L_{EF} is the supervised regression loss, L_{δ} is the Huber loss, L_{H1} is the EF-based supervised contrastive loss, and L_{H2} is the demographic hierarchical contrastive loss. The parameter λ_{H2} controls the contribution of the second contrastive stage and varies with training epoch ϵ . The model is optimized with AdamW under cosine learning-rate scheduling.

2.4. Benchmark and implementation

We evaluate on EchoNet-Pediatric, which comprises 7,643 apical four-chamber and parasternal short-axis clips drawn from 4,467 studies across 1,958 patients aged 0 to 18 years.² As a benchmark reference we train a video-only regression model that shares the frame encoder, temporal encoder, attention module, and regression head of ECHOStature. The two models do not receive the same input. The reference model observes the echocardiographic clip alone, whereas ECHOStature additionally observes the demographic vector and the synthetic clip rendered from it. The comparison therefore evaluates supplying demographic context in both scalar and rendered form against a video-only reference; it does not isolate either form individually, and under the multi-seed protocol reported in Section 3.2 it does not establish a pooled accuracy gain. Both models use AdamW, learning rate 3×10^{-4} , a cosine schedule, batch size 8, the same number of epochs, and the same clip-level split. The contrastive weight is 0.04 for the video-only model and 0.025 for ECHOStature, and ECHOStature adds the gating network, so this comparison is not a single-component ablation.

3. Experiments and results

3.1. Dataset and experimental setup

We evaluate on EchoNet-Pediatric, predicting left ventricular ejection fraction from pediatric echocardiography videos.² Each sample comprises a processed grayscale clip acquired in the apical four-chamber (A4C) or parasternal short-axis (PSAX) view, together with the corresponding EF, sex, age band (0–1, 2–5, 6–10, 11–15, and 16–18 years), and BMI category. BMI categories are derived from fixed body-mass-index thresholds rather than age- and sex-specific pediatric percentiles, since percentile references are undefined below two years of age and the cohort spans patients from birth. The resulting categories stratify patients by raw body mass index and do not correspond to clinical weight-status definitions. After preprocessing and quality filtering, the cohort contains 7,785 labeled clips, partitioned at the clip level into 6,227 training and 1,558 validation clips. The validation set is an internal validation set rather than an external cohort. Because the partition is performed at the clip level, multiple clips from the same patient may potentially occur across the training and validation sets. Therefore, the reported validation results should not be interpreted as evidence of patient-disjoint generalization. Table 1 summarizes the partitions, and all reported results are obtained on the validation set. Each clip is resized to 128×128 pixels, uniformly sampled to $T = 32$ frames, and normalized to $[0, 1]$. ECHOSTATURE additionally receives one demographically rendered synthetic clip for each real clip. The synthetic clip is paired with its real counterpart and shares the counterpart’s EF label and demographic attributes; it is therefore treated as demographic-conditioned visual context rather than as an independently labeled clinical observation, while the video-only reference model observes the echocardiographic clip alone. Each paired clip was compared with its source clip by structural similarity and peak signal-to-noise ratio. Across the 7,785 clips used for training and validation, mean structural similarity is 0.995 and mean peak signal-to-noise ratio is 49.0 dB. This comparison measures agreement with the source when the generator receives that source and its own demographic vector. It does not test whether a different demographic vector would change the clip, and it does not establish clinical realism or clinical validity.

3.2. Overall performance

Table 2 reports regression performance on the validation set. Under seven matched 100-epoch seeds, ECHOSTATURE reaches 4.49 ± 0.04 MAE, 6.17 ± 0.07 RMSE, and $R^2 = 0.587 \pm 0.010$, compared with 4.49 ± 0.05 , 6.17 ± 0.06 , and 0.588 ± 0.008 for the video-only reference. The pooled difference is therefore negligible under this protocol. The subgroup analysis that follows shows that age-group variation remains far larger than this pooled architecture gap.

3.3. Subgroup performance

Table 4 reports MAE (mean $\pm$ SD over seven seeds) across sex, age, and body-mass subgroups, and the variation it reveals exceeds every architectural effect measured in this work. Infants aged 0–1 years reach 5.79 ± 0.20 MAE against 3.89 ± 0.14 for the 2–5 year group. That difference of about 1.9 percentage points is far larger than the negligible pooled margin between

ECHOStATURE and the video-only reference. Where a patient falls in the pediatric age range therefore affects expected error far more than which model estimates it.

The observed performance gap may reflect characteristics of infant echocardiography, including differences in heart rate, chamber size, and the amount of anatomical information represented at the fixed spatial and temporal resolution used in this study. For example, higher infant heart rates could cause a fixed 32-frame sample to represent a different portion of the cardiac cycle than in older patients, while smaller chamber dimensions could occupy fewer pixels at the fixed input resolution. These factors are plausible explanations for the observed subgroup difference, but their individual contributions were not directly tested in the present analysis.

Differences across sex and body-mass strata are smaller than the age gap. Female patients reach 4.23 ± 0.07 MAE against 4.68 ± 0.07 for male patients, and the underweight stratum reaches 4.38 ± 0.11 against 5.17 ± 0.15 for overweight. Subgroup counts are unequal, so per-subgroup estimates carry correspondingly different precision, and the smaller strata warrant cautious interpretation.

The practical implication is that pooled accuracy misrepresents pediatric performance. A model reported near 4.49 MAE overall still performs near 5.8 MAE for the youngest patients, who are also those for whom echocardiography is most often the only available assessment of ventricular function.²

Reporting a single pooled figure conceals this variation entirely, and it conceals it for the subgroup least able to tolerate error. Pediatric EF regression therefore warrants subgroup-stratified reporting as standard practice, independently of the architecture under evaluation.

3.4. Ablation studies

Gated fusion. Table 3 compares variants trained under a matched 100-epoch budget without the contrastive objective. The three rows share that 100-epoch budget, AdamW, learning rate 3×10^{-4} , a cosine schedule, batch size 8, the same clip-level split, and a contrastive weight of zero, and are reported as mean $\pm$ SD over five seeds. The real-only and rendered-only rows use the same video-only network and differ in which clips are supplied. The real-plus-rendered row uses the ECHOStATURE network, including the gate. Under five matched 100-epoch seeds without the contrastive objective (Table 3), real-only reaches 4.44 ± 0.07 MAE, real-plus-rendered reaches 4.46 ± 0.15 , and rendered-only reaches 4.66 ± 0.03 . The gated stream therefore does not improve pooled MAE over real-only under this protocol, while rendered-only remains worse than real video, indicating that the rendered clips retain EF-relevant structure but are not a substitute for real acquisitions.

Contrastive objective. Under the same 100-epoch budget, the five-seed no-contrastive real-plus-rendered model reaches 4.46 ± 0.15 MAE (Table 3), compared with 4.49 ± 0.04 for the seven-seed full ECHOStATURE configuration (Table 2). Because the seed counts differ and several components change together, these figures do not isolate the contrastive term. The objective remains part of the full training configuration, and its contribution is not independently established.

Demographics-only control. Replacing the learned video representation entering the re-

gression head with zeros, while retaining the demographic embedding, degrades performance substantially: on a representative checkpoint from the seven-seed protocol in Table 2, MAE rises from about 4.49 to 7.61 for the video-only reference and from about 4.49 to 10.69 for ECHOStATURE. This control does not retrain either model. At inference, the pooled video vector is set to zero and the demographic embedding is left unchanged. Both models therefore rely predominantly on cardiac video rather than on demographic metadata.

Demographic conditioning.

Table 5 reports inference-time perturbation of the demographic vector with the video inputs held fixed. Across seven seeds, zeroing the vector raises MAE from 4.49 ± 0.04 to 4.53 ± 0.04 , and shuffling leaves MAE at 4.49 ± 0.04 , so the change is at most about 0.04 MAE. This perturbation affects only the scalar demographic pathway, since the synthetic clip is rendered before inference and is unchanged by the manipulation, so the result bounds the contribution of the post-pooling embedding rather than that of the rendered stream.

3.5. *Rendered-stream chamber structure*

Pixel-level agreement alone does not establish a coherent beating cavity in the rendered stream. On 100 random validation pairs, a classical dark-cavity connected-component check detects a cavity in all 32 frames of every real and rendered clip. Relative area change is 0.65 ± 0.11 (real) vs. 0.61 ± 0.12 (rendered; Kolmogorov–Smirnov $p = 0.054$), with associated within-pair ranks (Spearman $\rho = 0.72$). Figure 2 summarizes the check. We treat this as image-analysis evidence of a detectable cavity under paired reconstruction, not as clinical left-ventricle segmentation.

3.6. *Model interpretation*

Representation structure. Figure 3b(ii) presents UMAP projections of the regressor input representation, colored by sex, age bin, and body-mass stratum. Partial separation is visible for sex and age bin. Body-mass strata remain largely intermixed in this projection even when points are colored by the true weight/height BMI categories, which is consistent with the coarse threshold-based encoding described in Section 3.1. We present these projections as a qualitative description rather than as evidence of a learned effect, since demographic attributes are partly inferable from echocardiograms directly²⁴ and since UMAP geometry depends on the chosen neighborhood parameters.

Attribution and counterfactual stability. Grad-CAM maps from both models concentrate on the imaging sector over cardiac structures rather than off-sector background.²⁵ Figure 3b shows predictions under controlled perturbation of sex, age bin, and BMI with the video held fixed. Predicted EF shifts only marginally, consistent with Table 5, while attribution maps can still change under the same perturbations.

4. Discussion

The dominant source of error variation in pediatric EF regression is the patient, not the architecture. The difference between the best and worst age subgroups in our experiments is far larger than the negligible pooled difference between the two models, and infants incur the

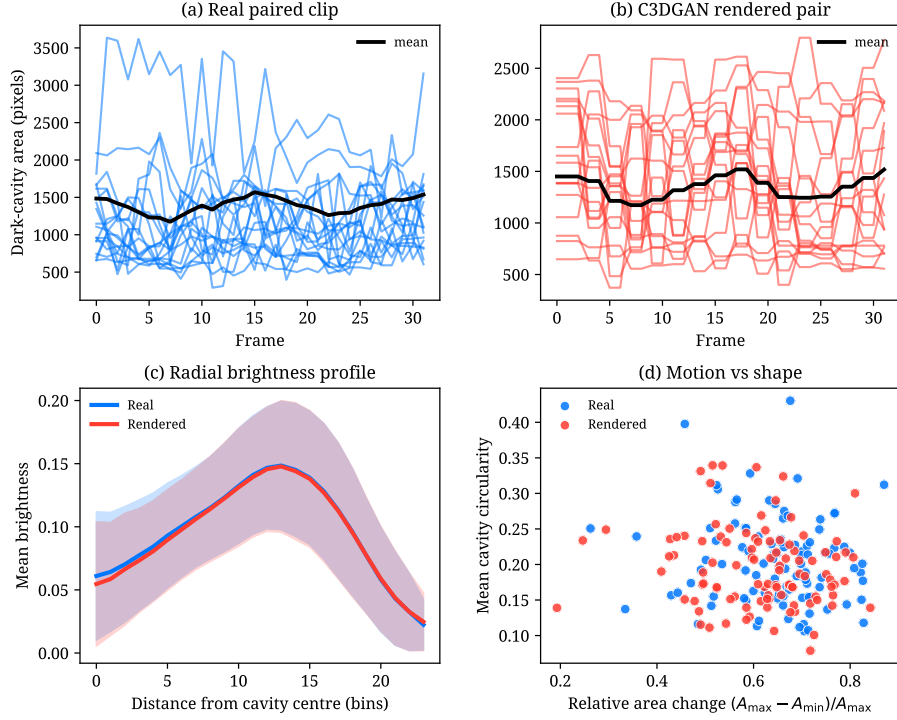


Fig. 2: Chamber structure in paired real and C3DGAN-rendered validation clips used by ECHOStATURE. (a, b) Dark-cavity area over 32 frames for 100 real clips and their rendered pairs (thin curves: random subset; black: mean). (c) Radial brightness profile from the cavity centre. (d) Relative area change versus mean cavity circularity.

Table 2: Overall EF regression on the validation set (mean $\pm$ SD over 7 seeds, 100 epochs).

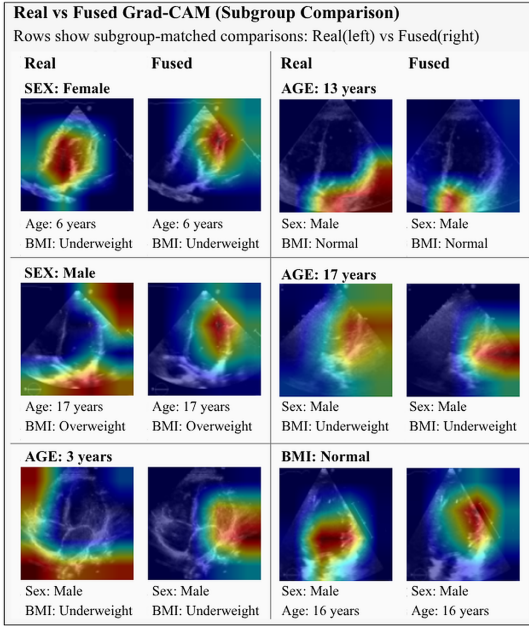
Model	MAE $\downarrow$	RMSE $\downarrow$	R ² $\uparrow$
Video-only	4.49 $\pm$ 0.05	6.17 $\pm$ 0.06	0.588 $\pm$ 0.008
ECHOStATURE	4.49 $\pm$ 0.04	6.17 $\pm$ 0.07	0.587 $\pm$ 0.010

Table 3: Input ablation under a matched 100-epoch schedule without the contrastive objective (mean $\pm$ SD over 5 seeds).

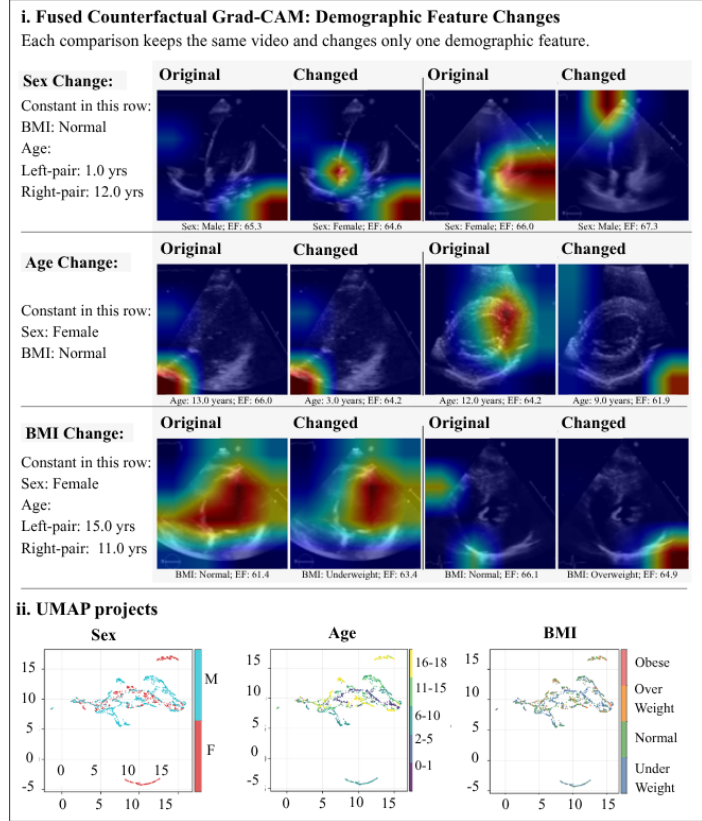
Input streams	Fusion	MAE $\downarrow$	RMSE $\downarrow$	R ² $\uparrow$
Real + rendered	✓	4.46 $\pm$ 0.15	6.15 $\pm$ 0.28	0.590 $\pm$ 0.037
Real only	—	4.44 $\pm$ 0.07	6.07 $\pm$ 0.10	0.601 $\pm$ 0.013
Rendered only	—	4.66 $\pm$ 0.03	6.38 $\pm$ 0.07	0.564 $\pm$ 0.010

highest error of any subgroup. Architectural comparison on pooled metrics therefore optimizes a quantity that obscures the failure mode most relevant to deployment, since a model deployed across a pediatric service encounters the full age range and performs measurably worse at one end of it. We report subgroup-stratified error for every configuration in this work and recommend the practice generally.

Under five matched 100-epoch seeds without the contrastive objective (Table 3), real-plus-rendered reaches 4.46 ± 0.15 MAE against 4.44 ± 0.07 for real-only, so the gated stream does not confer a pooled MAE gain in that ablation; rendered-only remains worse at 4.66 ± 0.03 . Under seven matched seeds with the full training objective, the pooled MAE of ECHOStATURE



(a) Grad-CAM comparison.



(b) Counterfactual analysis.

Fig. 3: Interpretability analyses. (a) Grad-CAM attribution maps from the video-only reference model and ECHOSTATURE on subgroup-matched validation clips. (b) Counterfactual analysis of demographic conditioning. (i) Grad-CAM maps under controlled perturbation of sex, age bin, and BMI category while the echocardiographic video is held fixed. (ii) UMAP projections of the learned regressor representation colored by sex, age bin, and BMI category. Sex and age show partial separation, whereas BMI categories remain largely intermixed.

and the video-only reference are both 4.49, so a pooled architectural advantage is not established there either. The convex gate is therefore best read as a mechanism that can admit complementary rendered information without treating rendered video as a substitute for real echocardiography, rather than as a source of a confirmed pooled gain.

Demographic attributes contribute auxiliary context rather than dominant signal. Zeroing the video representation while retaining the demographic embedding raises MAE from about 4.49 to 10.69 on a representative checkpoint, and perturbing the demographic vector at inference changes MAE by at most about 0.04 across seven seeds. Predictions consequently remain stable under controlled demographic perturbation, indicating that the model does not substitute metadata for visual evidence of ventricular mechanics. Attribution maps nonetheless shift under the same perturbations, so demographic context influences which spatial evidence the model weights without appreciably altering the resulting estimate. This dissociation between attribution sensitivity and prediction stability merits closer study, since attribution

Table 4: Subgroup MAE for ECHOStature on the validation set (mean $\pm$ SD over 7 seeds).

Subgroup	MAE $\downarrow$	Count
Sex: female	4.23 ± 0.07	653
Sex: male	4.68 ± 0.07	905
Age: 0–1	5.79 ± 0.20	132
Age: 2–5	3.89 ± 0.14	220
Age: 6–10	3.94 ± 0.13	394
Age: 11–15	4.53 ± 0.06	480
Age: 16–18	4.97 ± 0.12	332
BMI: underweight	4.38 ± 0.11	749
BMI: normal	4.44 ± 0.06	543
BMI: overweight	5.17 ± 0.15	159
BMI: obese	4.51 ± 0.20	107

Table 5: Demographic conditioning ablation for ECHOStature (mean $\pm$ SD over 7 seeds).

Mode	MAE $\downarrow$	MSE $\downarrow$	RMSE $\downarrow$	$R^2 \uparrow$
Normal	4.49 ± 0.04	38.12 ± 0.91	6.17 ± 0.07	0.587 ± 0.010
Zero	4.53 ± 0.04	39.09 ± 1.04	6.25 ± 0.08	0.576 ± 0.011
Shuffle	4.49 ± 0.04	37.98 ± 0.84	6.16 ± 0.07	0.589 ± 0.009

alone would suggest a stronger demographic dependence than the predictions exhibit.

Several limitations constrain these conclusions. Evaluation uses a fixed clip-level partition that does not guarantee patient-disjoint generalization, and model selection uses the same validation set. Main-model and demographic-perturbation results are reported as mean $\pm$ SD over seven seeds, and the matched no-contrastive input ablation as mean $\pm$ SD over five seeds, all on a single cohort; generalization across acquisition protocols remains untested. The comparison contrasts demographic-aware gated fusion against video-only regression, and it therefore does not isolate the contribution of the rendered stream from that of the demographic embedding. A no-contrastive configuration was trained under the same 100-epoch budget and summarized over five seeds, but because that protocol still differs from the seven-seed full model in more than the contrastive weight alone, the contribution of that objective remains unisolated. BMI categories rely on fixed thresholds rather than pediatric percentiles, which limits the granularity of body-size conditioning, and a percentile-based or continuous body-surface-area encoding would likely provide a stronger signal.¹¹ Grad-CAM and counterfactual analyses probe behavior without establishing clinical reliability, and expert evaluation and uncertainty estimation remain necessary. The dark-cavity connected-component check on real-rendered pairs likewise remains an image-analysis measurement rather than clinical left-ventricle segmentation.

The most consequential direction is to reduce error where it is largest. The higher infant error may be related to fixed-length, fixed-resolution encoding applied to a population with high heart rates and small chambers, but those mechanisms were not directly tested. Adaptive temporal sampling or resolution scaled to body size is therefore a hypothesis for

later experiments, rather than an established remedy. Whether accuracy and demographic independence trade off in pediatric EF regression is a second open question, and answering it requires perturbation analyses applied across architectures rather than to a single model.

Future evaluation should use patient-disjoint training, validation, and test partitions, so that performance reflects generalization to patients not seen in training. Validation on larger external cohorts, preferably from more than one center and acquired under different protocols, is also needed to determine whether the age-group error pattern and the gated-fusion comparison hold beyond EchoNet-Pediatric.

5. Conclusion

We presented ECHOStature, a pediatric EF regression model that fuses each real acquisition with a paired demographic-conditioned reconstruction through a learned gate. Under multi-seed evaluation the pooled difference from the video-only reference is negligible, and predictions remain stable under controlled demographic perturbation. Subgroup-stratified evaluation reveals larger variation across pediatric age groups than between architectures, with infants incurring the highest error. Pediatric EF regression therefore warrants subgroup-stratified reporting as standard practice, since pooled accuracy conceals precisely the population for which echocardiography is most often the only available assessment of ventricular function.

References

1. D. Ouyang, B. He, A. Ghorbani, N. Yuan and Others, Video-based ai for beat-to-beat assessment of cardiac function, *Nature* **580**, 252 (2020).
2. C. D. Reddy, L. Lopez, D. Ouyang, J. Y. Zou and B. He, Video-based deep learning for automated assessment of left ventricular ejection fraction in pediatric patients, *Journal of the American Society of Echocardiography* **36**, 482 (2023).
3. D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun and M. Paluri, A closer look at spatiotemporal convolutions for action recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
4. M. Mokhtari, T. Tsang, P. Abolmaesumi and R. Liao, EchoGNN: Explainable ejection fraction estimation with graph neural networks, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, , Lecture Notes in Computer Science Vol. 13434 (Springer, 2022) pp. 360–369.
5. S. Thomas, A. Gilbert and G. Ben-Yosef, Light-weight spatio-temporal graphs for segmentation and ejection fraction prediction in cardiac ultrasound, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, , Lecture Notes in Computer Science Vol. 13434 (Springer, 2022)
6. H. Li, Y. Wang, M. Qu, P. Cao, C. Feng and J. Yang, EchoEFNet: Multi-task deep learning network for automatic calculation of left ventricular ejection fraction in 2D echocardiography, *Computers in Biology and Medicine* **156**, p. 106705 (2023).
7. H. Reynaud, A. Vlontzos, B. Hou, A. Beqiri, P. Leeson and B. Kainz, Ultrasound video transformers for cardiac ejection fraction estimation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021.
8. A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone and E. Ferrante, Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis, *Proceedings of the National Academy of Sciences* **117**, 12592 (2020).
9. E. Puyol-Antón, B. Ruijsink, S. K. Piechnik, S. Neubauer, S. E. Petersen, R. Razavi and A. P. King, Fairness in cardiac MR image analysis: An investigation of bias due to data imbalance in deep learning based segmentation, in *MICCAI*, , LNCS Vol. 129032021.
10. L. Seyyed-Kalantari, H. Zhang, M. B. McDermott, I. Y. Chen and M. Ghassemi, Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations, *Nature medicine* **27**, 2176 (2021).
11. L. Lopez, S. Colan, M. Stylianou, S. Granger, F. Trachtenberg, P. Frommelt, G. Pearson, J. Camarda, J. Cnota, M. Cohen *et al.*, Relationship of echocardiographic z scores adjusted for body surface area to age, sex, race, and ethnicity: the pediatric heart network normal echocardiogram database, *Circulation: Cardiovascular Imaging* **10**, p. e006979 (2017).
12. S. Lee, J. Kim, K. Kang and M. Kang, Explainable integrative bipartite graph convolutional neural network for predicting ejection fraction in echocardiography, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, , Lecture Notes in Computer Science Vol. 15971 (Springer, 2025) pp. 331–341.
13. T. Park, M.-Y. Liu, T.-C. Wang and J.-Y. Zhu, Semantic image synthesis with spatially-adaptive normalization, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
14. E. Perez, F. Strub, H. de Vries, V. Dumoulin and A. Courville, FiLM: Visual reasoning with a general conditioning layer, in *Proceedings of the AAAI Conference on Artificial Intelligence*, (1)2018.
15. R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge and F. A. Wichmann, Shortcut learning in deep neural networks, *Nature Machine Intelligence* **2**, 665 (2020).

16. M. Mirza and S. Osindero, Conditional generative adversarial nets (2014).
17. D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, Learning spatiotemporal features with 3D convolutional networks, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.
18. H. Reynaud, A. Vlontzos, M. Dombrowski, C. Gilligan Lee, A. Beqiri, P. Leeson and B. Kainz, D'ARTAGNAN: Counterfactual video generation, in *MICCAI*, , LNCS Vol. 134382022.
19. R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. Williamson and F. Mahmood, Synthetic data in machine learning for medicine and healthcare, *Nature biomedical engineering* **5**, 493 (2021).
20. H. Reynaud, M. Qiao, M. Dombrowski, T. Day, R. Razavi, A. Gomez, P. Leeson and B. Kainz, Feature-conditioned cascaded video diffusion models for precise echocardiogram synthesis, in *International conference on medical image computing and computer-assisted intervention*, 2023.
21. T. Chen, S. Kornblith, M. Norouzi and G. Hinton, A simple framework for contrastive learning of visual representations, in *International conference on machine learning*, 2020.
22. P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu and D. Krishnan, Supervised contrastive learning, *Advances in neural information processing systems* **33**, 18661 (2020).
23. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.
24. A. Ghorbani, D. Ouyang, A. Abid, B. He, J. H. Chen, R. A. Harrington, D. H. Liang, E. A. Ashley and J. Y. Zou, Deep learning interpretation of echocardiograms, *NPJ digital medicine* **3**, p. 10 (2020).
25. J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt and B. Kim, Sanity checks for saliency maps, *Advances in neural information processing systems* **31** (2018).
26. S. Wachter, B. Mittelstadt and C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, *Harvard Journal of Law & Technology* **31**, 841 (2017).