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: andreah@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

© 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 BYNC) 4.0 License.

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 $r_s \approx 0.25–0.39$, 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.

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.