JAMIA Open, 2024, **7(4)**, ooae106
https://doi.org/10.1093/jamiaopen/ooae106
Research and Applications

Research and Applications

Artificial intelligence-generated feedback on social signals in patient–provider communication: technical performance, feedback usability, and impact

Manas Satish Bedmutha, BS[1], Emily Bascom, MS[2], Kimberly R. Sladek, BA[1], Kelly Tobar, BS[1], Reggie Casanova-Perez, MS[3], Alexandra Andreiu, BS[1], Amrit Bhat, MS[3], Sabrina Mangal, PhD, RN[4], Brian R. Wood, MD[5], Janice Sabin, PhD, MSW[3], Wanda Pratt, PhD[3,6], Nadir Weibel, PhD[1], Andrea L. Hartzler, PhD*[,3]

[1]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA 92093, United States, [2]Department of Human Centered Design and Engineering, School of Engineering, University of Washington, Seattle, WA 98195, United States, [3]Department of Biomedical Informatics and Medical Education, School of Medicine, University of Washington, Seattle, WA 98195, United States, [4]Department of Biobehavioral Nursing and Health Informatics, University of Washington School of Nursing, Seattle, WA 98195, United States, [5]Department of Medicine, School of Medicine, University of Washington, Seattle, WA 98195, United States, [6]Information School, University of Washington, Seattle, WA 98195, United States

*Corresponding author: Andrea L. Hartzler, PhD, Department of Biomedical Informatics and Medical Education, School of Medicine, University of Washington, 850 Republican St., Bldg. C SLU 358047, Seattle, WA 98109, United States (andreah@uw.edu)

Abstract

Objectives: Implicit bias perpetuates health care inequities and manifests in patient–provider interactions, particularly nonverbal social cues like dominance. We investigated the use of artificial intelligence (AI) for automated communication assessment and feedback during primary care visits to raise clinician awareness of bias in patient interactions.

Materials and Methods: (1) Assessed the technical performance of our AI models by building a machine-learning pipeline that automatically detects social signals in patient–provider interactions from 145 primary care visits. (2) Engaged 24 clinicians to design usable AI-generated communication feedback for their workflow. (3) Evaluated the impact of our AI-based approach in a prospective cohort of 108 primary care visits.

Results: Findings demonstrate the feasibility of AI models to identify social signals, such as dominance, warmth, engagement, and interactivity, in nonverbal patient–provider communication. Although engaged clinicians preferred feedback delivered in personalized dashboards, they found nonverbal cues difficult to interpret, motivating social signals as an alternative feedback mechanism. Impact evaluation demonstrated fairness in all AI models with better generalizability of provider dominance, provider engagement, and patient warmth. Stronger clinician implicit race bias was associated with less provider dominance and warmth. Although clinicians expressed overall interest in our AI approach, they recommended improvements to enhance acceptability, feasibility, and implementation in telehealth and medical education contexts.

Discussion and Conclusion: Findings demonstrate promise for AI-driven communication assessment and feedback systems focused on social signals. Future work should improve the performance of this approach, personalize models, and contextualize feedback, and investigate system implementation in educational workflows. This work exemplifies a systematic, multistage approach for evaluating AI tools designed to raise clinician awareness of implicit bias and promote patient-centered, equitable health care interactions.

Lay Summary

Although effective communication between patients and clinicians improves the quality of health care, implicit bias in clinicians worsens communication and perpetuates inequities. Clinician implicit bias based on a patient's race manifests in nonverbal signals in their communication with patients, such as dominating the interaction. Through a 3-stage approach, we used artificial intelligence (AI) to automatically detect such nonverbal signals and deliver feedback to clinicians to raise their awareness of implicit bias and opportunities to improve their communication with patients. First, we built an AI pipeline that feasibly detected nonverbal signals, such as dominance, warmth, engagement, and interactivity, in patient–provider communication during primary care visits. Second, we engaged clinicians to design personalized feedback dashboards that visualize patterns in those signals across visits. Third, we evaluated the impact of our AI-based communication assessment and feedback system in a new set of visits. The AI demonstrated fairness with respect to patient race and was better able to detect health care provider dominance, provider engagement, and patient warmth. Stronger clinician implicit race bias was associated with less provider dominance and warmth. Overall, clinicians were interested in this promising AI approach, but suggested improvements to make it more acceptable and feasible in clinical practice and medical education.

Key words: nonverbal communication; social interaction; interpersonal relations; primary health care/patient-centered care, artificial intelligence; "prejudice/bias, implicit.".

Received: April 2, 2024; **Revised:** August 29, 2024; **Editorial Decision:** September 25, 2024; **Accepted:** October 14, 2024

Background and significance

Patient-centered communication that reflects a strong rapport/trust and empathy during patient–provider interactions is linked with care quality and outcomes.[1] In contrast, clinician implicit bias—based on a person's gender, race, ethnicity, socioeconomic status and other aspects of a person's physical characteristics and perceived identity—is associated with adverse outcomes and poor quality of care, particularly among patients from historically marginalized groups.[2,3] Implicit bias is often expressed through nonverbal patient–provider communication, such as vocal patterns and body language.[4] For example, Cooper et al.[5] found that clinicians with stronger implicit race attitudes favoring White people over Black people express greater verbal dominance (ie, greater conversational control reflected by the ratio of clinician to patient statements) with Black patients compared to White patients. Yet, assessing such nuances in the quality of clinical interactions is complex. With advances in ambient clinical intelligence, such speech-to-text translation of patient–provider interactions into clinical notes,[6,7] there is a critical opportunity to build artificial intelligence (AI) systems that extract signals associated with implicit bias from those interactions.

Social signals are expressions of one's attitude toward a social situation manifested through nonverbal cues, such as turn-taking, eye contact, and interruptions.[8,9] Social signals in patient–provider interactions, such as warmth and dominance, have traditionally been assessed using manual coding systems like Roter Interaction Analysis System (RIAS).[10] Roter Interaction Analysis System is a widely used manual coding scheme to characterize patient-centered interactions during medical encounters. The RIAS Global Affect Rating codes for signals associated with affect, such as warmth, dominance, interactivity, and engagement, on a scale from low (1) to high (6). Although RIAS is one of the most common systems for studying social signals in patient–provider interactions, its application to annotate patient visits is labor-intensive. Annotation requires trained human observers to code visits for each social signal manually. This process can require hours to analyze visits, which limits scalability.

In contrast to traditional manual assessment, automated social signals processing (SSP) methods[9] show promise for understanding social interactions through machine analysis of nonverbal behavior.[11–14] Outside of health care, SSP has been used to assess conversational dynamics in several contexts,[8] ranging from hiring negotiations[15] to team cohesion.[16] Automated assessment of clinical conversations can ensure that clinicians receive efficient feedback about the quality of communication without manual assessment. Advances in SSP have created opportunities to raise clinician awareness of communication patterns with patients. For example, prior work on the "EQClinic"[12,13] and "ReflectLive"[14] systems describe the use of SSP to assess and improve nonverbal communication with patients during clinical consultations. Prior work on "Entendre" demonstrated clinician acceptance of the use of social signals as a communication feedback mechanism based on nonverbal cues.[17] However, the feasibility of SSP for detecting signals associated with implicit bias in patient–provider communication has not been explored. There is an opportunity for AI to automate this assessment and facilitate research on implicit bias in health care.

There is a need to explore the use of AI to develop SSP tools that automatically assess and deliver feedback to clinicians about their communication with patients. Potential generalizability of such models could help to demonstrate the potential for improving clinician awareness of implicit bias in their communication and fostering health care equity. We describe our approach and the opportunity for equity-focused AI in the context of ambient clinical computing.

Objective

We report on the 3-staged investigation of our AI-generated communication assessment and feedback system to: (1) assess technical AI performance, (2) design usable AI-generated communication feedback for clinicians' workflow, and (3) evaluate the impact of our AI approach.

Stage 1: Technical performance of AI

In Stage 1, we annotated social signals in audio-recorded primary care visits and used the annotations to build and test the performance of a machine-learning pipeline that automatically predict social signals from patient–provider interactions. By investigating the capability of AI models, we addressed the following research question (RQ): RQ1: Can meaningful social signals that reflect patient-centered care and potential implicit bias be feasibly extracted from audio-recorded nonverbal patient–provider interactions? Preliminary results were reported as a Stage 1 AMIA AI showcase presentation.[18] Our previous work[18,19] focused on a smaller subset of signals in a reduced dataset.

Methods

We built an SSP pipeline by annotating 12 social signals in patient–provider interactions in the secondary use of 145 recorded primary care visits from the "Establishing Focus" (EF) study.[20] The EF study investigated the impact of a collaborative agenda setting intervention on the quality of primary care encounters. Although the intervention was found not to impact visit content or time use,[21] the study generated data from 2002 to 2006 from audio-recorded study visits and post-visit questionnaires from patients and providers we used for secondary analysis. See Supplementary material S1 for details on social signals and participant demographics from that study. Study procedures were approved by the University of Washington (UW) Institutional Review Board (IRB) #00005436. The 12 social signals were labeled by 3 trained coders based on RIAS Global Affect Ratings[10]: dominance, attentiveness, warmth, engagement, empathy, respect, interactivity, irritation, nervousness, *hurriedness, **sadness, and **emotional distress (**provider only,* **patient only).

To capture the granularity of nonverbal communication behaviors throughout visits, we used the "thin slice" approach. A thin slice refers to "a brief excerpt of expressive behavior sampled from the behavioral stream".[22] When applied to thin 1-minute slices of recorded medical encounters, RIAS was shown to be an efficient predictor of visit outcomes while capturing local variation in communication quality.[23] Informed by this approach, we annotated social signals in thin slices of recorded visits. We split each visit into successive thin slices of 3-minute intervals, resulting in 690 slices across the 145 visits, each locally annotated by 3 trained coders. Informed by domains of RIAS Global Affect

Ratings,[10] coders rated each slice across the 12 social signals on a scale from 1 (low) to 6 (high). For irritation, nervousness, sadness, and emotional distress, a rating of 1 was assigned when there were no signs of the affect. Because these 4 social signals were rarely rated with a score other than 1, we excluded all 4 from our models. For the remaining 8 social signals, a rating of 3 was considered average affect. Because few slices deviated significantly from the average (score of 3), we observed extreme class imbalance where most slices fell near the mean score for each signal. To improve granularity in our modeling, we clustered ratings for each social signal into a 3-class system: "low" (rating below 3), "average" (rating of 3), and "high" (rating above 3).

Since vocalic nonverbal cues have been empirically associated with social signals observed in patient–provider interactions,[17] we investigated their use as features in our AI models. Given prior work linking verbal dominance with implicit bias[5] and patient-centered communication,[17] we explored audio features, such as talk-time, interruptions, and conversational turn-taking.[19] These audio features rely on knowing who spoke when, so we first distinguished speakers with speech diarization using diart.[24] We then extracted statistical features, such as mean, min, max, and SD of turns; interruptions; and pauses, to drive our AI model. We corrected the class imbalance by oversampling the minority labels to equal the number of majority class labels using SMOTE.[25] We then evaluated all models using leave-one-subject-out cross-validation. Figure 1 shows our social signal extraction pipeline.

Results

Despite the clustering with respect to average (rating of 3), we found that for most social signals there were few of the 690 slices annotated with ratings below 3. Thus, we excluded the "low" class from modeling. This exclusion transformed our task into a binary classification that can distinguish between "average" and "high" ratings. For this classification task, we trained different machine-learning models, namely logistic regression (LR), support vector machine (SVM), decision tree classifier (DTC), random forests (RF), and gradient boosted trees (GBDT). We compared the performance with a majority label (baseline) predicting a dummy classifier. Since our labels were extremely imbalanced (the ratio of minority to majority class samples had a mean = 0.192 and SD = 0.178), we evaluated all models based on F1 scores, given their robustness to class imbalance.[26] We chose the macro-F1 score since it considers performance on both classes equally important.

Table 1 shows the performance averaged over all splits of our cross-validation. Seven signals beat the baseline classifier: provider dominance, provider warmth, patient warmth, provider engagement, patient engagement, provider interactivity, and patient interactivity. For the remaining signals, the majority classifier either had scores higher than 0.5, which may be attributed to the small number of or no samples from the minority class in the held-out sets. Despite the class imbalance with few slices deviating from average in the coded dataset, our models demonstrate it is feasible to identify social signals in nonverbal patient–provider communication, such as dominance, warmth, engagement, and interactivity.

Stage 2: Usability and workflow of AI

In Stage 2, we engaged 24 primary care providers in design sessions to investigate the usability and workflow integration of AI-generated feedback on patient–provider communication. Through this work, we addressed RQ2: What are primary care providers' design preferences to improve the usefulness and usability of tools that provide AI-generated feedback about patient–provider communication during or after clinical encounters? Preliminary results were reported as a Stage 2 AMIA AI showcase presentation.[27]

Methods

Based on prior work,[28,29] we depicted 3 design concepts as wireframes for participant feedback on implicit patient–provider communication biases during design critique sessions: data-driven feedback (eg, via traditional dashboards), real-time digital nudge (eg, alerts and ambient feedback), and guided reflection (eg, via conversational agents and human mediation) (Figure 2). The wireframes visualize nonverbal communication features between patients and providers during clinical interactions, such as body language, eye contact, and speech patterns, with the goal of raising clinician awareness of potential implicit bias in their communication with patients.

Each design critique session was conducted remotely via Zoom by 2 researchers, lasted 45-60 minutes, and consisted of participants discussing desired wireframe features, how the depicted tools might fit into their clinical workflow with a focus on how to make the feedback most usable, and anticipated personal and institutional barriers for implementing tools like the ones depicted in the wireframes. Sessions were analyzed using qualitative content analysis to characterize and compare participants' relative preferences among the 3 design concepts. Study procedures were reviewed by UW IRB and determined exempt (#00008252).

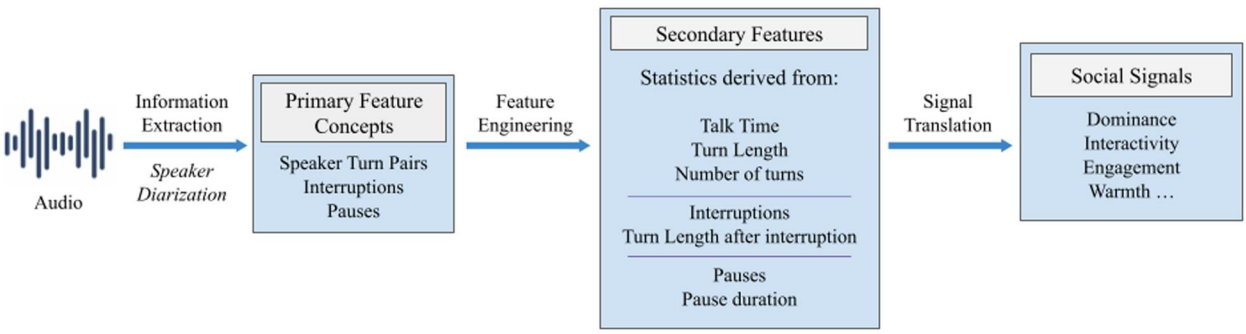

Figure 1. Social signal processing pipeline to predict levels of social signals based on audio-recorded patient–provider interaction.

Table 1. Model performance outputs averaged over leave-one-subject-out cross-validation.

Social signal	Rating < 3	Rating of 3	Rating > 3	Majority (baseline)	LR	SVM	DTC	RF	GBDT
Provider dominance	0	231	459	0.230	**0.516**	0.491	0.449	0.489	0.489
Patient dominance	7	614	69	0.686	0.452	0.556	0.427	0.438	0.450
Provider attentiveness	0	607	83	0.640	0.450	0.451	0.405	0.415	0.431
Patient attentiveness	0	633	57	0.689	0.472	0.494	0.437	0.440	0.446
Provider warmth	4	450	236	0.397	0.492	0.492	0.486	0.486	**0.519**
Patient warmth	0	472	218	0.406	0.453	0.441	0.427	**0.495**	0.451
Provider engagement	0	128	562	0.116	0.436	**0.477**	0.433	0.470	0.470
Patient engagement	2	88	600	0.097	0.453	0.469	0.456	0.469	**0.469**
Provider empathy	2	649	39	0.571	0.490	0.482	0.432	0.458	0.457
Patient empathy	0	688	2	0.916	0.916	0.916	0.916	0.916	0.916
Provider respect	5	673	12	0.831	0.831	0.831	0.831	0.831	0.831
Patient respect	0	686	4	0.873	0.873	0.873	0.873	0.873	0.873
Provider interactivity	0	536	154	0.438	0.444	0.435	0.406	**0.474**	0.461
Patient interactivity	0	532	158	0.435	0.466	0.450	0.465	**0.493**	0.490
Provider hurry	0	686	4	0.831	0.831	0.831	0.831	0.831	0.831

Bold indicates models that beat the majority baseline classification model.
Abbreviations: DTC, decision tree classifier; GBDT, gradient boosted trees; LR, logistic regression; RF, random forests; SVM, support vector machine.

Results

Table 2 summarizes the characteristics of the 24 participants. From our content analysis regarding the 3 design concepts of data-driven dashboards, real-time nudges, and guided reflections, we found that participants preferred seeing personalized communication patterns depicted in data-driven dashboards that included in-depth patient demographics and trends across time, rather than digital nudges or guided reflection. Fifteen participants (63%) found it insufficient to only display raw nonverbal cues (eg, eye contact, interruptions) when describing their communication with patients because this presentation disregards the context in which the conversation occurred. PCP02 elaborates, stating, "Not all interruptions are bad. The reality of the visit is that I've got 15 minutes... sometimes [interruptions] are what you do to get the information." Participants desired educational materials and further context about why raw nonverbal communication behaviors matter, indicating that social signals may be a more effective way of providing this feedback. They also expressed a desire to improve the tool's usability with quick-tip resources presented in a way that encourages clinicians wellness and helps them improve their communication. Participants also described personal and institutional barriers they anticipated for implementing AI-driven feedback on communication. Twenty participants (83%) were worried about having enough time to meaningfully engage with the tool, especially if clinicians are not incentivized or compensated for this time as continuing education or diversity, equity, and inclusion training. While these findings are highlights that prompted our design choices, Bascom et al.[30] discuss Stage 2 results in depth.

Stage 3: Evaluation of AI impact

In Stage 3, we assessed our AI models on a new cohort of 108 primary care visits that we prospectively collected from 15 clinicians across 4 primary care clinics at UW Medicine and University of California San Diego (UCSD) Health. We answered 2 RQs—RQ3.1 How well do the AI models generalize from the dataset they were trained on? and RQ3.2 What are clinicians' perspectives on the acceptability, feasibility, and barriers to implementing the AI-based approach in clinical practice? Preliminary results were reported as a Stage 3 AMIA AI showcase presentation.[31]

Methods

To evaluate how the developed AI models from Stage 1 generalize, we tested our algorithms against a new dataset comprising a new cohort of primary care visits. This test dataset consists of recorded patient–provider interactions. For this study, we chose clinics with diverse patient populations who historically experience discrimination in health care, Black, Indigenous, and people of color (BIPOC) and lesbian, gay, bisexual, transgender, and queer people (LGBTQ+). Clinicians were recruited through convenience sampling at UW Medicine and UCSD Health. Patients of enrolled clinicians were recruited in advance of scheduled visits. After obtaining consent from the patient and clinician, each visit was recorded for analysis. We characterized implicit bias in clinician participants with the Implicit Association Test (IAT). The IAT is a widely used indirect measure of implicit social cognition measuring the relative strength of positive and negative associations toward one social group compared to another.[32] It produces a score that identifies the strength of implicit bias.[33] We used 2 IATs, the race (Black/White) IAT and sexuality (gay/straight) IAT. We chose to do our analysis using the race IAT because we found a moderate pro-White bias among our sample and little to no implicit gay–straight bias.

Four coders trained by the trained coders followed the same procedure as Stage 1 to manually code visits for social signals. Mean social signal ratings computed for each provider were correlated with race IAT scores using Kendall's Tau-b, a nonparametric method.

Each visit was then processed with our SSP pipeline for the 4 social signals that beat baseline (Table 1). Since the pipeline is audio-based, we used the audio component of these recorded visits for evaluation. We evaluated model performance with macro-F1, which is a strict metric toward class imbalance. Given the association of communication measured by RIAS with implicit race bias,[5] we also analyzed social signal prediction for fairness with respect to patient race.

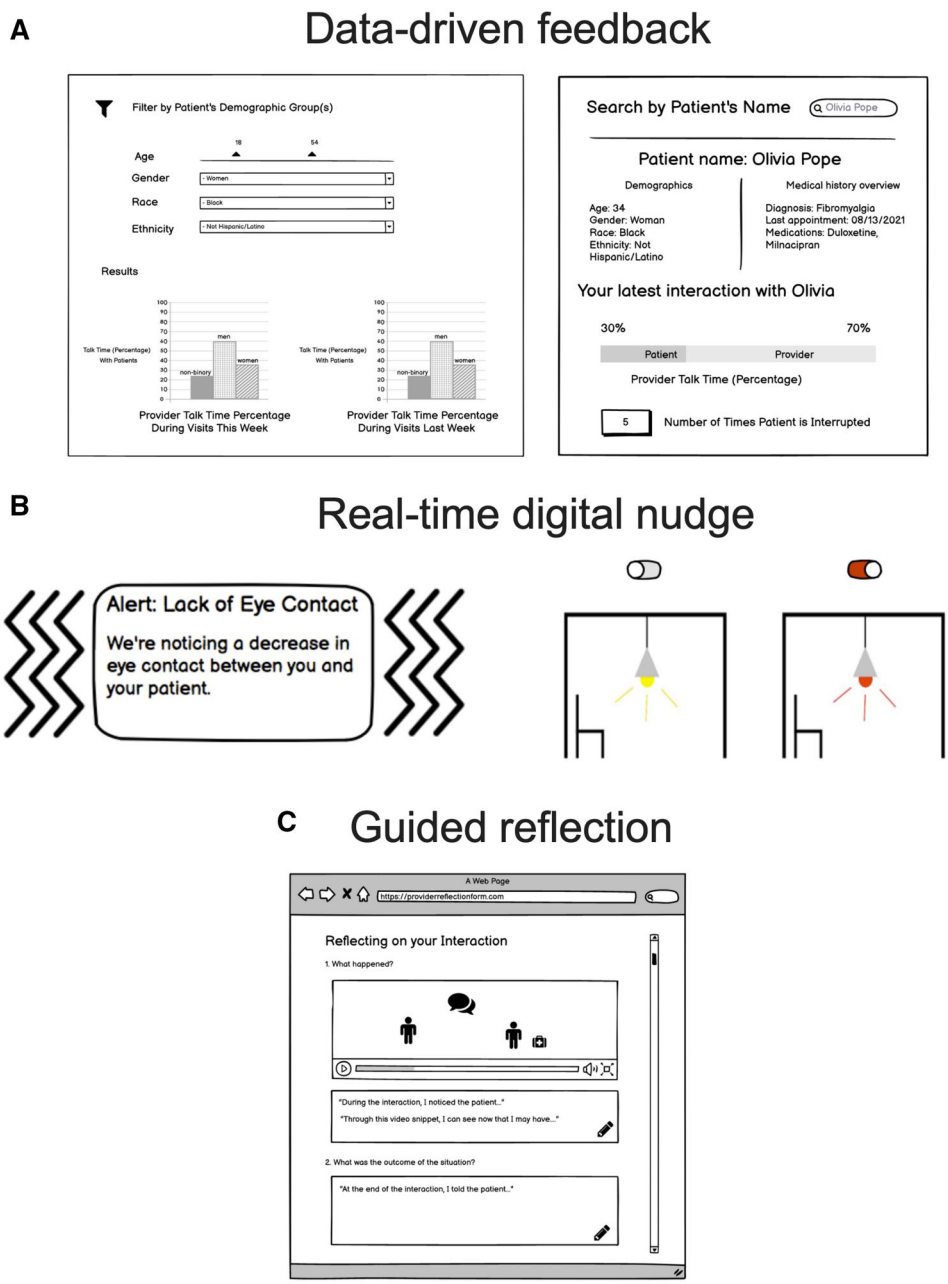

Figure 2. Wireframes for clinician feedback on implicit patient–provider communication biases during design critique sessions: (A) Data-driven feedback delivered via quantitative metrics on a dashboard. (B) Real-time digital nudge feedback delivered via smartwatch alerts or changes in ambient room lighting. (C) Guided reflection feedback delivered via clips of recorded interactions that prompt self-reflection with conversational agents or human mediation.

We calculated the demographic parity ratio across patients who reported White race compared to those who reported race other than White.

After recording at least 7 visits, we invited each clinician for a 1-hour Zoom interview to receive visual feedback on their communication in a personalized feedback report (Figure 3). During the interview, participants reviewed the report for feedback about its acceptability and shared their perceptions about the feasibility of our approach to automated communication assessment and feedback into their workflow, considering potential barriers to implementing in clinical practice. The data was analyzed deductively to explore the 3 a priori dimensions: acceptability, feasibility, and implementation considerations. Study procedures were approved by IRBs at UW (#00012188) and UCSD (#210932).

Results

Cohort description

The cohort included 15 providers and 108 patients. See Supplementary material S1 for participant demographics. Figure 4 shows the distribution of the 12 social signals rated by trained coders when observing patients and providers across visits. Except for provider hurriedness, social signals associated with less patient-centeredness (ie, irritation, nervousness, patient sadness, and distress) were low with mean

Table 2. Design session participant characteristics, including demographics and clinical experience.

	24 (100%)
Age—mean (SD), range	45 (11), 31–68
Gender	
Woman	12 (50.0%)
Man	12 (50.0%)
Race—count (%)	
White	16 (66.7%)
Black or African American	2 (8.3%)
Asian: Chinese, Asian Indian	2 (8.3)%
Another race (self-described): "Mixed," "Latina"	2 (8.3%)
Decline to state	2 (8.3%)
Ethnicity—count (%)	
Hispanic or Latino/a/x/e	2 (8.3%)
Not Hispanic or Latino/a/x/e	20 (83.3%)
Decline to state	2 (8.3%)
Self-selected identity—count (%)	
BIPOC: Black, Indigenous, and people of color	5 (20.8%)
LGBTQ+: lesbian, gay, bisexual, trans, queer, and other identities	3 (12.5%)
LATINX	2 (8.3%)
None of these	13 (54.2%)
Decline to state	1 (4.2%)
Clinical role—count (%)	
Nurse practitioner (NP)	2 (8.3%)
Doctor of osteopathic medicine (DO)	1 (4.2%)
Medical doctor (MD)	21 (87.5%)
Number of years in role—mean (SD), range	16 (12), 2–42
Approximate panel size (number of patients)—mean (SD), range	369 (457), 0–2000

ratings ranging from 1.0 to 1.1. In contrast, the remaining social signals are moderate, with mean ratings for patients ranging from 3.1 to 3.5 and mean ratings for providers slightly higher ranging from 3.3 to 3.8.

Performance of AI models

Evaluation of our AI models on the test dataset showed potential in using nonverbal vocal cues to generate social signal predictions. Table 3 shows the performance of our models. Since our models are binary classifiers and the F1 scores are computed in a class weighted manner, we use 0.5 as a threshold. Three models show stronger potential to generalize outside the distribution they were trained: provider dominance, provider engagement, and patient warmth. Based on balanced accuracy threshold of 0.5, 5 of 7 models performed better than chance. All models show a demographic parity ratio higher than 0.8 in prediction performance between people who identified as White versus race other than White, satisfying the "rule of four-fifths" threshold for fairness.[34] However, further feature-level analysis is warranted for signals where the models could not generalize.

Association of social signals with implicit bias

On average, provider implicit race attitudes showed a slight preference for White over Black people ($D = 0.29$, SD = 0.42, range −−0.56 to 0.98) that was significantly different from 0 (ie, no race bias) ($P = .02$). Table 4 shows the correlations between social signals and provider IAT scores. Although correlations were not statistically significant, some were moderate in strength with medium to large effect sizes for Kendall's Tau. In particular, provider implicit race bias was moderately associated with greater patient sadness and distress, and less provider dominance, attentiveness, warmth, engagement, and interactivity. Among these signals, our AI models can capture changes in provider dominance and warmth.

Exit interviews

We describe themes related to the acceptability and feasibility of integrating our AI-based communication assessment and feedback approach that surfaced in exit interviews with 14 of the 15 clinicians who contributed to the cohort for our study in Stage 3.

Acceptability

Most participants expressed interest in our approach to understand their communication behavior since opportunities for feedback were rare after residency. However, participants suggested several design enhancements to make AI-driven communication feedback more acceptable. For example, participants suggested embedding clips from recorded visits into the line graphs to help them remember the context of visits. Some participants suggested adding comparison with peers to feedback, but others had varied opinions. One participant proposed friendly competition as a motivating factor for improving communication skills, while another participant thought that feedback might be "scary" for clinicians who are vulnerable to peer pressure and may befit from a "safety context" and "strengths-based lens" for feedback.

Feasibility

Participants suggested that willingness to record visits may depend on clinician preference and organizational policy. To some participants, the recording setup in exam rooms felt unnatural and lacked portability requiring successive visits in the same exam room, hampering scalability. Although some participants initially worried about the "Hawthorne effect,"[38] others reflected that their behavior did not change, and they quickly forgot about the recording. Another barrier to recording was privacy and consent. Some participants were worried that patients might not consent to be recorded, fearing a lack of privacy. Even among those who would consent, one participant expressed concern that recording might inhibit patients from discussing issues freely. Clinician time spent reviewing feedback was a common concern. One participant mentioned that they might only look at feedback once or twice a month since reviewing a visit could be a significant time burden. They recommended selecting specific visits to record (eg, "challenging visits") as opposed to recording all visits, for example.

Implementation barriers

Several participants described barriers to implementing our AI approach in clinical practice. Given the challenges with recording in exam rooms, some participants also suggested telehealth as a context that might facilitate implementation. Four participants suggested medical education as an alternative use case to clinical practice implementation. For example, they proposed using standardized patients with whom clinicians could interact to improve communication skills and get feedback within a safe training context. A couple of participants suggested using our AI approach early in medical training, such as incorporating into Observed Structured

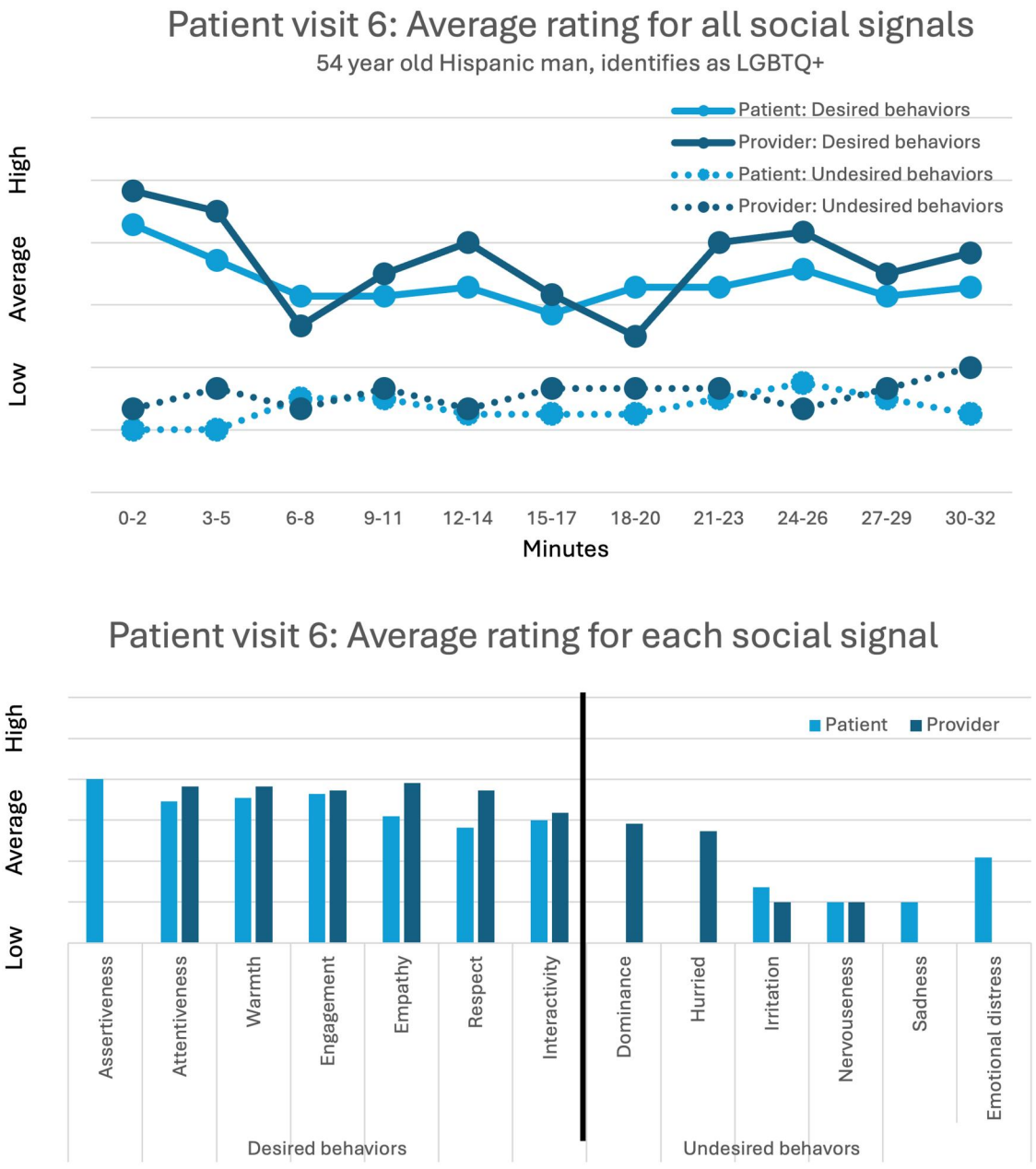

Figure 3. Example of personalized feedback report visualizing social signals across a provider's visit with line graphs (top) and bar charts (bottom). The line graph shows the average of all social signal ratings for every 3-minute slice of the visit for patients and providers. The bar chart shows the average rating for each social signal across the entire visit. Dominance, hurry, irritation, nervousness, sadness, and emotional distress were considered "undesired" communication behaviors and the remaining social signals were considered "desired" behaviors. The vertical separator distinguishes the desired (bar chart, left) and undesired (bar chart, right) communication behaviors.

Clinical Examination (OSCE) exams[39] during medical school.

Discussion

Principal findings

Through a 3-stage approach, we explored the potential of AI-generated communication assessment and feedback in primary care. Through our technical performance study (Stage 1), our AI models demonstrate it is feasible to identify social signals in nonverbal patient–provider communication, such as dominance, warmth, engagement, and interactivity. Through our design sessions with clinicians to visualize AI-generated communication feedback (Stage 2), AI-generated communication feedback delivered in personalized dashboards was preferred over digital nudges and guided repletion. A key learning was that raw nonverbal cues (eg, counts of interruptions, gaze drifts, etc.) were difficult for clinicians to interpret and that social signals may be an effective alternative feedback mechanism in data-driven dashboards. Finally, through our impact evaluation in a new cohort (Stage 3), we found 3 AI models to show stronger potential to generalize, namely provider dominance, provider engagement, and patient warmth. While all AI models demonstrated fairness through demographic parity, provider dominance and provider warmth were negatively associated with stronger clinician implicit race bias. Although participants expressed interest in the approach, they recommend design

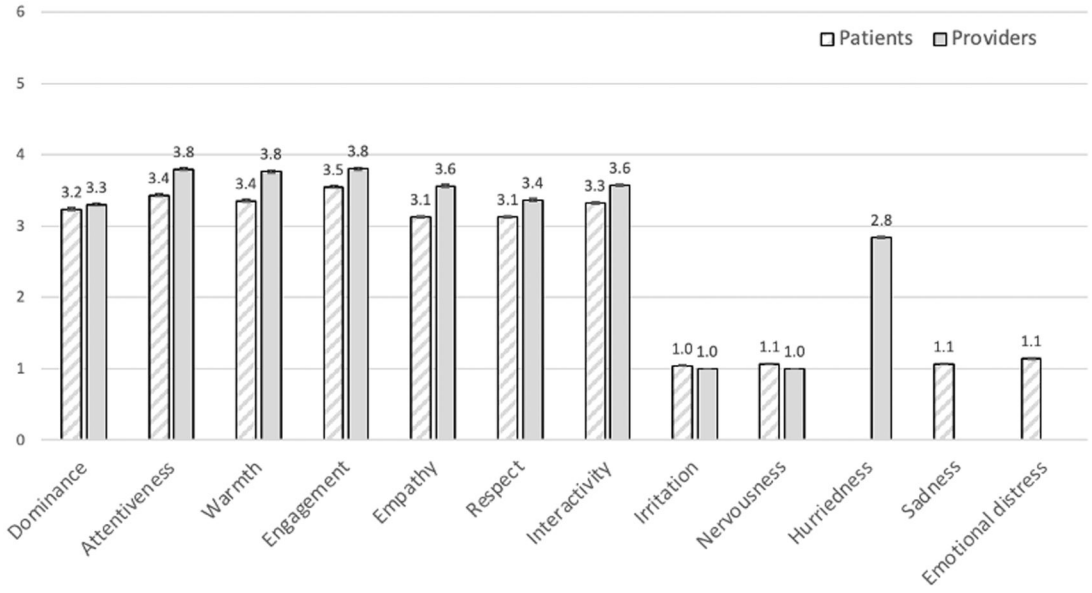

Figure 4. Distribution of mean ratings (1 "low" to 6 "high") across visits. Error bars show SE. For each visit, we calculated the mean rating for each social signal across slices for patients and providers. We used this distribution to calculate the mean rating for each social signal across the 108 visits.

Table 3. Performance of AI models trained on the EF study dataset when evaluated on the test dataset.

Social signal	F1 score (macro) (↑)	Balanced accuracy	Demographic parity ratio (∼1)
Provider dominance	**0.54**	0.577	0.995
Provider engagement	**0.51**	0.513	0.959
Patient engagement	0.32	0.496	0.941
Provider interactivity	0.43	0.511	0.881
Patient interactivity	0.36	0.517	0.918
Provider warmth	0.31	0.448	0.994
Patient warmth	**0.51**	0.512	0.911

Based on F1 scores higher than 0.5, the three models in bold show the potential to generalize. Balanced accuracy scores higher than 0.5 show that models perform better than chance. A demographic parity ratio higher than 0.8 is treated as the model being fair toward the sensitive attributes, here patient race.
Abbreviations: AI, artificial intelligence; EF, establishing focus.

improvements to enhance acceptability, feasibility, and implementation in alternative contexts, such as telehealth and medical education.

Limitations and future work

Our work is an early attempt to envision a functional AI-based communication assessment and feedback system and has limitations, selection bias among patients and clinicians who consented to have visits recorded could have resulted in capturing more effective communicators. Although our models suggest it is possible to identify certain social signals, further work is necessary to identify these signals reliably. Another limitation was the lack of variability in social signal ratings which limited the granularity of the SSP pipeline. Future work should also expand on classical machine-learning models to more advanced approaches, leveraging audio transformers[40] or foundation models.[41,42]

In our design sessions, most participating clinicians were non-Hispanic White cisgender individuals, which could have resulted in overlooking barriers or design ideas that clinicians who are part of marginalized racial, ethnic, or gender groups find important when receiving feedback about communication. Future work should explore how the perceptions and design suggestions of participants from diverse demographic backgrounds, clinic types, and patient populations served might vary to provide a deeper contextual lens. This lens is essential for ensuring effective interventions that promote equity and benefit clinicians and patients from all backgrounds.

Finally, through our impact evaluation, we found that while some of our models performed better than a baseline on a new cohort (test dataset), several social signals did not generalize. This could be due to differences in the distribution of features and labels. Future work should focus on developing novel features that can generalize out of distribution. Further, future work should investigate implementation in contexts that might enhance feasibility and acceptability to clinicians, such as telehealth and medical education.

Communication assessment and feedback in medical education

Participants across study stages recommended the use of communication assessment and feedback in medical education from medical school to residency to midcareer professionals. Since feedback may be more common during medical school and residency, clinical communication skills training

Table 4. Kendall's Tau correlation (τ_b) between social signals and provider race IAT scores.

Dominance	0.16
Patient	**−0.26***
Provider	
Attentiveness	−0.05
Patient	**−0.24**
Provider	
Warmth	−0.09
Patient	**−0.28***
Provider	
Engagement	−0.07
Patient	**−0.28**
Provider	
Empathy	−0.01
Patient	0.01
Provider	
Respect	−0.08
Patient	−0.03
Provider	
Interactivity	−0.03
Patient	**−0.22**
Provider	
Patient irritation	0.11
Patient nervousness	0.19
Provider hurriedness	−0.07
Patient sadness	**0.20**
Patient distress	**0.27**

Because the nonverbal cues and social signals were not normally distributed (ie, data was skewed), we used Kendall's Tau (τ_b) correlation for its performance advantage over Spearman correlation.[35] Interpretation of Kendall's Tau (τ_b) correlation: correlation strength—weak (0.1-0.19), moderate (0.20-0.29), strong (0.30 and greater)[36]; effect size—small (0.06), medium (0.15), large (0.24).[37] Bold indicates moderate Kendall's Tau correlations with medium to large effect sizes. Asterisks indicate social signals predicted by AI models in Stage 3.
Abbreviations: AI, artificial intelligence; IAT, implicit association test.

could be an avenue to introduce AI-based communication assessment and feedback. For example, AI tools could be used with standardized patients to practice communication skills over time. For example, these simulated environments can mimic cognitive stressors that may trigger implicit race bias.[43] Integrating AI-based assessment and feedback tools into simulated training could advance implicit bias recognition and management systems.[44,45]

Toward personalization in modeling and feedback

Our models tend to fare well within the data distribution they were trained on. Recent work in activity sensing proposes that injecting small amounts of data from an individual can greatly improve the accuracy of predictions for that individual.[46] Future research should explore whether injecting slices of manual coding can improve performance. Further, participants sought more context and personalized insights on their communication behavior for self-reflection. Prior work has shown utility[12–14] and acceptability[47] to visualize raw nonverbal cues, such as counts of interruptions and turn-taking, as means for communication feedback for clinicians. Personalized dashboards with more context on these cues should be further explored.

Conclusion

We investigated how AI can be used to understand patient–provider interactions in primary care visits using social signals and how such systems can be designed to provide visual feedback for integration into clinical workflows. Using a primary care dataset, we demonstrated the feasibility of AI to predict changes in social signal levels from nonverbal communication during visits. We then designed social signal feedback for provider-facing tools that visualize those communication patterns in personalized dashboards. Finally, we evaluated the impact of our AI-based system on a new cohort of visits, garnering formative feedback from clinicians who experienced the approach. This work is an example of systematic and multistaged methods for evaluating AI tools and a jumping-off point for using equity-focused AI to develop SSP tools that deliver clinicians feedback on improving communication with patients.

Acknowledgments

We would like to thank our Clinical Champions and the rest of the UnBIASED project team, clinical champions for their invaluable contributions to our conversations about this work: Hollie David, Lisa Dirks, Veen Doski, Niyat Efrem, Colleen Emmeneger, Charles Goldberg, Linda Hill, J.P. Lopez, Sean Mooney, Tom Payne, Steven Rick, Debra Roter, Cynthia E. Schairer, Angad Singh, Pooja Thorali, Anuujin Tsedenbal, Jeremiah (J.W.) Wiebe-Anderson, and staff in our partnering clinics. Most importantly, we would like to thank study participants for sharing their input and experiences.

Author contributions

Manas Satish Bedmutha, Emily Bascom, Reggie Casanova-Perez, and Sabrina Mangal contributed to data acquisition, data analysis and interpretation, drafting of the manuscript, and critical revision of the manuscript for important intellectual content. Kimberly R. Sladek, Kelly Tobar, and Alexandra Andreiu contributed to data acquisition and critical revision of the manuscript for important intellectual content. Amrit Bhat contributed to data analysis and interpretation, and critical revision of the manuscript for important intellectual content. Brian R. Wood, Wanda Pratt, and Nadir Weibel contributed to conception and design, and critical revision of the manuscript for important intellectual content. Janice Sabin contributed to conception and design, data analysis and interpretation, and critical revision of the manuscript for important intellectual content. Andrea L. Hartzler contributed to conception and design, data acquisition, data analysis and interpretation, drafting of the manuscript, and critical revision of the manuscript for important intellectual content.

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

This research was supported by the National Library of Medicine at the National Institutes of Health 1R01LM013301. S. M. was supported by the National Institute of Nursing Research at the National Institutes of Health T32NR016913.

Conflicts of interest

The authors have no competing interests.

Data availability

The data underlying this article cannot be shared publicly due to the privacy of individuals that participated in the study. The data will be shared on reasonable request to the corresponding author.

References

1. Zolnierek KB, DiMatteo MR. Physician communication and patient adherence to treatment: a meta-analysis. *Med Care*. 2009;47:826-834.
2. FitzGerald C, Hurst S. Implicit bias in healthcare professionals: a systematic review. *BMC Med Ethics*. 2017;18:Article 19.
3. Hall WJ, Chapman MV, Lee KM, et al. Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. *Am J Public Health*. 2015;105: e60-e76.
4. Hagiwara N, Lafata JE, Mezuk B, Vrana SR, Fetters MD. Detecting implicit racial bias in provider communication behaviors to reduce disparities in healthcare: challenges, solutions, and future directions for provider communication training. *Patient Educ Couns*. 2019;102:1738-1743.
5. Cooper LA, Roter DL, Carson KA, et al. The associations of clinicians' implicit attitudes about race with medical visit communication and patient ratings of interpersonal care. *Am J Public Health*. 2012;102:979-987.
6. Tran BD, Latif K, Reynolds TL, et al. "Mm-hm," "Uh-uh": are non-lexical conversational sounds deal breakers for the ambient clinical documentation technology? *J Am Med Inform Assoc*. 2023;30:703-711.
7. Yim WW, Fu Y, Ben Abacha A, Snider N, Lin T, Yetisgen M. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Sci Data*. 2023;10:586.
8. Vinciarelli A, Pantic M, Bourlard H. Social signal processing: survey of an emerging domain. *Image and Vision Computing*. 2009;27:1743-1759.
9. Burgoon JK, Magnenat-Thalmann N, Pantic M, Vinciarelli A, eds. *Social Signal Processing*. Cambridge University Press; 2017.
10. Roter D, Larson S. The Roter interaction analysis system (RIAS): utility and flexibility for analysis of medical interactions. *Patient Educ Couns*. 2002;46:243-251.
11. Riku A, Yakura H. REsCUE: a framework for REal-time feedback on behavioral CUEs using multimodal anomaly detection. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Paper 572. ACM; 2019:1-13.
12. Liu C, Scott KM, Lim RL, Taylor S, Calvo RA. EQClinic: a platform for learning communication skills in clinical consultations. *Med Educ Online*. 2016;21:31801.
13. Wu K, Liu C, Calvo RA. Automatic nonverbal mimicry detection and analysis in medical video consultations. *Int J Hum Comput Interact*. 2020;36:1379-1392.
14. Faucett HA, Lee ML, Carter S. I should listen more: real-time sensing and feedback of non-verbal communication in video telehealth. *Proc ACM Hum Comput Interact*. 2017; 1:1-9.
15. Curhan J, Pentland A. Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. *J Appl Psychol*. 2007;92:802-811.
16. Lehmann-Willenbrock N, Hung H. A multimodal social signal processing approach to team interactions. *Organ Res Methods*. 2024; 27:477-515.
17. Hartzler AL, Patel RA, Czerwinski M, et al. Real-time feedback on nonverbal clinical communication. *Methods Inf Med*. 2014;53:389-405.
18. Bedmutha MS, Sladek K, Bascom E, et al. Extracting meaningful social signals from patient–provider interactions. In: *Proceedings of the AMIA 2023 Informatics Summit: AI Showcase (Stage 1)*, March 14, 2023. Seattle, WA: AMIA; 2023:941-942.
19. Bedmutha MS, Tsedenbal A, Tobar K, et al. ConverSense: an automated approach to assess patient–provider interactions using social signals. *Proc SIGCHI Conf Hum Factor Comput Syst*. 2024;2024:448.
20. AHRQ. Effects of Establishing Focus in the Medical Interview (R01HS 013172 PI Lynne Robins). AHRQ; March 2006. Accessed October 9, 2024. https://www.ahrq.gov/sites/default/files/2024-07/robins-report.pdf
21. Brock DM, Mauksch LB, Witteborn S, Hummel J, Nagasawa P, Robins LS. Effectiveness of intensive physician training in upfront agenda setting. *J Gen Intern Med*. 2011; 26:1317-1323.
22. Ambady N, Bernieri FJ, Richeson JA. Toward a histology of social behavior: judgmental accuracy from thin slices of the behavioral stream. *Adv Exp Soc Psychol*. 2000;32:201-271.
23. Roter DL, Hall JA, Blanch-Hartigan D, Larson S, Frankel RM. Slicing it thin: new methods for brief sampling analysis using RIAS-coded medical dialogue. *Patient Educ Couns*. 2011;82:410-419.
24. Coria JM, Bredin H, Ghannay S, Rosset S. Overlap-aware low-latency online speaker diarization based on end-to-end local segmentation. In: *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, December 13, 2021. IEEE:1139-1146.
25. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-357.
26. PlÖtz T. Applying machine learning for sensor data analysis in interactive systems: common pitfalls of pragmatic use and ways to avoid them. *ACM Comput Surv*. 2021;54:1-25.
27. Bascom E, Bedmutha MS, Casanova-Perez R, et al. Healthcare providers' perspectives on implicit communication bias feedback. In: *Proceedings of the AMIA 2023 Clinical Informatics Conference: AI Showcase (Stage 2)*, Chicago, IL. May 2023.
28. Dirks L, Beneteau E, Sabin J, et al. Battling bias in primary care encounters: informatics designs to support clinicians. In: *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. ACM; April 27, 2022:1-9.
29. Loomis A, Montague E. Human-centered design reflections on providing feedback to primary care physicians. In: Kurosu M, ed. *Proceedings of Human-Computer Interaction. Design and User Experience Case Studies: Thematic Area, HCI 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event*, July 24–29, 2021. Proceedings, Part III. Vol. 23. Springer International Publishing; 2021:108-118.
30. Bascom E, Bedmutha M, Casanova-Perez Tobar K, et al. Designing communication feedback systems to reduce healthcare providers' implicit biases in patient encounters. *Proc SIGCHI Conf Hum Factor Comput Syst*. 2024;2024:452.
31. Bedmutha MS, Bhat A, Mangal S, et al. Towards inferring implicit bias in clinical interactions using social signals. In: *Proceedings of the AMIA Annual Symposium: AI Showcase (Stage 3). November 2023*, New Orleans, LA.
32. Greenwald AG, McGhee DE, Schwartz JL. Measuring individual differences in implicit cognition: the implicit association test. *J Pers Soc Psychol*. 1998;74:1464-1480.
33. Greenwald AG, Nosek BA, Banaji MR. Understanding and using the implicit association test I: an improved scoring algorithm. *J Pers Soc Psychol*. 2003;85:197-216.
34. Holzer H, Neumark D. Assessing affirmative action. *J Econ Lit*. 2000;38:483-568.
35. Arndt S, Turvey C, Andreasen NC. Correlating and predicting psychiatric symptom ratings: Spearman's r versus Kendall's tau correlation. *J Psychiatr Res*. 1999;33:97-104.
36. Gilpin AR. Table for conversion of Kendall's Tau to Spearman's Rho within the context of measures of magnitude of effect for meta-analysis. *Educ Psychol Meas*. 1993;53(1):87-92.

37. Gilpin AR. Table for conversion of Kendall's tau to Spearman's rho within the context of measures of magnitude of effect for meta-analysis. *Educ Psychol Meas*. 1993;53:87-92.
38. Adair JG. The Hawthorne effect: a reconsideration of the methodological artifact. *J Appl Psychol*. 1984;69:334-345.
39. Zayyan M. Objective structured clinical examination: the assessment of choice. *Oman Med J*. 2011;26:219-222.
40. Gong Y, Chung YA, Glass J. Ast: audio spectrogram transformer. In: *Proceedings of Interspeech 2021*. August 30–September 3, 2021;2021:571-575.
41. Baevski A, Zhou Y, Mohamed A, Auli M. wav2vec 2.0: a framework for self-supervised learning of speech representations. *Adv Neural Inf Process Syst*. 2020;33:12449-12460.
42. Hsu WN, Bolte B, Tsai YH, Lakhotia K, Salakhutdinov R, Mohamed A. Hubert: self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans Audio Speech Lang Process*. 2021;29:3451-3460.
43. Gonzalez CM, Ark TK, Fisher MR, et al. Racial implicit bias and communication among physicians in a simulated environment. *JAMA Netw Open*. 2024;7:e242181.
44. Sukhera J, Watling C. A framework for integrating implicit bias recognition into health professions education. *Acad Med*. 2018;93:35-40.
45. Sukhera J, Watling CJ, Gonzalez CM. Implicit bias in health professions: from recognition to transformation. *Acad Med*. 2020;95:717-723.
46. Bin Morshed M, Haresamudram HK, Bandaru D, Abowd GD, Ploetz T. A personalized approach for developing a snacking detection system using earbuds in a semi-naturalistic setting. In: *Proceedings of the 2022 ACM International Symposium on Wearable Computers*. ACM; September 11, 2022:11-16.
47. LeBaron V, Flickinger T, Ling D, et al. Feasibility and acceptability testing of CommSense: a novel communication technology to enhance health equity in clinician–patient interactions. *Digit Health*. 2023;9:20552076231184991.

JAMIA Open, 2024, 7, 1–11
https://doi.org/10.1093/jamiaopen/ooae106
Research and Applications