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

Aishwarya Raj[1], Kenneth Yi[1], Rohith Palli MD, PhD[1,3], Pavankumar Suresh[1], Catherine Chia MS[1], Marvin Tran[1], 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 pipeline[19] 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.

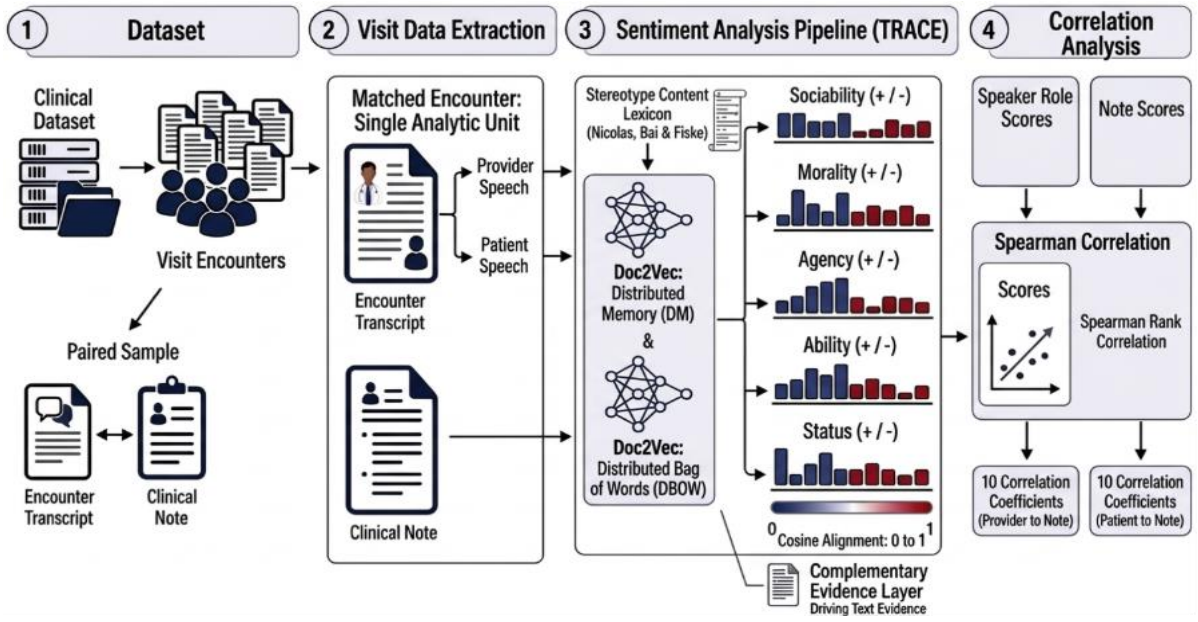


Fig 1. Schematic of study design. Clinical visit transcripts and associated notes were sampled from the Observer Repository. Notes and transcripts were extracted, provider and patient speech were separated, and then TRACE was applied to each of the three sources of clinical text to quantify scores across 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 Repository[20] 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 Fiske[18] 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 (rs) between note and transcript sentiment scores by speaker role (N = 89 visits). *p.adj (adjusted p value) < 0.05 after Benjamini–Hochberg correction.

	Provider speech vs. note		Patient speech vs. note	
Dimension	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.

		Transcripts (Provider-Speech)			Visit notes		
Dimension	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 requirements[29], 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 condensed[5] and task-oriented representation[31] 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 s 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'Conor 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, Schutze 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