Journal of Patient Experience
Volume 12: 1-6

DOI: 10.1177/23743735251314843
journals.sagepub.com/home/jpx
Sage

Natural Language Processing (NLP): Identifying Linguistic Gender Bias in Electronic Medical Records (EMRs)

Site Xu, MPH[1] and Mu Sun, MD[1]

Abstract

With the rise of feminism, women report experiencing doubt or discrimination in medical settings. This study aims to explore the linguistic mechanisms by which physicians express disbelief toward patients and to investigate gender differences in the use of negative medical descriptions. A content analysis of 285 electronic medical records was conducted to identify 4 linguistic bias features: judging, reporting, quoting, and fudging. Sentiment classification and knowledge graph with ICD-11 were used to determine the prevalence of these features in the medical records, and logistic regression was applied to test gender differences. A total of 2354 descriptions were analyzed, with 64.7% of the patients identified as male. Descriptions of female patients contained fewer judgmental linguistic features but more fudging-related linguistic features compared to male patients (judging: OR 0.69, 95% CI 0.54-0.88, $p < 0.01$; fudging: OR 1.38, 95% CI 1.09-1.75, $p < 0.01$). No significant differences were found in the use of reporting (OR 0.95, 95% CI 0.61-1.47, $p = 0.81$) and quoting (OR 0.99, 95% CI 0.72-1.36, $p = 0.96$) between male and female patients. This study highlights how physicians may express disbelief toward patients through linguistic biases, particularly through the use of judging and fudging language. Both male and female patients may face different types of systematic bias from physicians, with female patients experiencing more fudging-related language and less judgmental language compared to male patients. These differences point to a potential mechanism through which gender disparities in healthcare quality may arise, underscoring the need for further investigation and action to address these biases.

Keywords

linguistic gender bias, EMR, sentiment classification, knowledge graph, ICD-11

Introduction

Previous studies have shown that female patients may face clinicians who are more likely to dismiss, ignore, or downplay their concerns compared to male patients. In the United States, such experiences are considered forms of epistemic injustice, where women's inequitable healthcare experiences undermine their authority to manage their own health and bodies.[1] Women are often let down by medical professionals, and their symptoms are frequently misunderstood as psychosomatic before proper treatment is provided.[2,3] In Canada, women are often victims of gaslighting within the healthcare system, suffering when they are repeatedly told to question their reality, particularly when considering alternative medicine.[4]

Research indicates that gender bias, like racial and ethnic disparities, is prevalent in healthcare as in other fields.[5] However, detecting the effects of this bias on clinical care can be challenging because it is typically unconscious and subtle. In healthcare, the primary source of detectable biases is electronic medical record (EMR). According to research in social psychology, language can reflect underlying attitudes.[6–8] Consequently, unconscious biases may manifest in the language used to describe patients in clinical records. Building on this, previous studies have identified several linguistic features that convey a positive or negative attitude toward patients.[9]

Few studies have critically examined EMR to identify the presence of biased language. Research has shown that even highly trained mental health professionals exhibit systematically different judgments when exposed to 2 commonly used terms. Specifically, the term "substance abuser," as opposed

[1] Department of Statistics, Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, China

Corresponding Author:
Mu Sun, Department of Statistics, Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai 200025, China.
Email: schuster_ter@163.com

to "substance use disorder," may reinforce stigmatizing attitudes.[10,11] Moreover, a qualitative analysis of medical records from patients suffering acute pain related to sickle cell disease revealed that notes containing stigmatizing language were associated with more negative attitudes towards these patients. Additionally, such notes were linked to a less aggressive approach to managing the patient's pain.[12] These findings suggest that the use of biased language in medical records can detrimentally affect the quality of future healthcare.

Under the background of the rise of feminism in developing countries like China caused by the key investment of American foundations such as the National Endowment for Democracy,[13] it is worth paying attention to whether there is similar gender bias in the healthcare field of a developing country.

Therefore, to further explore gender bias in EMR, we sampled descriptions suggesting negative attitudes toward patients and then found gender differences in the use of biased language within medical records. This study aims to examine the linguistic mechanisms through which physicians express disbelief toward patients in the EMR. Specifically, it investigates gender differences in the use of 4 linguistic features—judging, reporting, quoting, and fudging—in medical descriptions. The study also seeks to understand how these biased linguistic features may contribute to gender disparities in healthcare quality.

Methods

Data Collection

We studied 2354 medical descriptions from EMRs written in Chinese by physicians in 2021 about patients in the internal medicine departments of Ruijin Hospital. Ruijin Hospital is located in Shanghai, China, and is a tertiary-grade A general hospital that provides data from multiple departments. Medical descriptions refer to the information recorded in the EMR by physicians, which includes patients' symptoms, diagnoses, treatment plans, and other relevant clinical information. These descriptions are used for clinical decision-making and communication among healthcare providers. Although the EMR does contain some template text, it also contains free text that was written by physicians during the encounter. Descriptions included notes written by both attending and resident physicians. The independent variables mainly included gender, age, and place. Wherein, the place referred to as the birthplace of the patient, as a confounding reflecting the potential antiforeign sentiment of physicians.

Linguistic Features

To examine specific linguistic elements, we applied the principles of epistemic modality and evidentiality in a content analysis of 600 internal medicine physician descriptions selected at random. Epistemic modality and evidentiality involve how speakers use linguistic resources to express their level of commitment to the truth of their statements.[14] Record-keepers also express their endorsement or skepticism regarding the source of their information through these linguistic tools.[15,16] Our analysis identified 4 linguistic features indicative of stigmatizing language,[17] as established by prior research: Judging, Reporting, Quoting, and Fudging. We engaged 6 physicians, including 2 senior physicians and 4 attending physicians, with equal gender representation, to evaluate these features using the Delphi method. In the first round, they assessed the relevance and application of each linguistic feature based on their clinical experience. The results were then summarized and provided as feedback for the second round, allowing for refinement of the definitions. After 2 rounds of evaluation, consensus was reached on the final definitions of the linguistic features, ensuring the reliability and validity of our findings.[18]

Judging. Judging refers to a statement that not only identifies the source of information but also creates a distance between the physician and the information, casting doubt on its credibility. This linguistic feature involves negative evaluations of a patient's narrative or provided medical information. Examples include terms such as "claims," "insists," "states," and "denies," as well as qualifiers like "vague," "unknown," "irregular," and "external diagnosis." Additional indicators include phrases like "lack of attention" or the use of a "question mark" to imply uncertainty or skepticism. For example, a physician might write: "The patient claims to have taken the medication, but there is no evidence to support this," creating distance and expressing doubt about the information.

Reporting. Reporting is a grammatical feature that signifies the source of one's knowledge.[14,19–21] A simple declarative statement such as "the patient's stomachache started yesterday" conveys certainty. However, when a physician uses reporting, for example, "the patient reports that the stomachache started yesterday," it attributes the information to another source and slightly distances the physician from committing to its accuracy. Consequently, this study hypothesizes that the more frequently a physician employs reporting, the greater their skepticism towards the patient's account. The category of "reporting" includes various tenses of verbs such as "reports," "informs," "mentions," and "expresses."

Quoting. Quoting is a complicated grammatical feature intended to promote accuracy by citing directly.[22] In fact, medical training encourages patients to be quoted to make EMR more patient-oriented.[23,24] However, due to the evolution of societal usage of quotes, quoting has become an indication of dubious. When physicians record, "she reports she had a 'stomachache' due to the medication," they probably

do not trust the 'stomachache' occurred. In popular culture, these are referred to as scare quotes.

Fudging. Fudging refers to the phenomenon of providing vague, incomplete, or unsubstantiated entries in medical records. Physicians may use perfunctory phrases such as "treatment plan as before" or simply write "none," or even directly copy from previous medical records without updating or verifying the information. In some cases, parts of the medical record may be left blank or unfilled, which is also categorized as fudging. For example, a physician might write: "No significant changes," without specifying what was observed or why further details were omitted. This lack of detailed documentation can undermine the accuracy and transparency of medical records.

Statistical Analysis

Based on NLP (natural language processing) methods, the study attempted to identify 4 linguistic features—judging, reporting, quoting, and fudging—and to output number of appearances of each feature in each description.

The study used the Ernie 3.0 algorithm, a pre-trained language model developed by Baidu, to build a sentiment classification model.[25,26] Ernie 3.0 leverages a multi-layer transformer architecture and incorporates knowledge enhancement techniques to improve its performance on downstream tasks. For this study, we fine-tuned Ernie 3.0 on 285 labeled samples extracted from 2354 medical records. The labeling process involved 5 physicians above the level of attending physician, who annotated the data based on predefined linguistic bias features. The dataset was split into training, validation, and test sets in an 8:1:1 ratio. During model training, hyperparameters such as learning rate, batch size, and the number of epochs were optimized using grid search. The evaluation metrics included accuracy, precision, recall, and F1-score to assess the model's performance. The best-performing model achieved an F1-score of 0.92 on the test set, indicating high reliability. The trained model was applied to all 2354 medical descriptions to classify linguistic bias features. Each output was further verified by 5 physicians using a majority voting principle to determine the final classification. Additionally, named entity recognition (NER) was applied to extract biased entities from the medical records. Relationship extraction techniques were used to link these entities, and the sentiment knowledge graph was constructed to confirm sentiment orientations. This process demonstrated the effectiveness of combining Ernie 3.0, NER, relationship extraction, and knowledge graph construction for mining Chinese EMRs. The ICD-11, especially its independent numbering of URIs and its entity-like nature, played an important role throughout the entire modeling process. To address potential sampling and

Table 1. Medical Description Characteristics.

		N(%) Descriptions
Gender	Man	1524(64.74%)
	Woman	830(35.26%)
Place	Local	1885(80.08%)
	Non-local	469(19.92%)
Age	[20,45]	180(7.65%)
	[45,70]	1133(48.13%)
	[70,95]	1041(44.22%)

Table 2. Linguistic Feature Characteristics.

	Gender (mean ± std)		Place (mean ± std)		Age (mean ± std)		
Linguistic features	Male	Female	Local	Non-local	[20,45]	[45,70]	[70,95]
Judging	2.63 ± 2.48	2.02 ± 2.21	2.33 ± 2.39	2.75 ± 2.44	2.49 ± 3.08	2.31 ± 2.04	2.52 ± 2.62
Reporting	0.07 ± 0.28	0.08 ± 0.28	0.06 ± 0.27	0.1 ± 0.3	0.08 ± 0.27	0.07 ± 0.29	0.07 ± 0.27
Quoting	0.89 ± 1.39	1.01 ± 1.5	0.97 ± 1.52	0.79 ± 0.98	0.77 ± 1.11	0.73 ± 1.35	1.18 ± 1.52
Fudging	0.75 ± 0.96	1.02 ± 1.13	0.78 ± 0.94	1.12 ± 1.29	0.89 ± 1.39	0.92 ± 1.08	0.76 ± 0.87
Sum	4.34 ± 3.75	4.13 ± 3.52	4.14 ± 3.71	4.76 ± 3.46	4.23 ± 4.16	4.03 ± 3.12	4.53 ± 4.09

Table 3. Prevalence of Linguistic Features by Gender.

	Gender (mean ± std)		Unadjusted female-male difference	Adjusted female-male difference
Linguistic features	Male	Female	OR (95% CI)	OR (95% CI)
Judging	2.63 ± 2.48	2.02 ± 2.21	0.15(0.13, 0.18)*	0.69(0.54, 0.88)*
Reporting	0.07 ± 0.28	0.08 ± 0.28	0.04(0.03, 0.06)*	0.95(0.61, 1.47)
Quoting	0.89 ± 1.39	1.01 ± 1.5	0.08(0.07, 0.11)*	0.99(0.72, 1.36)
Fudging	0.75 ± 0.96	1.02 ± 1.13	0.21(0.17, 0.25)*	1.38(1.09, 1.75)*
Sum	4.34 ± 3.75	4.13 ± 3.52	0.22(0.18, 0.26)*	0.89(0.71, 1.1)

* $P < .01$.

labeling biases, we ensured that the 5 physicians who labeled the data were from diverse clinical backgrounds and underwent standardized training on the linguistic bias definitions. The labeling process was further reviewed to minimize inconsistencies. These measures aimed to reduce subjectivity and improve the robustness of the final model outputs.[27–31]

Results

Study Sample

Table 1 displays the number of medical descriptions by gender, place, and age. Of 2354 descriptions, 1524 were male patients and 830 were female patients.

Characteristics of Linguistic Features

As the medical records we used did not include any gender diverse categories, the data under the categories of gender and sex are effectively equivalent in this study. Table 2 displays the number of different linguistic features by gender, place, and age. In terms of the total number of characteristics, male patients were more than female patients, non-local patients were more than local patients, and elderly patients (aged 70 to 95) were more than other patients. "Sum" refers to the total count of all 4 linguistic features.

Linguistic Differences by Gender

The linguistic features used in descriptions by gender are shown in Table 3. In unadjusted analyses, all features significantly appeared more in male patients than in female patients. In adjusted analyses that incorporated confounding factors of place and age, we found that judging was less likely to be used in descriptions for female compared to male patients (OR 0.69, 95% CI 0.54-0.88), and fudging was more likely to be used in descriptions for female compared to male patients (OR 1.38, 95% CI 1.09-1.75). "Sum" refers to the total count of all 4 linguistic features.

Discussion

The study found labels of disbelief in EMR of both male and female patients, indicating that patients of different genders may be confronted with different kinds of linguistic biases from physicians. People whose self-knowledge is questioned may suffer harm as a result of this bias, which could also lead to the lack of trust from patients to their clinicians and have a negative impact on healthcare quality.

Linguistic bias, as a kind of testimonial injustice, occurs when the listener's prejudice results in an unfair deficit of credibility for the speaker.[32] Negative occurrence, such as delayed diagnosis, inappropriate treatment, and even death, happens when patients suffer from this kind of bias in the healthcare field.

Credibility doubts can be caused by 3 different reasons: queries about the competency of interpreting accurately, concerns about deliberate deception without sincerity, or be perfunctory simply due to contempt. All of these may be in use in the healthcare field, which could explain the gender differences we observed. The first reason above is often accompanied by a stronger emotion, which is mainly manifested through the category of judging, and mostly occurred in male patients. The second reason above exists in categories of reporting and quoting. The third reason above often leads to the category of fudging, which usually occurred in female patients. The results hinted that there are different tendencies of discrimination against patients of different genders. Male patients are more likely to meet stronger opposition and questioning, while female patients are more likely to meet more contempt and neglect.

The study may have some potential methodological limitations. It is significant to note that the linguistic features in the study might not be accurate indicators of linguistic bias. Particularly, the use of reporting or quoting to describe the experiences of patients is not necessarily negative and may be beneficial for objectivity. Take quoting for example, the use of it may have different motivations in different situations. It's not always wrong to quote a patient in their descriptions, and it could even be done with good intentions. It may occur more frequently when there is a greater cultural distance between physicians and patients. In addition, on the statistical level, firstly, due to the insufficient quality of EMR data, the information of medical recorders was not well preserved, so this study did not consider the mixed effect caused by the same clinician's recording of multiple descriptions. Secondly, we did not have data on the demographic or other characteristics of the physicians writing the descriptions we analyzed. As such, we were unable to determine whether the gender, place of birth, age, or training status of physicians affected how they used the linguistic features we focused on. Thirdly, socioeconomic status was not taken into account in our analysis, because EMR does not readily contain information about income or education. Finally, among the confounding factors, the regional discrimination shown by birthplace is not accurate enough, and the combination of ancestral place and birthplace may be better.

However, despite the possibility of mislabeling, the fact that we discovered gender differences suggests that our findings represent a cautious estimate of linguistic bias in EMR. Different from previous studies and the general impression, the victims of gender discrimination are not only women, which provides a more reasonable direction for controlling medical quality in the future.

It is important to consider that the use of these linguistic features is more likely to reflect unconscious gender bias. This means that physicians are more likely to be doing so without realizing it. There is a lot of speculation about how physicians' implicit bias contributes to healthcare disparities between males

and females, but there is little evidence about how implicit bias affects healthcare decision-making and delivery. One possible direction that could be the focus of interventions to reduce the implicit bias is clarified by the findings of our study.

It is also worth noting that, unlike the unilateral persecution claimed by ultra-feminists in developing countries like China funded by typical organizations, gender discrimination is a problem that exists widely both in males and females. Whether feminism is alienated by organizations with ulterior motives as a weapon to intensify contradictions in developing countries deserves our vigilance. Lastly, it is important to emphasize that, during the course of the study, we identified the dynamic characteristics of ICD-11-like entities, which provided significant support for the mining of EMR data.[33]

In the whole process of diagnosis and treatment, physicians must record medical descriptions in a more respectful way to eliminate the influence of gender or other hidden prejudices that are unwarranted and to avoid letting patients at risk for lower-quality care. Further research is needed to mine linguistic bias in EMR, and interventions or standards should be developed to lessen its impact.[33]

Acknowledgments

We gratefully acknowledge each medical worker for the help in data collection.

Author Contributions

All authors contributed to designing the study. Xu and Sun were responsible for the investigation and consultation. Xu S was responsible for data collection and analysis. Xu S was responsible for writing the manuscript. The corresponding author Sun attested that all listed authors meet authorship criteria. No other individuals meeting the criteria have been omitted. Sun is the guarantor. All authors have read and approved the final manuscript.

Availability of Data and Materials

The datasets analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request with permission.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The work was supported by Young Talent Development Program in the Humanities of Shanghai Jiao Tong University (2025QN036) and Youth Talent Development Program of Ruijin Hospital (2024PY065). The funding agencies had no role in the design and conduct of the study; or in the collection, management, data analysis, result interpretation, preparation and review of the manuscript.

Ethics Approval and Consent to Participate

As a retrospective study, the data were derived from medical records, did not involve intervention and follow-up, and were without patients' identifiable information. The study has no ethical issues involved.

Informed Consent

Informed consent was obtained from all individual participants included in the study.

ORCID iD

Site Xu https://orcid.org/0000-0003-4317-6041

References

1. Thompson CM, Babu S, Makos S. Women's experiences of health-related communicative disenfranchisement. *Health Commun.* 2022;38(14):3135–46.
2. Chen EH, Shofer FS, Dean AJ, et al. Gender disparity in analgesic treatment of emergency department patients with acute abdominal pain. *Acad Emerg Med.* 2008;15(5):414–18.
3. Macphee M. Hysterical women: pain bias and epistemic injustice in medicine. *Granite J Postgrad Interdiscip*. 2019;3(2):12–9.
4. Ashley F. The doctor doesn't listen to her. But the media is starting to. *The Atlantic*. 2018. https://www.theatlantic.com/family/archive/2018/08/womens-health-care-gaslighting/567149/
5. Santry HP, Wren SM. The role of unconscious bias in surgical safety and outcomes. *Surg Clin*. 2012;92(1):137–51.
6. Newman ML, Groom CJ, Handelman LD, Pennebaker JW. Gender differences in language use: an analysis of 14,000 text samples. *Discourse Process.* 2008;45(3):211–36.
7. Lindquist KA, MacCormack JK, Shablack H. The role of language in emotion: predictions from psychological constructionism. *Front Psychol.* 2015;6:444. doi:10.3389/fpsyg.2015.00444
8. Beukeboom CJ, Burgers C. Linguistic bias. In: *Oxford Research encyclopedia of communication*. Oxford University Press; 2017, pp.1–19.
9. Park J, Saha S, Chee B, Taylor J, Beach MC. Physician use of stigmatizing language in patient medical records. *JAMA Network Open*. 2021;4(7):e2117052–e2117052.
10. Kelly JF, Westerhoff CM. Does it matter how we refer to individuals with substance-related conditions? A randomized study of two commonly used terms. *Int J Drug Policy*. 2010;21(3):202–7.
11. Kelly JF, Dow SJ, Westerhoff C. Does our choice of substance-related terms influence perceptions of treatment need? An empirical investigation with two commonly used terms. *J Drug Issues.* 2010;40(4):805–18.
12. Goddu AP, O'Conor KJ, Lanzkron S, 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
13. National Endowment for Democracy. Mainland CHINA 2020. National Endowment for Democracy. 2021, < Mned.org/region/asia/mainland-china-2020/>.
14. Wu Y, Ma W. The handbook of conversation analysis. *Aust J Ling*. 2015;35:418–20.

15. Schegloff EA. 1987. Some sources of misunderstanding in talk-in-interaction.
16. Pomerantz A. 1988. Constructing skepticism: Four devices used to engender the audience's skepticism.
17. Park J, Saha S, Chee B, Taylor J, Beach MC. Physician use of stigmatizing language in patient medical records. *JAMA Network Open*. 2021;4(7):e2117052–e2117052.
18. Linstone HA, Turoff M (Eds) *The delphi method*. MA: Addison-Wesley, 1975, pp. 3–12.
19. Nuyts J. Subjectivity as an evidential dimension in epistemic modal expressions. *J Pragmat.* 2001;33(3):383–400.
20. Stivers T, Mondada L, Steensig J. Knowledge, morality and affiliation in social interaction. *Moral Knowl Convers*. 2011;2011:3–24. doi:10.1017/CBO9780511921674.002
21. Vickers CH. Occasioned membership categorization in a transnational medical consultation: interaction, marginalization, and health disparities. *J Socioling*. 2020;24(5):574–92.
22. Finnegan RH. *Why do we quote? The culture and history of quotation*. Open Book Publishers; 2011.
23. Srithar K, Harrow S, Nadarajasundaram A, Mensah A. Comments on 'seven reasons why the physical examination remains important'. *J R Coll Phys Edinburgh*. 2021;51(4):424–7.
24. Bickley LS, Hoekelman RA. Bates' guide to physical examination and history taking. In: *Bates' guide to physical examination and history taking*. Lippincott Williams & Wilkins; 1995, p.789.
25. Sun Y, Wang S, Li Y, et al. Ernie: Enhanced representation through knowledge integration. arXiv preprint arXiv:1904.09223, 2019.
26. Hsieh YH, Zeng XP. Sentiment analysis: an ERNIE-BiLSTM approach to bullet screen comments. *Sensors*. 2022;22(14):5223.
27. Xu S, Sun M. Assessment of EMR ML mining methods for measuring association between metal mixture and mortality for hypertension. *High Blood Press Cardiovasc Prev*. 2024;31(5):473–83 doi:10.1007/s40292-024-00666-w
28. Xu S, Sun M. The interpretable machine learning model associated with metal mixtures to identify hypertension via EMR mining method. *J Clin Hypertens*. 2024;26(2):187–96. doi:10.1111/jch.v26.2
29. Xu S, Zhang T, Sheng T, Liu J, Sun M, Luo L. Cost supervision mining from EMR based on artificial intelligence technology. *Technol Health Care*. 2023;31(3):1077–91. doi:10.3233/THC-220608
30. Xu S, Sun M. Covid-19 vaccine effectiveness during Omicron BA.2 pandemic in Shanghai: a cross-sectional study based on EMR. *Medicine*. 2022;101(45):e31763. doi:10.1097/MD.0000000000031763
31. Xu S-T, Sun M, Xiang Y. Global, regional, and national trends in type 2 diabetes mellitus burden among adolescents and young adults aged 10 -24 years from 1990 to 2021: a trend analysis from the Global Burden of Disease Study 2021. *World J Pediatr*. 2025. doi:10.1007/s12519-024-00861-8
32. Medina J. *The epistemology of resistance: gender and racial oppression, epistemic injustice, and resistant imaginations*. Oxford University Press; 2012.
33. McDowell MJ, Goldhammer H, Potter JE, Keuroghlian AS. Strategies to mitigate clinician implicit bias against sexual and gender minority patients. *Psychosomatics*. 2020;61(6):655–61. doi:10.1016/j.psym.2020.04.021