The Human Side of Service Engineering, Vol. 221, 2026, 333–343
https://doi.org/10.54941/ahfe1007737

Exploring Necessary Conditions for High and Low Patient Ratings in Online Healthcare Consultation: An LLM-Based Weak Supervision Approach

Hanshu Wang and Xiuzhu Gu

Department of Industrial Engineering and Economics, School of Engineering, Institute of Science Tokyo, Tokyo, Japan

ABSTRACT

The prevalence of the "positive rating plateau" in Online Healthcare Consultation (OHC) hinders a clear understanding of the critical constraints required for service improvement. This study aims to identify the necessary conditions for high and low patient ratings. We utilized a dataset of 517,026 patient narrative reviews and adopted an LLM-based weak supervision approach to quantify 12 emotion categories. Subsequently, Necessary Condition Analysis (NCA) was conducted on patient reviews for 625 physicians, to identify necessary emotion categories and their bottleneck levels for high (5-star) and low (1-3 star) ratings. The results revealed that positive emotions are not necessary for high ratings. In contrast, Disappointment and Distrust (d = 0.390, $p < 0.001$) emerged as the core necessary condition for low ratings. Additionally, Price Complaints (d = 0 .196, $p < 0.001$) and Poor Communication (d = 0.180, $p < 0.001$) exhibited distinct bottleneck characteristics across varying low-rating rates. Beyond informing a "stepwise improvement strategy" for physicians to place greater emphasis on caring practices and optimizing bilateral service design for platforms, this study establishes a low-cost and high-precision analytical paradigm fusing LLM-based weak supervision with NCA to address the semantic complexity and data sparsity inherent in unstructured medical text.

Keywords: Online healthcare consultation (OHC), Patient ratings, Large language model (LLM), Necessary condition analysis (NCA), Weak supervision

INTRODUCTION

With the rapid development of digital technologies in the healthcare sector, Online Healthcare Consultation (OHC) has become a critical channel for alleviating the uneven distribution of medical resources and improving service accessibility (Wang et al., 2025). For healthcare providers, OHC serves not only as a tool to extend service boundaries but also as an important platform for obtaining economic returns (Li and Hu, 2025). Against this background, patients' Online Reviews have become one of the key indicators for users to evaluate healthcare service quality (Zhang et al., 2025). These massive amounts of unstructured textual data not only reflect patients' experiences

Received February 2, 2026; Revised April 16, 2026; Accepted May 3, 2026; Available online July 20, 2026

and sentiments but also influence the medical decision-making of subsequent patients and the service performance of healthcare providers (Yang et al., 2025). Therefore, deeply mining semantic information within patient reviews holds significant theoretical and practical value for understanding the mechanism of online physician-patient interaction and optimizing platform services.

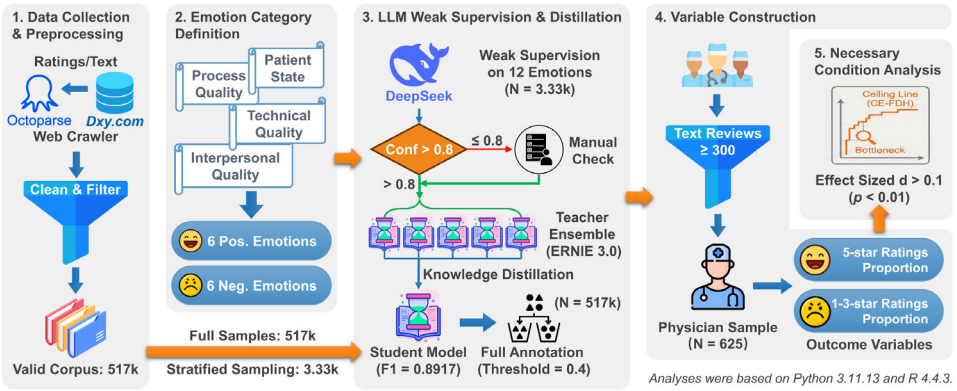

Figure 1: Workflow for LLM-based weak supervision and NCA.

The current online review system faces the severe challenge of a "positive rating plateau". Existing research indicates that patient reviews in OHC exhibit a significant "inverse L-shaped" distribution, meaning that the vast majority of ratings tend to be high, while low ratings are scarce (Fan et al., 2022). This prevalence of high ratings creates a bottleneck for traditional linear analysis methods (such as regression analysis) in explaining "how to further improve services", as these methods focus on demonstrating the net effects of variables and struggle to identify the bottleneck factors required to achieve high performance (Dul et al., 2023). Specifically, it is unknown which emotional elements are necessary for obtaining high ratings, and which serve as the baseline for avoiding low ratings.

Furthermore, traditional hand-coding or dictionary-based text analysis methods often face limitations in efficiency or accuracy when processing massive volumes of reviews, particularly when dealing with complex contexts (Madanay et al., 2024). Although Large Language Models (LLMs) have demonstrated immense potential in medical text processing (Luo et al., 2024), how to integrate them with necessity analysis to systematically explore the necessity constraints behind patient ratings remains a gap in the current literature.

To address these limitations, this study aims to integrate LLM-based weak supervision with Necessary Condition Analysis (NCA) to deeply investigate the necessary factors contributing to high and low patient ratings in OHC. By leveraging this novel analytical structure to reveal the necessity mechanisms underlying the formation of high and low ratings, this study provides specific improvement paths for healthcare providers to break through the "positive rating plateau" and offers a reference for OHC platforms to design more adaptive service feedback mechanisms.

METHODOLOGY

The research process is illustrated in **Figure 1**. This process integrates large-scale unstructured text processing with necessity inference, comprising five main phases: (1) Data collection and preprocessing; (2) Literature-based definition of fine-grained emotion categories; (3) LLM-based weak supervision and knowledge distillation; (4) Physician-level variable construction; and (5) Necessity and bottleneck analysis.

Data Collection and Preprocessing

The data for this study were sourced from a mainstream Chinese OHC platform, Dingxiang (Dxy.com). This platform hosts a vast number of practicing physicians, allowing patients to leave ratings and textual reviews after consultations. Using the web crawler Octoparse (Octoparse.com), we collected publicly available anonymous patient ratings and textual reviews between August and September 2025 for physicians with over 200 rating records, in accordance with the platform's terms of service, and without collecting any identifiable or sensitive personal information. Our preprocessing procedures included: (1) removing reviews with low semantic load (fewer than 5 Chinese characters or 10 characters); (2) cleaning non-textual noise such as special symbols, continuous punctuation, and garbled text; and (3) eliminating duplicate content and reviews containing significant non-Chinese text. Ultimately, we obtained a corpus of 517,026 valid patient reviews.

Definition of Emotion Categories

To extract emotion from unstructured reviews, this study constructed a fine-grained emotion classification based on prior research regarding OHC service quality, online physician-patient relationships and interaction.

First, drawing on the SERVQUAL model and related theories, we mapped emotion categories to four core dimensions: Technical Quality, Interpersonal Quality, Process Quality, and Patient State, to cover the full spectrum of patient emotion after the OHC (Shan et al., 2019; Wu and Lu, 2018; Yang et al., 2025). On this basis, we defined 12 specific emotion categories. These include six positive categories (Reassured and Relieved, Professional and Trustworthy, Caring and Supportive Attitude, Patient Gratitude, Convenience and Efficient Process, Clear and Effective Communication) and six negative categories (Disappointment and Distrust, Perceived Medical Error and Safety Risk, Price Complaints, Waiting Time and Delay Frustration, Patient Anxiety and Worry, Poor Communication). These categories comprehensively cover the diverse emotions embedded in patient textual reviews.

LLM-Based Weak Supervision and Knowledge Distillation

Given the massive volume of the corpus, traditional manual annotation was cost prohibitive. Furthermore, negative reviews were significantly scarcer than positive ones (Class Imbalance). Therefore, this study designed a training

strategy integrating stratified sampling, LLM-based weak supervision, and knowledge distillation to enhance the model's ability to identify sparse negative emotions.

First, to ensure training data covered all emotion categories, especially rare negative reviews, we performed stratified sampling based on star ratings on the raw corpus, extracting a total of 33,300 representative review samples. We selected DeepSeek-chat for weak supervision due to its superior performance in understanding Chinese contexts. We designed a prompt containing role definition, task instructions, and output constraints to instruct the model to perform multi-label classification and output a Confidence Score. The prompt example is: "You are an emotion analysis expert. Please read the review and annotate based on the predefined 12 categories. Strictly output JSON format containing labels and confidence score...". To ensure the quality of pseudo-labels, we implemented a "Human-in-the-loop" calibration strategy. We inspected samples based on the confidence scores output by the LLM's results. It showed that when confidence exceeded 0.8, the annotation highly aligned with human judgment, requiring almost no modification. Consequently, we manually corrected all pseudo-labels with confidence below 0.8 and combined them with samples above 0.8 to form a "Silver Standard" dataset.

Addressing the instability of model training caused by the scarcity of negative emotion samples, we adopted a 5-Fold cross-validation strategy to construct a teacher model group. Specifically, we divided the 33,300 annotated samples into five folds and trained five independent ERNIE 3.0 (ernie-3.0-xbase-zh) classifiers as the "Teacher Ensemble". We then transferred the ensemble knowledge from these five teacher models to a single ERNIE 3.0 Student Model via distillation techniques. Ultimately, this student model achieved a weighted F1 Score of 0.8917 on the test set, demonstrating that the method could balance recognition rates for both positive and negative categories in highly imbalanced emotional data. Specifically, the student model produces probability scores for each of the 12 emotion categories at the review level. To convert probabilistic outputs into binary emotion indicators, we applied a fixed decision threshold (prob_threshold = 0.4), whereby emotions with predicted probabilities above this threshold were retained for subsequent analysis. Subsequently, we employed this student model for full corpus annotation.

All analyses described above were conducted using Python 3.11.13.

Variable Construction

To explore the necessary conditions for breaking through the "positive rating plateau", we defined and calculated two outcome variables for each physician based on patient star ratings: (1) the high-rating ratio, defined as the proportion of 5-star ratings; and (2) the low-rating ratio, defined as the proportion of 1–3-star ratings. For condition variables, we calculated the proportion score for each physician across the 12 emotion categories by dividing count of each category's reviews by the physician's total review count. To minimize statistical bias from small review sizes, we only included physicians with more than 300 textual reviews. Ultimately, 625 physicians were included in the NCA analysis sample.

NECESSARY CONDITION ANALYSIS (NCA)

This study employs Necessary Condition Analysis (NCA) to identify the necessary conditions producing high ratings or low ratings. Unlike traditional regression analyses that focus on the net effects of variables, NCA aims to identify necessary rather than sufficient relationships, where a certain condition (e.g., a good attitude) must reach a minimum level for the outcome (e.g., a high positive-rating ratio) to occur.

Specifically, we performed the analysis using the NCA package in R. The Ceiling Line was estimated using the Ceiling Envelopment-Free Disposal Hull (CE-FDH) technique. This method generates a stepwise, non-decreasing envelope line along the upper-left boundary of the scatter plot observations, offering flexibility for various data distributions without presuming a functional form. We quantified the strength of necessity by calculating the effect size d. According to established standards, a condition was deemed a necessary condition if $d > 0.1$ and the permutation test showed statistical significance ($p < 0.01$) (Dul et al., 2023). Subsequently, we conducted a bottleneck analysis to quantify the minimum threshold levels required for each emotion category to achieve varying levels of rating outcomes (e.g., a 90% positive rating rate).

RESULTS

Distribution of Emotion Categories Across Review Star Ratings

Figure 2 illustrates the distribution of the 12 fine-grained emotion categories across review star ratings. From an overall perspective, positive emotions overwhelmingly dominate 5-star reviews; 4-star reviews begin to exhibit a mixed state of positive and negative, whereas from 3-star to 1-star reviews, negative emotions rapidly occupy most of the content.

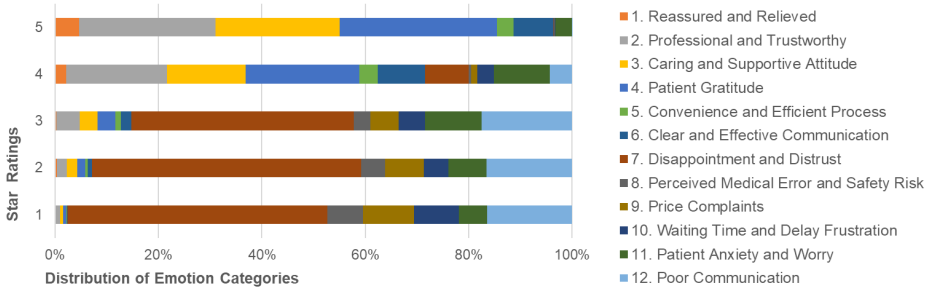

Figure 2: Distribution of fine-grained emotion categories across different star ratings.

Specifically, within the 5-star and 4-star reviews, the three most prevalent emotion categories are Professional and Trustworthy, Caring and Supportive Attitude, and Patient Gratitude. The proportions of these three categories are relatively balanced, constituting the core of patient satisfaction. In contrast, among 3-star to 1-star reviews, Disappointment and Distrust consistently accounts for the largest proportion, followed by Poor Communication. A longitudinal comparison across low-star ratings reveals that the proportions

of Disappointment and Distrust and Poor Communication remain relatively stable between 3-star and 1-star reviews. The primary dynamic variation is that as the rating drops from 3-star to 1-star, the proportions of Price Complaints and Waiting Time and Delay Frustration, which relate to the OHC platform experience, show a distinct upward trend. Notably, Patient Anxiety and Worry is one of the few emotion categories that consistently appears across all star ratings and remains visibly present in each rating group, although its proportion varies across rating levels.

Distribution of Physician Sample Characteristics

Figure 3 presents the characteristics of the 625 physicians included in the NCA analysis. Regarding job title distribution, the sample is dominated by Attending Physicians (n = 269) and Associate Chief Physicians (n = 263). In comparison, junior Physicians (n = 9) and senior Chief Physicians (n = 84) are relatively fewer in number.

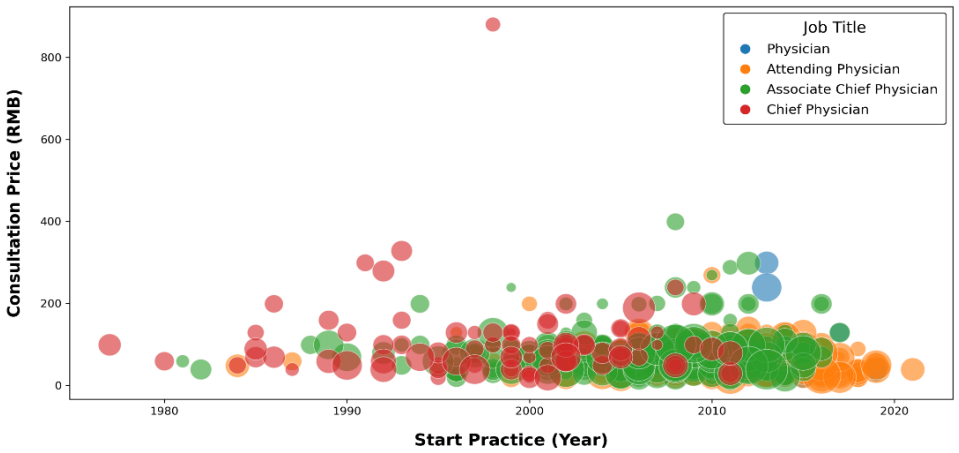

Figure 3: Distribution by price, practice duration, job title, and total consultation volume.

Regarding total consultation volume (ball size), there is a substantial disparity among individuals (from 1,395 to 64,902). Data indicate that the primary contributors to online consultation volume are concentrated among physicians who began practicing between 1995 and 2020. This interval coincides with the densest distribution of Attending and Associate Chief Physicians, suggesting that middle-aged and young physicians exhibit the highest activity levels on the OHC platform.

Regarding consultation price, the sample pricing ranges from 10 to 880 RMB but is primarily concentrated in the 19 to 199 RMB band. Within this mainstream price range, no significant linear relationship was observed between price and total consultation volume. Furthermore, there was no significant correlation between the physicians' start of practice and their pricing. Relatively high-priced samples (>200 RMB) mainly originate from the Associate Chief Physician and Chief Physician groups, reflecting the price premium capability associated with higher professional titles.

Descriptive Statistics of Variables

Figure 4 visually displays the proportional distribution of the twelve emotion categories, the 5-star ratings, and the 1–3-star ratings at the physician level via boxplots, with the extremes of each variable shown in the square brackets.

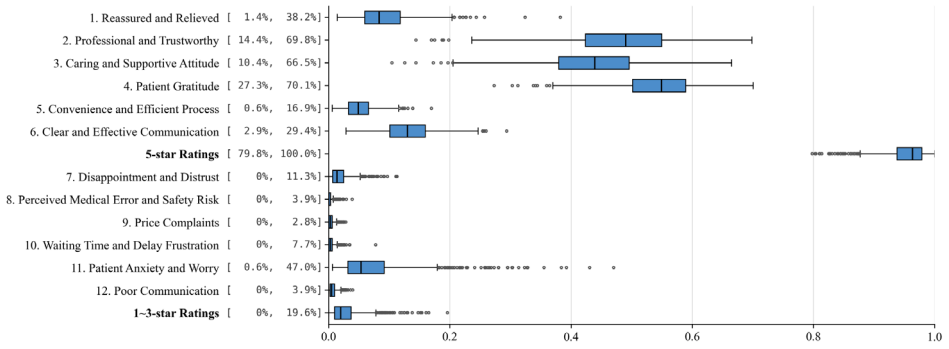

Figure 4: Descriptive statistics of proportions of emotion category and ratings.

First, the results reveal an extreme asymmetry in rating distribution within the OHC. The rate at which physicians receive 5-star ratings exhibits a significant ceiling effect, with the distribution concentrated in the high range of 79.8% to 100%. In contrast, the rate of 1–3-star ratings are compressed within the low range of 0% to 19.6%. Second, among positive emotion categories, Patient Gratitude, Professional and Trustworthy, and Caring and Supportive Attitude are the three most frequently mentioned categories, constituting the core content of positive patient feedback. In contrast, Convenience and Efficient Process has the lowest mention rate among positive emotions (with an extremely low median and a distribution range of only 0.6%–16.9%). This indicates that although process convenience is a core competency of OHC services, patients rarely actively praise it. Finally, among the negative emotion categories, most types (e.g., Price Complaints) exhibit low and narrowly distributed mention ratios, whereas Patient Anxiety and Worry stands out with both a comparatively higher prevalence and substantially greater inter-individual heterogeneity, with the maximum mention rate reaching 47.0%. This significant fluctuation range seems to imply substantial differences among physicians in their ability to address and alleviate patient anxiety.

Necessity and Bottleneck Analysis of Emotion Categories

Figure 5 presents the results of the NCA for high and low ratings (with effect size d and significance p listed on the right).

For positive emotion categories, according to mainstream NCA standards ($d > 0.1$, $p < 0.01$), no emotion category was identified as a necessary condition to be improved in the context of aiming to break through the "positive rating plateau". Regarding negative emotion categories, the results reveal significant necessary conditions. First, Disappointment and Distrust is the core necessary factor for the generation of low ratings, with a large

effect size (d = 0.390, $p < 0.001$). Subsequently, conditions with medium effects are Price Complaints (d = 0.196, $p < 0.001$), Poor Communication (d = 0.180, $p < 0.001$), and Perceived Medical Error and Safety Risk (d = 0.136, $p < 0.001$).

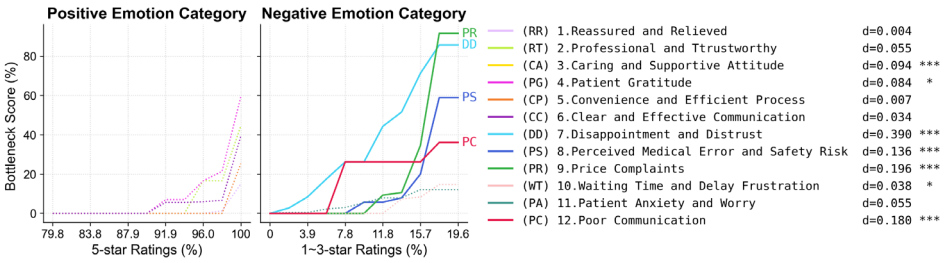

Figure 5: NCA results (Right: Effect Sizes and Significance) and bottleneck scores of emotion categories for high and low rating outcomes (Left) (***p<0.001, *p<0.05).

Further bottleneck analysis reveals the dynamic characteristics of these necessary conditions at different review-rate thresholds. As a direct manifestation of dissatisfaction, the bottleneck level of Disappointment and Distrust exhibits an approximately linear growth trend as the 1–3-star review rate increases. For other categories, the constraining effects demonstrate certain phasic characteristics: (1) When the 1–3-star review rate is in a high range (i.e., many low ratings), the bottleneck level of Price Complaints is very high, followed by Perceived Medical Error and Safety Risk, and finally Poor Communication. This implies that physicians with relatively lower ratings must have received a higher volume of price complaints. (2) As the 1–3-star review rate drops to a medium range, the bottleneck level of Poor Communication becomes relatively higher, while the levels of Perceived Medical Error and Safety Risk and Price Complaints decline rapidly. This suggests that after reducing medical errors and price disputes, physicians who still have relatively high ratio of low ratings must have limited communication skills. The above results indicate that the formation of low ratings exbibits stepwise constraint characteristics.

DISCUSSION

Given the relative prevalence of patient anxiety among negative emotions, physicians must realize that patients in OHCs, particularly those with chronic diseases, often carry severe psychological burdens (Liu et al., 2022). Consequently, the medical service paradigm needs to place greater emphasis on caring practices that include emotional support, alongside clinical treatment. Research demonstrates that emotional support and patience explanation are key drivers for enhancing patient satisfaction (Yang et al., 2025).

Based on the bottleneck analysis results from NCA, we propose a "stepwise improvement strategy" for physicians with a high ratio of low ratings. First, mitigating "Price Complaints" is the primary strategy; pricing that far exceeds the patient's estimated cost can easily trigger a sense of injustice (Wu and Lu, 2018). This strategy does not imply a simple price reduction, as our results

found no significant linear relationship between pricing and consultation volume within mainstream price ranges. Physicians should scrutinize whether their pricing strategy matches the complexity of the condition and the depth of service provided (e.g., response word count, interaction frequency) to alleviate patient dissatisfaction regarding cost-effectiveness. Second, after addressing pricing complain points, improving communication serves as the advanced path. For example, physicians can enhance patients' willingness to consult by providing high-quality content responses, appropriate emotional expression, suitable response times, and appropriately in-depth interaction (Kong et al., 2025).

Additionally, platforms should optimize bilateral service design. From the patient perspective, platforms should guide users to provide structured feedback on dimensions such as communication experience and cost-effectiveness through designing review tags. This helps overcome the sparsity of textual data and capture full-dimensional emotional signals. From the physician perspective, given physicians' limited time and energy, the platform should provide customized improvement pathways for physicians at different development stages based on NCA results. For instance, the platform could push reasonable pricing reference ranges to physicians in the price-complaint bottleneck stage (Wu and Lu, 2018), or provide high quality physician-patient dialogue templates to physicians constrained by insufficient communication experience, thereby strengthening their communication skills (Li and Hu, 2025). Furthermore, the platform should utilize these fine-grained emotion profiles to achieve more precise physician-patient matching, thereby reducing low ratings caused by expectation mismatches.

This study innovatively proposes a research method fusing LLM-based weak supervision with NCA, thereby providing a cost-effective and high-precision paradigm to address challenges such as semantic complexity, class imbalance, and data sparsity in unstructured medical texts (Madanay et al., 2024). Notably, we revisit the formation mechanisms of patient ratings from a "necessity" perspective, precisely identifying the necessary conditions and their bottleneck levels for the formation of low ratings, thereby offering a novel perspective for explaining phenomena such as the " positive rating plateau".

There are two limitations in this study. First, the use of cross-sectional data limits inferences regarding dynamic relationships, failing to capture the dynamic impact of changes in physician pricing strategies or platform policy adjustments on reviews (Yang et al., 2025). Second, data from a single platform may contain sample selection bias. Future research could combine longitudinal data across multiple platforms to further explore the dynamic evolutionary patterns of necessary conditions across different departments or disease types.

CONCLUSION

This study innovatively integrates LLM-based weak supervision with NCA to uncover the necessary conditions underlying the formation of OHC patient ratings. The results indicate that positive emotions are not necessary

conditions for obtaining high ratings. In contrast, the formation of low ratings is premised on specific emotions: Disappointment and Distrust serve as the core necessary condition, whereas Price Complaints and Poor Communication exhibit distinct phased bottleneck characteristics across different low-rating rates. Therefore, physicians with high rates of low ratings should follow a stepwise improvement strategy ranging from reasonable pricing to communication optimization, thereby placing greater emphasis on caring practices. Furthermore, platforms can leverage NCA insights to provide bilateral services for both patients and physicians. This study not only provides a cost-effective, high-precision analytical paradigm to address the challenges of semantic complexity and data sparsity in medical texts but also offers a scientific basis for understanding the online physician-patient interaction process and improving service quality.

ACKNOWLEDGMENT

This work was supported by the Japan Society for the Promotion of Science (Grant Number 24K07926).

REFERENCES

Dul, J., Hauff, S., Bouncken, R.B., 2023. Necessary condition analysis (NCA): review of research topics and guidelines for good practice. Rev. Manag. Sci. 17, 683–714. https://doi.org/10.1007/s11846-023-00628-x

Fan, J., Geng, H., Liu, X., Wang, J., 2022. The effects of online text comments on patients' choices: the mediating roles of comment sentiment and comment content. Front. Psychol. 13. https://doi.org/10.3389/fpsyg.2022.886077

Kong, M., Wang, Y., Li, M., Yao, Z., 2025. Mechanism Assessment of Physician Discourse Strategies and Patient Consultation Behaviors on Online Health Platforms: Mixed Methods Study. J. Med. Internet Res. 27, e54516. https://doi.org/10.2196/54516

Li, P., Hu, Y., 2025. Evaluating online doctor–patient communication quality and its effects: text analysis approach. DIGITAL HEALTH 11, 20552076251395547. https://doi.org/10.1177/20552076251395547

Liu, X., Hu, M., Xiao, B.S., Shao, J., 2022. Is my doctor around me? Investigating the impact of doctors' presence on patients' review behaviors on an online health platform. J. Assoc. Inf. Sci. Technol. 73, 1279–1296. https://doi.org/10.1002/asi.24632

Luo, X., Deng, Z., Yang, B., Luo, M.Y., 2024. Pre-trained language models in medicine: a survey. Artif. Intell. Med. 154, 102904. https://doi.org/10.1016/j.artmed.2024.102904

Madanay, F., Tu, K., Campagna, A., Davis, J.K., Doerstling, S.S., Chen, F., Ubel, P.A., 2024. Classification of Patients' Judgments of Their Physicians in Web-Based Written Reviews Using Natural Language Processing: Algorithm Development and Validation. J. Med. Internet Res. 26, e50236. https://doi.org/10.2196/50236

Shan, W., Wang, Y., Luan, J., Tang, P., 2019. The influence of physician information on patients' choice of physician in mHealth services using china's chunyu doctor app: eye-tracking and questionnaire study. JMIR Mhealth Uhealth 7, e15544. https://doi.org/10.2196/15544

Wang, J., Fu, L., Huang, Z., Hu, K., Lin, Z., Tang, Q., 2025. How AI-powered consultation services in internet hospitals influence patient satisfaction: a structural analysis. DIGITAL HEALTH 11, 20552076251358673. https://doi.org/10.1177/20552076251358673

Wu, H., Lu, N., 2018. Service provision, pricing, and patient satisfaction in online health communities. Int. J. Med. Inf. 110, 77–89. https://doi.org/10.1016/j.ijmedinf.2017.11.009

Yang, B., Lu, W., Xuan, Y., Hao, C., Huang, X., 2025. The influences of social support expressed from doctors and disclosed from peers on patient decision-making: an analysis from the online health community. Sci. Rep. 15, 2703. https://doi.org/10.1038/s41598-024-85023-6

Yang, F., Cheng, Y., Yao, R., Zhang, X., 2025. What key factors affect patient satisfaction on online medical consultation platforms? A case study from China. Health Care (Don Mills) 13, 540. https://doi.org/10.3390/healthcare13050540

Zhang, X., Sun, J., Li, X., Liu, Y., Li, C., 2025. Developing a framework for online review-based health care service quality assessment: text-mining study. J. Med. Internet Res. 27, e66141. https://doi.org/10.2196/66141