**Original Article /** *Özgün Makale*

OPEN ACCESS

# AI-based assessment of abstract quality in otorhinolaryngology journals

*Kulak burun boğaz dergilerinde özet kalitesinin yapay zeka tabanlı değerlendirilmesi*

Kadir Şinasi Bulut[1] , Fatih Gül[2]

*¹Department of Otolaryngology, Ankara Yıldırım Beyazit University, Ankara, Türkiye*
*²Department of Otolaryngology, Lokman Hekim University, Ankara, Türkiye*

**ABSTRACT**

**Objectives:** This study aims to analyze 2024 otorhinolaryngology journal abstracts indexed in the Web of Science (WoS) using an artificial intelligence (AI)-based structured rubric (ChatGPT) to assess quality and explore associations with journal metrics.

**Materials and Methods:** A methodological analysis was conducted on 515 comparative-study abstracts from 66 WoS-indexed journals. Each abstract was evaluated by an AI language model (ChatGPT-5, OpenAI) using a 10-item rubric derived from international reporting standards, scoring 0-100 across originality, aim, design, methods, statistics, results, interpretation, flow, and impact. Journal metrics (SCI/ESCI, quartile, Journal Citation Indicator [JCI]) were retrieved from the WoS database.

**Results:** The mean total quality score was 75.3±7.6 (range, 50 to 94). Highest scores were for clarity of aim and results (91.0±5.6%), while lowest were for study design and sample size. Abstracts in SCI journals (76.0±7.6) scored higher than ESCI (70.2±5.1, p<0.001). Higher quality was also associated with Q1-2 journals and JCI >1 (p<0.001 for both). Quartile ranking showed the highest predictive value (area under the curve [AUC] =0.76).

**Conclusion:** Abstract quality in otorhinolaryngology journals is variable but correlates positively with journal impact. AI-based evaluation offers an objective, efficient approach to assess scientific reporting quality.

**Keywords:** Abstracts, artificial intelligence, otorhinolaryngologic diseases.

**ÖZ**

**Amaç:** Bu çalışmada, yapay zeka (YZ) tabanlı bir değerlendirme aracı (ChatGPT) kullanarak Web of Science (WoS) kapsamında indekslenen 2024 yılı otorinolaringoloji dergi özetleri analiz edildi ve özetlerin dergi metrikleri ile olan ilişkileri incelendi.

**Gereç ve Yöntemler:** Web of Science kapsamında indekslenen 66 dergiden 515 karşılaştırmalı çalışma özeti üzerinde metodolojik bir analiz yapıldı. Her özet, uluslararası raporlama standartlarından türetilen 10 maddelik bir değerlendirme ölçeği kullanılarak bir yapay zeka dil modeli (ChatGPT-5, OpenAI) tarafından değerlendirildi ve özgünlük, amaç, tasarım, yöntemler, istatistikler, sonuçlar, yorumlama, akış ve etki açısından 0-100 arasında puanlandı. Dergi metrikleri (SCI/ESCI, çeyreklik, Journal Citation Indicator [JCI]), WoS veri tabanından alındı.

**Bulgular:** Ortalama toplam kalite puanı 75.3±7.6 (dağılım, 50-94) idi. En yüksek puanlar amaç ve sonuçların açıklığı için alınırken (%91.0±5.6), çalışma tasarımı ve örneklem büyüklüğü için en düşük puanlar alındı. SCI dergilerindeki özetler (%76.0±7.6) ESCI (%70.2±5.1, p<0.001) dergilerine kıyasla daha yüksek puan aldı. Daha yüksek kalite, Q1-Q2 dergilerde yayımlanma ve JCI >1 ile de ilişkiliydi (her ikisi için p<0.001). Çeyreklik sıralama, en yüksek öngörücü değeri gösterdi (eğri altında kalan alan [AUC] =0.76).

**Sonuç:** Otorinolaringoloji dergilerinde özet kalitesi değişkendir, ancak dergi etkisi ile pozitif korelasyon göstermektedir. Yapay zeka tabanlı değerlendirme, bilimsel raporlama kalitesini değerlendirmek için objektif ve etkili bir yaklaşım sunmaktadır.

**Anahtar sözcükler:** Özetler, yapay zeka, kulak burun boğaz hastalıkları.

The abstract section of a scientific article presents the essential information of a study in a concise, clear, and understandable manner, serving as the first point of contact for both readers and editors.[1,2] With the rapid increase in the number of scientific publications, the time that researchers and editors can allocate for evaluation has become increasingly limited.[3,4] Therefore, the abstract plays a critical role in determining whether a manuscript is worth reading, citing, or publishing.[2,5]

In the field of medicine in particular, the quality of the abstract directly affects the visibility of the article in the literature, the accurate understanding of its scientific contribution, and its usability by other researchers.[4,6,7] A sufficient and well-structured abstract effectively reflects the scientific value of a study, whereas an insufficient abstract may reduce the visibility of even a high-quality research work.

Journals indexed in the Web of Science (WoS) database are mainly categorized under the Science Citation Index (SCI) and the Emerging Sources Citation Index (ESCI), with their positions in the scientific community determined by impact factor, journal citation indicator, and quartile ranking. However, comprehensive studies investigating the relationship between abstract quality and a journal's index status or quartile level remain limited.

In the present study, we aimed to evaluate the abstracts of articles published in 2024 in otorhinolaryngology journals indexed in WoS using a systematic checklist based on international standards and artificial intelligence (AI)-based analysis and to analyze the relationship between abstract quality and SCI/ESCI indexing, quartile ranking, and Journal Citation Indicator.

## MATERIALS AND METHODS

### Study design

This study was planned as a methodological analysis examining the abstracts of articles published in 2024 in journals categorized under otorhinolaryngology in the WoS database. For the literature search, a total of 66 journals listed in the WoS database were identified. Among these, research articles published in 2024 employing "comparative study" design were filtered using the PubMed database. From the obtained publications, a total of 515 abstracts that adhered to the IMRAD structure (Introduction, Methods, Results, and Discussion) were included in the analysis. Ethics committee approval was not required for this study because it involved analysis of publicly available, previously published journal abstracts and did not include human participants, animals, or identifiable personal data.

### Evaluation tool and criteria

The abstracts were evaluated using a 10-item scoring rubric developed based on international standards of scientific writing and editorial guidelines including the International Committee of Medical Journal Editors (ICMJE), Consolidated Standards of Reporting Trials (CONSORT), Strengthening the Reporting of Observational Studies in Epidemiology (STROBE), Committee on Publication Ethics (COPE), and Standards for Reporting of Diagnostic Accuracy Studies (STARD).[8-10] The rubric was adapted from literature reviews, international guidelines, and peer review forms, and was further structured with the support of ChatGPT (OpenAI).

Inclusion criteria were as follows: originality and contribution to the literature (Originality & Contribution), which assessed whether the study addressed a familiar topic without novelty or instead filled a clear gap in the literature with a meaningful contribution. Clarity of the study aim (Clarity of Aim) examined whether the research objective was explicitly stated, measurable, and clearly defined. Strength of study design and consistency with the literature (Study Design) evaluated whether the methodology was weak and inconsistent or instead reflected a robust, prospective, and literature-aligned design. Adequacy of sample size and statistical power analysis (Sample Size & Power) considered whether the study relied on an insufficient or arbitrary sample, or whether it reported an adequate sample size with an appropriately conducted power analysis. Clarity and reproducibility of methods (Methodological Detail) focused on whether the methods were vaguely described and non-reproducible or instead transparently and comprehensively reported with bias-reducing strategies. Appropriateness and accuracy of statistical analyses (Quality of Statistical Analysis) assessed whether inadequate or incorrect statistical approaches were applied versus appropriate methods accurately performed and interpreted. Clarity of results and presentation of numerical data (Clarity of Results) distinguished between abstracts that reported vague or incomplete results and those that provided clear, comprehensive findings including p-values, confidence intervals, and effect sizes. Interpretation of results and clinical relevance (Interpretation of Results) evaluated whether the discussion was speculative and inconsistent or instead critically aligned with existing evidence while emphasizing clinical implications. Coherence, flow, and consistency across sections of

the abstract (Scientific Flow & Structure) assessed whether the abstract was fragmented and disorganized or demonstrated strong logical integrity and smooth transitions across sections. Finally, potential clinical or practical impact of the findings (Practical/Clinical Impact) reflected whether the study offered little or no practical relevance or whether it provided highly applicable insights with the potential to influence clinical practice and guideline recommendations (Table 1).

Each criterion was scored on a six-point scale ranging from 0 (very poor) to 10 (excellent), with detailed descriptors defined for each level (e.g., "0 = well-known topic; no originality" to "10 = clear literature gap; strong contribution"). This approach ensured that the evaluation was based on systematic and comparable measures rather than subjective judgment.

### Evaluation process

All abstracts were independently evaluated and scored by an AI-based language model (ChatGPT, OpenAI) according to the 10-item rubric. No human observer intervention was involved; the scoring was performed solely by AI. A total quality score (0-100) was calculated for each abstract.

### Additional variables

For each abstract, the following journal metrics were retrieved from the WoS database and included in the analysis: SCI or ESCI indexing status, Quartile ranking (Q1, Q2, Q3, Q4), Impact factor, Journal Citation Indicator (JCI) score.

### Statistical analysis

Statistical analysis was performed using the SPSS Statistics version 29.0 software (IBM Corp., Armonk, NY, USA). Descriptive data were calculated for each of the 10 evaluation criteria and expressed in mean ± standard deviation (SD), median (min-max), or number and frequency. The normality of the distribution of total abstract quality scores was assessed using the Shapiro-Wilk test. As the data showed a normal distribution, parametric tests were applied. Comparisons of total abstract quality scores between groups were performed using independent samples t-tests. Differences were evaluated according to citation index status (SCI vs. ESCI), quartile groups (Q1-2 *vs.* Q3-4), and Journal Citation Indicator (JCI <1 *vs.* >1). Results were presented with mean ± SD, 95% confidence intervals (CI), and p-values. The predictive ability of abstract quality scores for journal metrics was further assessed using

| Table 1 Criteria for assessing abstract quality | | Score |
|---|---|---|
| Originality & contribution | To what extent does the study fill a gap in the literature and provide an original contribution? | 0-10 |
| Clarity of aim | Is the research aim clear, measurable, and well defined? | 0-10 |
| Study design | Does the study design demonstrate methodological strength and alignment with existing literature? | 0-10 |
| Sample size & power | Is the sample size sufficient, and has an appropriate power analysis been conducted and reported? | 0-10 |
| Methodological detail | Are the methods described in sufficient detail to ensure reproducibility, and were bias-reducing strategies implemented? | 0-10 |
| Quality of statistical analysis | Are the statistical analyses appropriate, accurate, and do they include advanced methods when necessary? | 0-10 |
| Clarity of results | Are the results transparent, well-structured, and adequately supported by complete reporting of statistical values? | 0-10 |
| Interpretation of results | Are the results interpreted consistently with the literature, and is their clinical significance clearly highlighted? | 0-10 |
| Scientific flow & structure | Does the manuscript demonstrate logical flow and coherence between sections? | 0-10 |
| Practical/clinical impact | What is the potential practical or clinical impact of the study-can it influence practice guidelines or decision-making? | 0-10 |

the receiver operating characteristic (ROC) curve analysis. Optimal cut-off values were determined by the Youden index. For each indicator (JCI, quartile, citation index), sensitivity, specificity, area under the curve (AUC), and 95% CI were reported. Additionally, binary logistic regression analyses were performed to determine the independent association between total abstract quality scores and journal metrics (JCI, quartile status, and citation index). Results were expressed as regression coefficients (B), odds ratios (OR) with 95% CI, and Nagelkerke $R^2$ values to indicate model fit. A $p$ value of <0.05 was considered statistically significant.

## RESULTS

Descriptive statistics of the quality scores of the 515 included abstracts showed a mean total score of 75.32±7.58 (range, 50 to 94), with a wide distribution.

At the criterion level, the highest mean scores were observed for Clarity of Aim (8.72±0.98), Clarity of Results (8.37±1.02), and Scientific Flow & Structure (8.23±0.88). In contrast, the lowest mean scores were recorded for Sample Size & Power (6.35±1.44) and Study Design (6.59±1.10). A particularly wide variability was noted for the sample size and statistical

**Table 2**
Descriptive statistics of abstract quality scores across evaluation criteria

| | Mean±SD | Min-Max | Range |
|---|---|---|---|
| Questions | | | |
| Originality & contribution | 7.00±1.08 | 4-10 | 6 |
| Clarity of aim | 8.72±0.98 | 6-10 | 4 |
| Study design | 6.59±1.10 | 4-10 | 6 |
| Sample size & power | 6.35±1.44 | 2-10 | 8 |
| Methodological detail | 7.17±0.99 | 4-8 | 4 |
| Quality of statistical analysis | 7.20±1.18 | 2-10 | 8 |
| Clarity of results | 8.37±1.02 | 6-10 | 4 |
| Interpretation of results | 7.92±1.14 | 6-10 | 4 |
| Scientific flow & structure | 8.23±0.88 | 6-10 | 4 |
| Practical/clinical impact | 7.71±0.95 | 4-10 | 6 |
| Total | 75.32±7.58 | 50-94 | 44 |

SD: Standard deviation.

**Table 3**
Comparison of mean scores according to citation index, quartile, and JCI score

| | n | % | Mean±SD | 95% CI Lower | Upper | $p$ |
|---|---|---|---|---|---|---|
| Citation index | | | | | | |
| SCI | 455 | 88 | 76.00±7.60 | 3.81 | 7.78 | <0.001 |
| ESCI | 60 | 12 | 70.20±5.11 | | | |
| Quartile | | | | | | |
| Q1-2 | 391 | 76 | 76.85±7.42 | 4.91 | 7.78 | <0.001 |
| Q3-4 | 124 | 24 | 70.50±5.93 | | | |
| JCI score | | | | | | |
| <1 | 257 | 50 | 72.85±7.13 | 6.17 | 3.68 | <0.001 |
| >1 | 258 | 50 | 77.78±7.23 | | | |

CI: Confidence interval; SD: Standard deviation; SCI: Science Citation Index; ESCI: Emerging Sources Citation Index; JCI: Journal citation indicator.

**Table 4**

Logistic regression results for the association between total score and journal metrics

|  | B | OR | 95% CI Lower | 95% CI Upper | p | Nagelkerke R square |
|---|---|---|---|---|---|---|
| JCI | 0.097 | 1.102 | 1.07 | 1.13 | <0.001 | 0.142 |
| Quartile | 0.121 | 0.886 | 0.859 | 0.915 | <0.001 | 0.185 |
| Citation index | 0.098 | 0.907 | 0.874 | 0.94 | <0.001 | 0.110 |

CI: Confidence interval; OR: Odds ratio; JCI: Quartile, citation index.

**Table 5**

ROC curve analysis of abstract quality scores for predicting JCI, quartile, and citation index

|  | Cut off value | Sensitivity | Specificity | Area under curve | 95% CI Lower | 95% CI Upper | p |
|---|---|---|---|---|---|---|---|
| JCI | 76.50 | 75.5 | 58.1 | 0.691 | 0.646 | 0.737 | <0.001 |
| Quartile | 75.00 | 79.8 | 62.7 | 0.762 | 0.719 | 0.806 | <0.001 |
| Citation Index | 73.50 | 70 | 66.6 | 0.748 | 0.695 | 0.801 | <0.001 |

CI: Confidence interval; JCI: Quartile, citation index.

power criterion, indicating inconsistent reporting in this domain (Table 2).

According to journal indexing, abstracts published in SCI journals had significantly higher quality scores (76.0±7.6) compared with those in ESCI journals (70.2±5.1, p<0.001). Quartile analysis revealed that abstracts in Q1-2 journals (76.85±7.42) scored significantly higher than those in Q3-4 journals (70.50±5.93, p<0.001). Similarly, abstracts from journals with JCI >1 (77.78±7.23) had significantly higher quality scores than those from journals with JCI <1 (72.85±7.13, p<0.001) (Table 3).

Binary logistic regression analysis revealed that all three journal metrics were significantly associated with total score. Quartile ranking demonstrated the strongest predictive effect, with higher quartiles (Q3-4) being associated with lower total scores (B=0.121, OR=0.886, 95% CI: 0.859-0.915, p<0.001, Nagelkerke $R^2$=0.185). The JCI score also emerged as a significant predictor, where each unit increase in JCI was associated with a 10% increase in the odds of achieving a higher total score (B=0.097, OR=1.102, 95% CI: 1.07-1.13, p<0.001, Nagelkerke $R^2$=0.142). Citation index (SCI *vs.* ESCI) showed the weakest but still significant effect, with ESCI journals exhibiting lower total scores (B=0.098, OR=0.907, 95% CI: 0.874-0.94, p<0.001, Nagelkerke $R^2$=0.110). Overall, quartile ranking accounted for the greatest

proportion of variance in total score, followed by JCI score and citation index (Table 4).

According to the ROC analysis, specific cut-off values of abstract quality scores demonstrated significant discriminatory accuracy for journal metrics. For JCI, a cut-off value of 76.5 yielded a sensitivity of 75.5% and a specificity of 58.1%,
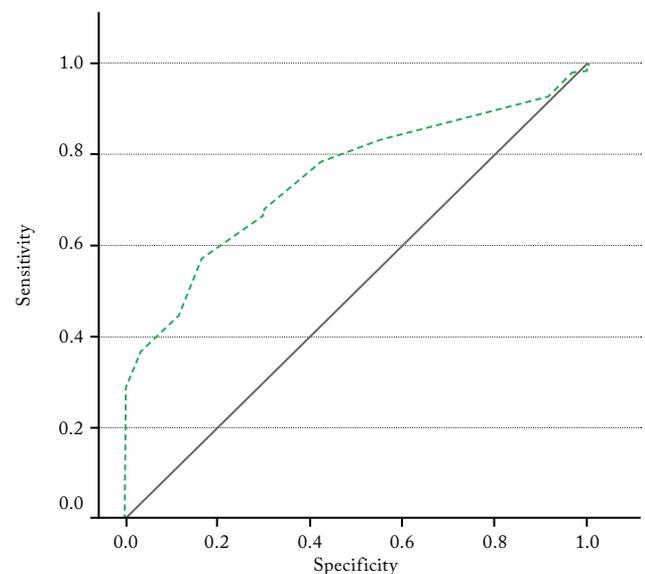


**Figure 1.** ROC curve for abstract quality scores in predicting citation index status (SCI *vs.* ESCI).

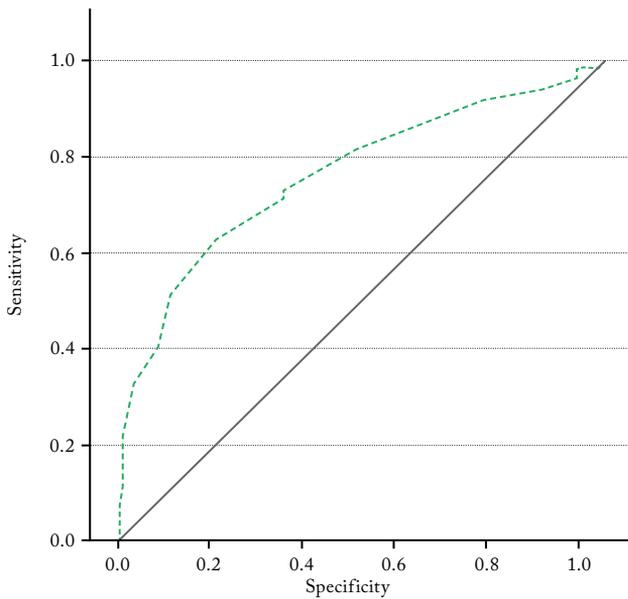ROC: Receiver operating characteristic; SCI: Science Citation Index; ESCI: Emerging Sources Citation Index.

**Figure 2.** ROC curve for predicting quartile classification (Q1-2 *vs.* Q3-4) using abstract quality scores.

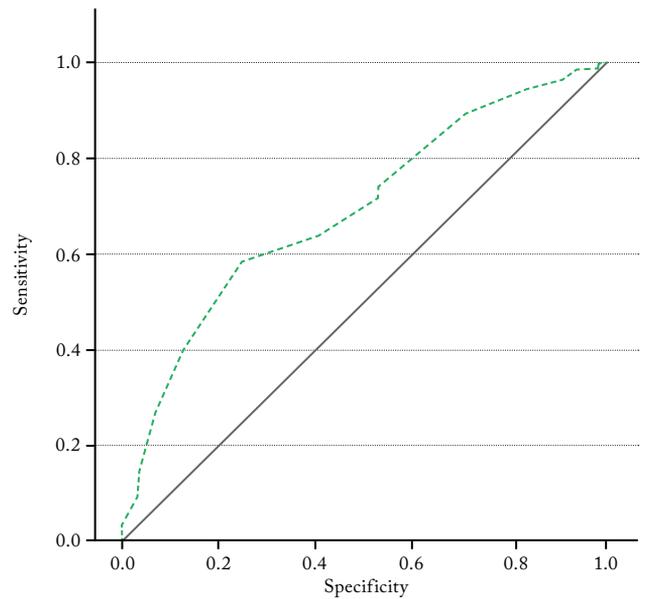ROC: Receiver operating characteristic; Q: Quartile.



**Figure 3.** ROC curve for predicting journal citation indicator (JCI >1 *vs.* <1) based on abstract quality scores.

ROC: Receiver operating characteristic; JCI: Journal citation indicator.

significantly distinguishing journals with higher JCI values (AUC: 0.691, p<0.001). For quartile classification, a cut-off of 75.0 provided the highest discriminative power, with 79.8% sensitivity and 62.7% specificity, successfully predicting Q1-2 journals (AUC: 0.762, p<0.001). For citation index (SCI *vs.* ESCI), a cut-off of 73.5 achieved 70% sensitivity and 66.6% specificity, also indicating significant discriminatory ability (AUC: 0.748, p<0.001) (Table 5, Figures 1-3).

## DISCUSSION

In the present study, we systematically evaluated the abstracts of articles published in otorhinolaryngology journals covered by WoS in 2024 using an AI-based language model (ChatGPT-5, OpenAI) using a 10-item structured rubric based on international editorial and reporting guidelines. Our study findings demonstrated that overall abstract quality was moderate to high, with a mean score of 75.3 out of 100, yet with substantial variability across criteria and journal metrics.

Among individual criteria, clarity of aim, clarity of results, and scientific flow emerged as the strongest domains, suggesting that most abstracts succeed in presenting the study objectives and findings in a coherent manner. Conversely, the lowest scores were observed in the domains of sample size and power, as well as study design. This indicates that essential

methodological details, such as justification of sample size or explicit mention of statistical power, are frequently underreported in abstracts, despite their importance for evaluating scientific rigor. Similar deficiencies have been highlighted in prior evaluations of medical abstracts, underscoring a persistent gap between reporting standards and actual practice.[2,4]

The comparative analyses revealed that abstracts published in SCI-indexed, higher-quartile (Q1-2), and higher-JCI (>1) journals consistently demonstrated significantly higher quality scores. These results suggest that abstract quality is not only a reflection of individual author practices, but also linked to editorial standards and peer-review stringency within higher-impact journals.[2,9] In particular, the quartile classification showed the strongest discriminative performance in ROC analysis (AUC: 0.762), indicating that abstract quality could serve as a practical marker of journal stratification.

Logistic regression analysis further confirmed that total abstract quality score was independently associated with journal citation indicators. Abstracts with higher scores were more likely to be published in SCI journals, in higher-quartile journals, and in journals with JCI >1. This provides empirical evidence supporting the hypothesis that well-structured and methodologically transparent abstracts contribute to dissemination in higher-ranking journals.[5]

The study has several important implications. First, systematic and rubric-based evaluation of abstracts can highlight strengths and weaknesses in scientific reporting and potentially serve as a quality improvement tool for authors. Second, the demonstrated association between abstract quality and journal metrics suggests that improving abstract reporting may enhance visibility, citation potential, and acceptance likelihood. Third, the successful application of an AI-based language model (ChatGPT-5, OpenAI) for independent scoring indicates that AI can be integrated into academic peer-review processes, particularly for large-scale, standardized evaluations.[4] Our findings align with Yin et al.,[11] who reported that medical journals were increasingly adopting guidelines for the use of generative AI, particularly in high-ranking journals. This supports the feasibility of integrating AI-based abstract evaluations, such as our ChatGPT-5 approach, into editorial workflows. In line with these implications, Fiorillo and Mehta[12] recently showed that the integration of AI-assisted peer review could significantly accelerate editorial workflows without undermining scientific rigor. Their study demonstrated that incorporating models such as ChatGPT into the early stages of manuscript evaluation could reduce review times, improve efficiency, and provide timely feedback to authors. This evidence supports our finding that AI-based evaluation of abstracts is not only feasible, but may also contribute to faster and more consistent publication processes in scientific journals.

Nevertheless, some limitations should be acknowledged. First, the evaluation relied solely on abstract content and did not include full-text analysis; therefore, methodological rigor may have been underrepresented. In addition, although the rubric was adapted from international standards, abstract assessment inherently carries some subjectivity, even when supported by AI algorithms. Furthermore, this was a cross-sectional study limited to 2024 publications in otorhinolaryngology, which may restrict the generalizability of results across other specialties or timeframes.

In conclusion, abstract quality in otorhinolaryngology journals is heterogenous, with notable strengths in clarity but persistent weaknesses in methodological reporting. Higher scores were significantly associated with SCI indexing, higher quartiles, and stronger citation indicators, suggesting that abstract quality plays a critical role in determining scientific impact. Integrating AI-based abstract assessment tools may provide an innovative approach to improving the efficiency and objectivity of scientific evaluation in the future.

## REFERENCES

1. Andrade C. How to write a good abstract for a scientific paper or conference presentation. Indian J Psychiatry 2011;53:172-5. doi: 10.4103/0019-5545.82558.

2. Timmer A, Sutherland LR, Hilsden RJ. Development and evaluation of a quality score for abstracts. BMC Med Res Methodol 2003;3:2. doi: 10.1186/1471-2288-3-2.

3. Millar BC, Lim M. The Role of visual abstracts in the dissemination of medical research. Ulster Med J 2022;91:67-78.

4. Korevaar DA, Cohen JF, Hooft L, Bossuyt PM. Literature survey of high-impact journals revealed reporting weaknesses in abstracts of diagnostic accuracy studies. J Clin Epidemiol 2015;68:708-15. doi: 10.1016/j.jclinepi.2015.01.014.

5. Simera I, Moher D, Hirst A, Hoey J, Schulz KF, Altman DG. Transparent and accurate reporting increases reliability, utility, and impact of your research: Reporting guidelines and the EQUATOR Network. BMC Med 2010;8:24. doi: 10.1186/1741-7015-8-24.

6. Bahadoran Z, Mirmiran P, Kashfi K, Ghasemi A. The principles of biomedical scientific writing: Abstract and keywords. Int J Endocrinol Metab 2020;18:e100159. doi: 10.5812/ijem.100159.

7. Mitchell NS, Stolzmann K, Benning LV, Wormwood JB, Linsky AM. Effect of a scoring rubric on the review of scientific meeting abstracts. J Gen Intern Med 2020;36:2483-5. doi: 10.1007/s11606-020-05960-6.

8. Hopewell S, Chan AW, Collins GS, Hróbjartsson A, Moher D, Schulz KF, et al. CONSORT 2025 statement: Updated guideline for reporting randomised trials. BMJ 2025;389:e081123. doi: 10.1136/bmj-2024-081123.

9. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. J Clin Epidemiol 2008;61:344-9. doi: 10.1016/j.jclinepi.2007.11.008.

10. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;351:h5527. doi: 10.1136/bmj.h5527.

11. Yin S, Huang S, Xue P, Xu Z, Lian Z, Ye C, et al. Generative Artificial Intelligence (GAI) usage guidelines for scholarly publishing: A cross-sectional study of medical journals. BMC Med 2025;23:77. doi: 10.1186/s12916-025-03899-1.

12. Fiorillo L, Mehta V. Accelerating editorial processes in scientific journals: Leveraging AI for rapid manuscript review. OOR 2024;10:100511. doi: 10.1016/j.oor.2024.100511.