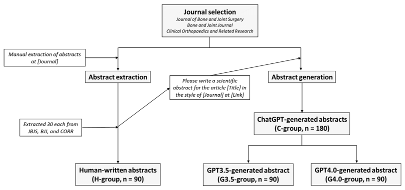
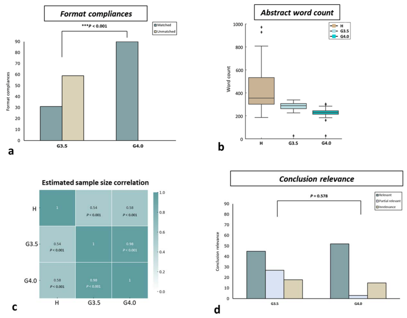
Can ChatGPT effectively generate abstracts in orthopedic surgery? A comparative analysis between human-written and ChatGPT-generated scientific abstracts

Hong Jin Kim[1], Pil Whan Yoon[2], Jae Youn Yoon[2], Jun-Ki Moon[3]
[1]Orthopedic Surgery, Inje University Sanggye Paik Hospital, [2]Orthopedic Surgery, Seoul Now Hospital, [3]Orthopedic Surgery, College of Medicine, Chung-Ang University Hospital

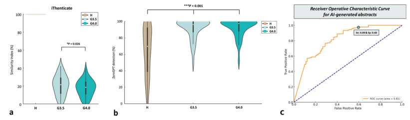
INTRODUCTION:
With the innovative development of large language models (LLMs) in artificial intelligence (AI), ChatGPT can generate logical and coherent text within just one minute using a simple prompt. The capabilities of ChatGPT have been demonstrated to be sufficient to pass several standardized assessments of medical knowledge such as the United States Medical Licensing Examination (USMLE) and Orthopedic In-Training Examination (OITE). Meanwhile, the plausible ability to tell a lie in ChatGPT was recognized as a hallucination, which is indistinguishable from humans. Given the limited evidence, little information has been documented on the capability of ChatGPT for writing an abstract in orthopedic surgery
METHODS:
This cross-sectional study aims to assess the reproducibility of chatGPT-generated abstracts in orthopedic surgery based on the titles that have already been published in the Journal of Bone and Joint Surgery (JBJS), the Bone and Joint Journal (BJJ), and the Clinical Orthopaedics and Related Research (CORR). A total of 90 human-written abstracts (H-group) were randomly extracted from JBJS (n = 30), BJJ (n = 30), and CORR (n = 30). With the prompt “Please write a scientific abstract for the article [title] in the style of [journal] at [link].”, we collected a total of 180 ChatGPT-generated abstracts, which were divided into two groups according to the GPT version: GPT3.5-generated abstracts (G3.5 group, n = 90) and GPT4.0-generated abstracts (G4.0 group, n = 90) (Figure 1). We compared the journal’s format compliance, word count, estimated sample size, and conclusion relevance between G3.5 and G4.0 groups based on the human-written abstracts. The similarity index using the iThenticate and AI detection rates using the ZeroGPT program were respectively measured to evaluate the plagiarism (in the case of similarity index ≥ 15%) and the detection capability of ChatGPT-generated abstracts. We also analyzed reliability using Cohen’s kappa to assess the distinguishability by humans between human-written and ChatGPT-generated abstracts.
RESULTS:
The ChatGPT-generated abstracts met the journal’s format guidelines in 34.4% of cases in the GPT3.5 group and 100% in the GPT4.0 group, with a statistically significant difference ($P < 0.001$) (Figure 2a). The mean word count was 432.9 in the H group, 285.2 in the G3.5 group, and 230.3 in the G4.0 group (Figure 2b). The ChatGPT-generated abstracts met the journal’s word count guidelines were 86.7% in the G3.5 group and 100% in the G4.0 group, with a statistically significant difference ($P < 0.001$). The ChatGPT-generated abstracts with predicted study designs that matched the human-written abstracts were 74.4% in the GPT3.5 group and 75.6% in the GPT4.0 group with no statistical difference ($P = 0.863$). The estimated sample size in ChatGPT-generated abstracts was significantly correlated with real sample size in human-written abstracts ($r = 0.54$, $P < 0.001$ between H group and G3.5 group; $r = 0.58$, $P < 0.001$ between H group and G4.0 group) (Figure 2c). Regarding the relevance of conclusions, there was no statistical difference between the two groups ($P = 0.578$) (Figure 2d). The mean similarity index using the iThenticate was 20.8% in the G3.5 group and 17.5% in the G4.0 group with a statistical significance ($P = 0.026$) (Figure 3a). However, there was no statistical difference in plagiarism between the two groups ($P = 0.160$). Regarding the ZeroGPT as a program for AI detection, the mean detection rate was 63.9% in the H group, 91.4% in the G3.5 group, and 92.3% in the G4.0 group (Figure 3b). The receiver operative characteristic curve presented a 0.81 curve area with a sensitivity of 0.89 and a specificity of 0.60 (Figure 3c). The Cohen’s kappa for one human assessment showed 0.25, indicating minimal agreement.
DISCUSSION AND CONCLUSION:
Writing a scientific article in the field of orthopedic surgery can be helpful by using ChatGPT, but relying solely on ChatGPT can be unethical considering the high plagiarism and AI detection rate. In particular, humans can not accurately distinguish between human-written and ChatGPT-generated abstracts. Therefore, our study indicate that the need for establishing ethical standards for the use of LLMs such as ChatGPT in writing scientific papers in the field of orthopedic surgery.



Journal selection
Journal of Bone and Joint Surgery
Bone and Joint Journal
Clinical Orthopaedics and Related Research
Manual extraction of abstracts at [Journal]
Abstract extraction
Please write a scientific abstract for the article [Title] in the style of [Journal] at [Link]
Abstract generation
Extracted 30 each from JBJS, BJJ, and CORR
ChatGPT-generated abstracts (C-group, n = 180)
Human-written abstracts (H-group, n = 90)
GPT3.5-generated abstract (G3.5-group, n = 90)
GPT4.0-generated abstract (G4.0-group, n = 90)



Format compliances
***P < 0.001
Matched
Unmatched
Format compliances
G3.5
G4.0
a
Abstract word count
H
G3.5
G4.0
Word count
b
Estimated sample size correlation
H
G3.5
G4.0
1
0.54 P < 0.001
0.58 P < 0.001
0.98 P < 0.001
c
Conclusion relevance
P = 0.578
Relevant
Partial relevant
Inrelevance
Conclusion relevance
G3.5
G4.0
d



iThenticate
*P = 0.026
Similarity index (%)
H
G3.5
G4.0
a
***P < 0.001
ZeroGPT detection (%)
b
Receiver Operative Characteristic Curve for AI-generated abstracts
Sn: 0.89 & Sp: 0.60
True Positive Rate
False Positive Rate
ROC curve (area = 0.81)
c