Comparative study of ChatGPT and human evaluators on the assessment of medical literature according to recognised reporting standards


ABSTRACT

Introduction Amid clinicians’ challenges in staying updated with medical research, artificial intelligence (AI) tools like the large language model (LLM) ChatGPT could automate appraisal of research quality, saving time and reducing bias. This study compares the proficiency of ChatGPT against human evaluation in scoring abstracts to determine its potential as a tool for evidence synthesis.

Methods We compared ChatGPT’s scoring of implant dentistry abstracts with human evaluators using the Consolidated Standards of Reporting Trials for Abstracts reporting standards checklist, yielding an overall compliance score (OCS). Bland-Altman analysis assessed agreement between human and AI-generated OCS percentages. Additional error analysis included mean difference of OCS subscores, Welch’s t-test and Pearson’s correlation coefficient.

Results Bland-Altman analysis showed a mean difference of 4.92% (95% CI 0.62%, 0.37%) in OCS between human evaluation and ChatGPT. Error analysis displayed small mean differences in most domains, with the highest in ‘conclusion’ (0.764 (95% CI 0.186, 0.280)) and the lowest in ‘blinding’ (0.034 (95% CI 0.818, 0.895)). The strongest correlations between were in ‘harm’ (r=0.32, p<0.001) and ‘trial registration’ (r=0.34, p=0.002), whereas the weakest were in ‘intervention’ (r=0.02, p<0.001) and ‘objective’ (r=0.06, p=0.001).

Conclusion LLMs like ChatGPT can help automate appraisal of medical literature, aiding in the identification of accurately reported research. Possible applications of ChatGPT include integration within medical databases for abstract evaluation. Current limitations include the token limit, restricting its usage to abstracts. As AI technology advances, future versions like GPT4 could offer more reliable, comprehensive evaluations, enhancing the identification of high-quality research and potentially improving patient outcomes.

INTRODUCTION

In the dynamic landscape of medical research, clinicians face the daunting challenge of staying abreast of the latest advancements amid their demanding clinical responsibilities. The rate and varying quality of emerging research further compounds this challenge. A number of appraisal tools exist to help readers assess the quality of the reported research, although these can also be time-consuming to employ and are at risk of user bias. The use of large language models (LLMs) like ChatGPT has the potential to automate this evaluation, thereby aiding clinicians in making informed decisions. However, the accuracy of LLMs compared with human expertise as a gold standard remains uncertain. In November 2023, OpenAI unveiled ChatGPT, a generative pretrained transformer (GPT) language model grounded in transformer architecture, which empowers it to process vast amounts of text data and generate coherent text outputs by discerning the relationships between input and output sequences. ChatGPT has been trained on extensive human language datasets, and several studies attest to its ability to produce high-quality, coherent text outputs. Clinical research applications of ChatGPT have yielded promising results, suggesting that artificial intelligence could potentially critically appraise abstracts and liberate valuable clinician time. The objective of this study is to compare the proficiency of ChatGPT3, the third iteration of OpenAI’s GPT model, in scoring abstracts against human evaluation as the benchmark. By determining the accuracy and efficiency of these LLMs in assessing research quality, we aim to explore their potential as valuable tools for clinicians in appraisal and evidence synthesis.

METHODS

In this study, we used a previously published paper as the basis of our comparison with ChatGPT. In their study, abstracts from a systematic review on implant dentistry were scored using the Consolidated Standards of
The OCS is a measure of how many of the CONSORT-A items below are included in a given abstract. Each item below is: completely reported, partially reported or not reported.

The 15 Items included in the OCS are as follows with definitions for each domain. Each domain can be completely reported, partially reported or not reported. Each domain is given a score dependent on how it is graded as:

- **Title**:
  - reported completely: the title must include “randomized”, “randomisation”, “RCT” in the title
  - not reported: no report about random assignment in the title
- **Trial Design**:
  - completely reported: must include the words/words “parallel”, “cluster”, “crossover”, “factorial”, “superiority”, “equivivalence”, “noninferiority” or combinations.
  - not reported: no report of the trial design
- **Participants**:
  - completely reported: eligibility criteria (health status) and location and timeframe the study was conducted or in the abstract.
  - partially reported: one eligibility criterion (health status) and location and timeframe are in the abstract.
  - not reported: no report of the eligibility criteria, location and timeframe.
- **Interventions**:
  - completely reported: description of test and control group treatment.
  - not reported: no description of treatment
- **Objective**:
  - completely reported: 1 objective or primary objective clearly indicated, clearly described.
  - not reported: no report of the objective
- **Outcome**:
  - completely reported: defined primary outcomes for the study or primary endpoints of the study reported.
  - partially reported: only 1 outcome assessed and clearly in the abstract.
  - not reported: no information about primary outcomes or endpoints
- **Randomisation**:
  - completely reported: information in the abstract about how they randomised the participants.
  - not reported: no information about the randomisation process
- **Blinding**:
  - completely reported: information about which people were blinded/masked (participants, caregivers and outcome assessors) in the abstract.
  - not reported: no information about masking
- **Numbers randomised**:
  - completely reported: must state the number of participants randomly allocated to each of the groups evident in the abstract or is easily understood.
  - partially reported: number can be added up in the abstract but is not directly reported.
  - not reported: number of participants in each group is not reported and cannot be calculated.
- **Numbers analysed**:
  - completely reported: must state the number of participants analysed in each of the groups evident in the abstract.
  - partially reported: number can be added up in the abstract but is not directly reported.
  - not reported: number of participants in each group is not reported and cannot be calculated.
- **Outcome 1**:
  - completely reported: reported a primary outcome, results for each group, effect size and a measure for its precision (confidence interval).
  - partially reported: one or more of the following items: results for each group, effect size, a measure for its precision (confidence interval).
  - not reported: no data reported or only descriptive (e.g., The survival rate at 6 months was lower in the test group. The survival rate at 12 months higher in the test group.)
- **Harms**:
  - completely reported: reported which specific adverse events, side-effects or complications occurred (e.g., pain, swelling, post-op bleeding, necrosis, chipping, fractures etc.) or that no such events occurred (can also be part of an outcome that is considered a measure for harms/side effects).
  - partially reported: reported about the existence of complications but did not further describe them (can also be part of an outcome that is considered a measure for harms/side effects).
  - not reported: no information about side-effects etc.
- **Conclusion**:
  - completely reported: conclusion stated.
  - not reported: no conclusion was stated.
  - Trial registration:
  - completely reported: Trial registration number was reported in the abstract.
  - not reported: no information about trial registration.
  - Funding:
  - completely reported: Source of funding was reported in the abstract.
  - not reported: no information about funding.

The OCS is calculated by taking the number of completely reported items (C), multiplying that by 1, taking the number of items that were partially reported (P), multiplying that by 0.5, and then adding together these three numbers. The maximum score is 15. Please give OCS and OCS%.

![Figure 1](A) Example prompt used to generate the OCS as per CONSORT-A criteria. (B) An example of the calculated OCS and OCS% as generated by ChatGPT. CONSORT-A, Consolidated Standards of Reporting Trials for Abstracts; OCS, overall compliance score.
Reporting Trials for Abstracts (CONSORT-A)\(^6\) statement by the human authors of the study. The processes of selection and data extraction were performed independently and in duplicate by two clinician reviewers across a sample of 30 abstracts. Discrepancies were systematically addressed through discussion until a consensus of at least 80\% was achieved. Subsequent data extraction was conducted solely by one reviewer. The CONSORT-A checklist scores abstract reporting standards based on well-defined definitions for subsections such as trial design, blinding and randomisation. The human evaluators scored each item as fully reported, partially reported or not reported. ChatGPT was used to score the same set of abstracts, using a prompt to assess for each domain within the CONSORT-A checklist (figure 1). Building on the methodology established, each constituent subgroup was subsequently scored and categorised into one of the three classifications (figure 1A). An overall compliance score (OCS) was given out of 15, along with an OCS percentage (figure 1B). This was performed using the GPT3.5 model.

Bland-Altman analysis was used to evaluate the overall agreement between human and ChatGPT-generated OCS percentage. For error analysis, the mean difference of the absolute OCS subscores, Welch’s two-sample t-test and Pearson’s correlation coefficient were undertaken. The mean difference provides information on the magnitude and direction of the differences in OCS between ChatGPT and human evaluators, while the Pearson’s correlation coefficient provides information on the strength and direction of the linear relationship between the two sets of scores. This provided complementary information on the agreement between ChatGPT and human evaluator. The Pearson’s correlation coefficient was interpreted based on magnitude: r, 0–0.19 very weak, 0.2–0.39 weak, 0.40–0.59 moderate, 0.6–0.79 strong and 0.8–1 very strong correlation. Statistical analysis was done in R (V.4.1.1). P<0.001 was deemed statistically significant.

RESULTS

Bland-Altman analysis revealed a mean difference of 4.92\% (95\% CI 0.62\%, 0.37\%) in OCS percentage (figure 2). Error analysis revealed small mean differences
between human evaluation and ChatGPT in most domains (table 1).

The mean difference in absolute OCS was highest for the ‘conclusion’ domain (0.764, 95% CI: 0.186, 0.280), indicating that ChatGPT differed the most from human evaluators in this domain. In contrast, the domain with the lowest mean difference in absolute OCS was ‘blinding’ (0.034, 95% CI: 0.818, 0.895), indicating that ChatGPT was most accurate in this domain. In terms of correlation, the study found varying levels of correlation between ChatGPT and human evaluators for different domains. For example, the domains with a strong positive correlation were ‘harms’ (r=0.52, p<0.001) and ‘trial registration’ (r=0.34, p=0.002), indicating a high level of consistency between ChatGPT and human evaluators in these domains. On the other hand, ‘intervention’ (r=0.02, p>0.001) and ‘objective’ (r=0.06, p<0.001) domains had very weak correlations, suggesting that ChatGPT’s performance was less consistent with human evaluators in these domains.

CONCLUSION

As the technology continues to evolve and improve, the next iteration of GPT, GPT4, may further enhance the accuracy and efficiency of the tool, allowing for even more reliable and comprehensive evaluations of research. While there are still limitations to this technology, the promise it holds for assisting in the evaluation and identification of high-quality research is a significant step towards improving patient care and outcomes.

REFERENCES


