

Challenges in Detecting Artificial Intelligence in Scientific Writing: An Analysis of Clinical Reviewers and AI Detection Tools in Plastic Surgery



Keenan S. Fine, MS, Emily E. Zona, BS, Aidan W. O'Shea, ScB, Samuel O. Poore, MD, PhD
Division of Plastic and Reconstructive Surgery, University of Wisconsin School of Medicine and Public Health. Madison, WI

Introduction

- ❖ Artificial intelligence (AI) tools are becoming increasingly integrated into the academic writing process. While these tools may assist with efficiency, integrity is a concern.
- ❖ Identifying AI-generated content remains challenging with potential risks including plagiarism, inaccuracies, and ethical concerns in scientific publications.

The primary objective of this study was to evaluate the ability of medical professionals and online AI-detection tools to accurately detect AI involvement in the writing of surgery manuscript passages.

Methods

- ❖ 8 manuscript passages were evaluated by medical students, residents, and attendings (n=24) across four topics.
- ❖ Each passage was either human-written, human-written with AI editing, or fully AI-generated
- ❖ Rater accuracy: calculated as % of passages correctly identified
- ❖ Interrater Reliability: assessed using Fleiss' kappa for human raters
- ❖ AI Detection Tools: Three online tools used to analyze human-written and AI-generated manuscripts.
- ❖ ROC Analysis: Performed to evaluate the accuracy of detection tools in identifying AI-generated content.
- ❖ Agreement Among Tools: Measured using Intraclass Correlation Coefficients (ICC) to assess consistency in AI-generated content percentage estimates.
- ❖ AI tool performance metrics: true positive rate (TPR), false positive rate (FPR), positive predictive value (PPV), and negative predictive value (NPV)
- ❖ Youden's Index: used to assess cutoff point

Results

Human raters

- Overall accuracy: Correctly identified origin 26.5% of the time (range 0% - 75%)
- No significant difference in accuracy among medical students, residents, and attending physicians
- Passages with reading levels above college level had a lower rater accuracy (30.8% vs. 19.4%)
- Highly technical passages had a lower accuracy than semi-technical/narrative passages (28.1% vs 20.8%)

Table 1: Rater Accuracy and Agreement among Human Raters

	Number of ratings	Accuracy	Fleiss's Kappa	Agreement
Human-written	44	14.5%	-0.348	Poor
AI-Generated	88	34.4%	-0.023	Poor
Human-written and AI-edited	44	22.9%	0.025	Poor

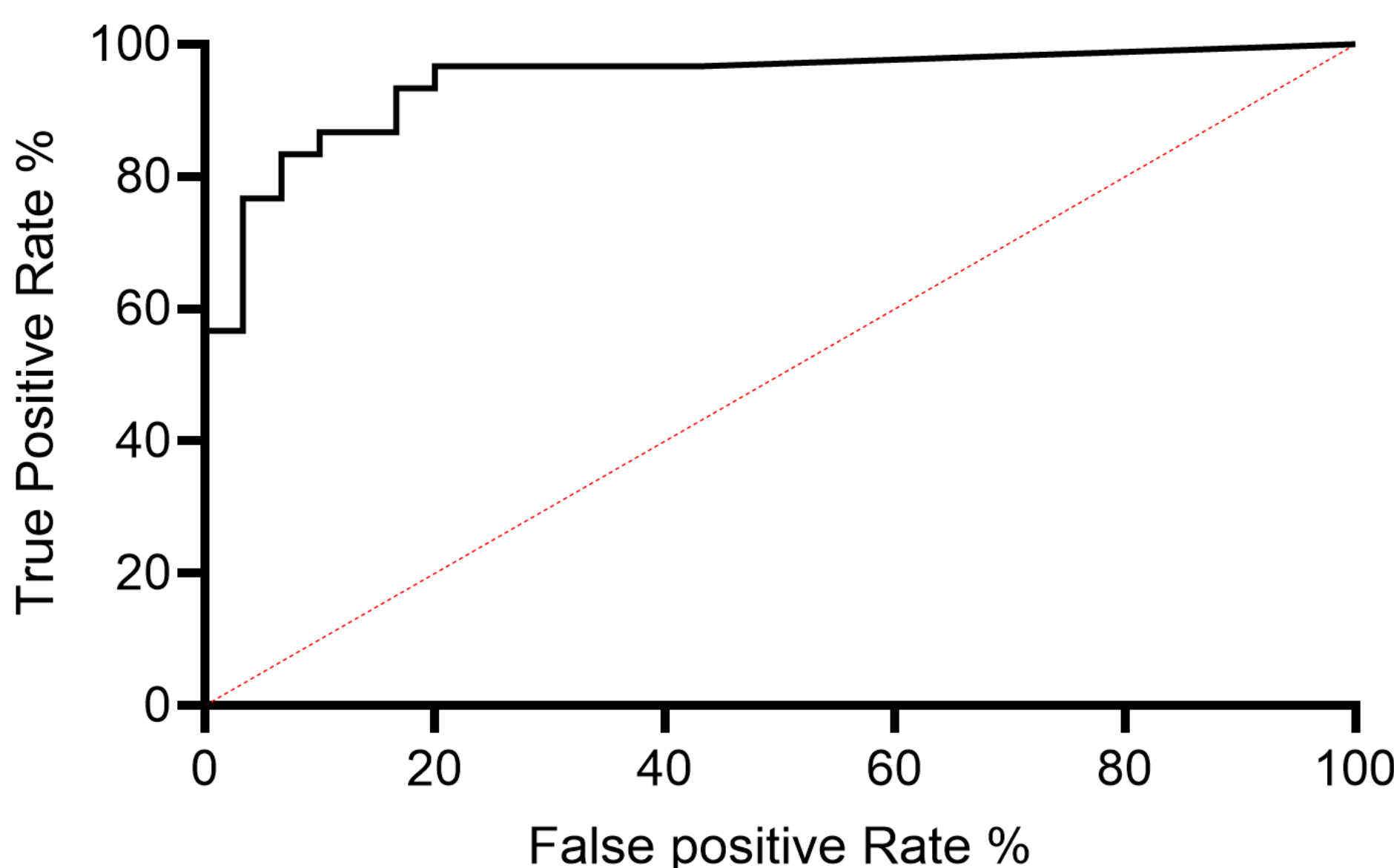
AI Detection Tools

- Tool agreement: ICC was 0.078 among the three detection tools, indicating poor agreement,
- Tool misclassification: High false positive rates at several different cutoff points.

Table 2: Receiver Operating Curve (ROC) Performance Metrics at Key Percentile Thresholds

Cutoff	Sensitivity	Specificity	PPV	NPV	Youden's Index
25%	96.7%	76.7%	80.6%	95.8%	0.73
50%	83.3%	90.0%	89.3%	84.4%	0.73
75%	66.7%	96.7%	95.2%	74.4%	0.63

Figure 1: ROC Curve Analysis of AI Detection Tools



Conclusions

- ❖ Low Detection Accuracy: Medical professionals and AI detection tools showed low accuracy in correctly identifying AI-generated content
- ❖ Interrater Agreement: Human raters demonstrated "poor" interrater reliability, suggesting high variability in their ability to detect AI involvement.
- ❖ Tool Performance: AI detection tools frequently misclassified human-written content as AI-generated indicated by a high FPR
- ❖ Reliability Issues: Poor agreement among AI detection tools underscores the current limitations in reliably distinguishing between human and AI-generated content

Future Directions

- ❖ Conduct a larger study evaluating a broader range of professional backgrounds and AI-detection tools.
- ❖ Assess the effectiveness of formal training programs for medical professionals in improving their ability to detect AI-generated content
- ❖ Study the ethical implications of using AI in scientific writing and develop guidelines for responsible AI use that can be adopted by scientific journals and institutions.