Can Artificial Intelligence Fool Orthopaedic Residency Selection Committees? Analysis of Personal Statements Created by Real Applicants and Generative AI: A Randomized Single-Blind Study

Zachary Lum[1], Lohitha Guntupalli, Augustine M Saiz, Holly Bee Leshikar, Hai Le, John Patrick Meehan, Eric Huish
[1]South Florida Institution Sports Med

INTRODUCTION: The potential capabilities of generative artificial intelligence (AI) tools have been relatively unexplored, particularly in the context of creating personalized statements for medical students applying to orthopaedic surgery residencies. This study aimed to investigate the ability of generative AI, specifically ChatGPT and Google BARD, to generate personal statements and assess whether faculty surgeons on residency selection committees could evaluate differences between real and AI statements.

METHODS: Fifteen real personal statements from fourth-year medical students, comprising 10 statements that were accepted into a residency program and 5 that were not, were selected as training data. These statements were used to train both ChatGPT and Google BARD. Subsequently, the generative AI chatbots were prompted to generate 15 unique and distinct personal statements each, resulting in a total of 45 statements. The statements were then randomized and blinded, and presented to a group of faculty reviewers who have served or are currently serving on a residency selection committee. The faculty members assessed the statements in sequential randomized order using a set of metrics including grammar, word usage, punctuation, sentence/paragraph structure, overall organization, originality, articulation, compelling nature, English proficiency, reasons for choosing orthopaedic surgery, personal interests, career goals, and relevance to the residency program using a Likert scale (ranging from 1 to 5). Finally, faculty were asked to determine whether each personal statement was AI-generated or real, written by a medical student. A comparison of all metrics was conducted between the personal statements, BARD-generated statements, ChatGPT-generated statements, and those written by medical students.

RESULTS:

Faculty correctly identified 88% (79/90) real statements, 90% (81/90) BARD, and 44% (40/90) ChatGPT statements. Accuracy of identifying real and BARD statements was 89%, but this dropped to 74% when including ChatGPT. Reviewers identified statements written by AI (ChatGPT or Bard) with 67% sensitivity and 88% specificity. Additionally, the accuracy did not increase as faculty members reviewed more personal statements, with an AUC of 0.498 (p=0.966).

Using ordinal logistic regression, BARD performed poorer than both REAL and ChatGPT across all metrics (p<0.001). Comparing REAL with ChatGPT, there were no differences in most metrics, except for Personal Interests, Reasons for Choosing Orthopaedic Surgery, Career Goals, Compelling Nature, and Originality, and favoring the REAL personal statements (p=0.05, p=0.028, p=0.015, p=0.001, p=0.001, respectively).

DISCUSSION AND CONCLUSION: Faculty members accurately identified real personal statements and those generated by BARD, but ChatGPT deceived them 56% of the time. Interestingly, accuracy didn't improve as faculty members reviewed more personal statements, with an AUC of 0.498 (p=0.966), indicating no significant learning curve. Real personal statements excelled over those by ChatGPT in aspects like originality, compelling nature, motivation for choosing orthopaedic surgery, personal interests, and career goals. Although AI can craft convincing statements that are sometimes indistinguishable from real ones, replicating personal nuances and individualistic elements found in real personal statements remains a challenge. Residency selection committees might want to prioritize these particular metrics while assessing personal statements, given the growing capabilities of AI in this arena.