Artificial Intelligence Generated Operative Reports for Common Spine Surgeries: How Close Are They to the Real Thing?

Rohan Iyer Suresh, Anthony K Chiu[1], John Carbone[2], Amit Ratanpal[3], Brian Shear[4], Alexander Ruditsky, Sennay Ghenbot, Jeffrey Weinreb, Stephen Daniel Lockey[5], Tyler J Pease, Idris Amin, Louis Bivona, Julio J Jauregui[6], Daniel Lee Cavanaugh, Eugene Young Koh, Steven C Ludwig

[1]The George Washington University School of Medicin, [2]University of Maryland Department of Orthopaedics, [3]Albany Medical College, [4]University of Maryland Medical Center, [5]University of Virginia, [6]University of Maryland / Shock Trauma

INTRODUCTION: Large language models (LLMs) are a form of generative artificial intelligence (AI) which have been rapidly popularized in recent years. These models are trained to understand and generate text in a way that is astoundingly similar to true human writing. In addition, LLMs can often demonstrate apparent knowledge of a topic through producing conceptually accurate information. The capability of AI to assist in increasing the efficiency of healthcare delivery is an area of major interest. One application of LLMs could be in the space of healthcare documentation, such as operative reports. The purpose of this paper is to investigate the level of similarity between AI generated operative reports and operative reports written by a fellowship-trained spine surgeon. We aim to assess whether or not participants are able to identify and distinguish operative reports that are generated from AI versus human input. Secondarily, we are interested in identifying qualities associated with AI-generated operative reports and understanding the level of surgeon interest in their utility.

METHODS: Operative reports for two common spine surgeries were generated on the popular LLM, Chat GPT-3 from OpenAI. The LLM was prompted to “write an operative report for” anterior cervical discectomy and fusion (ACDF) and lumbar microdiscectomy. A fellowship-trained spine surgeon then wrote operative reports for the same procedures. After obtaining informed consent, attending surgeons, fellows, and residents in orthopaedic surgery or neurosurgery were challenged to determine if a single randomized note was written by AI or a human, and differentiate between an AI-generated operative report and a human operative report. Finally, participants indicated their level of certainty and identified qualities of the operative reports. All data was obtained using REDCap software and analyzed using Microsoft Excel. Chi-squared tests of independence were conducted using R Software. The primary outcomes of interest were the ability of surgeons to 1) correctly identify a single random operative note as being AI-generated versus human-written, and 2) be able to distinguish an AI-generated operative note from a human-written note. Secondary outcomes were to evaluate surgeons’ level of certainty in their decision, to determine qualitative impressions of AI-generated operative notes, and to survey surgeon interest in AI technology for operative notes in the future. Qualitative impressions of interest included writing style, operative note components/structure, level of detail, inclusion of key procedural steps, and accuracy to real-world practice. Surgeons were asked if they thought AI could be implemented to assist with operative notes in the future and if they would personally be interested in using it.

RESULTS: A total of 52 respondents participated. All were orthopaedic surgeons or neurosurgeons. Overall, 69.2% of participants were able to identify an AI-generated operative report correctly for ACDF (P=0.050), and 61.5% for lumbar microdiscectomy (P=0.239). When comparing side-by-side, the rates of correct identification were 79.2% for ACDF (P=0.004) and 60% for lumbar microdiscectomy (P=0.317). Differentiation accuracy improved with level of training, from 55.6% at the resident level to 100% at the attending spine surgeon level. Most thought that AI had a human-like writing (86.3%), adequate detail (68.6%), key steps (78.4%), and accurate procedure description (68.6%). Finally, most (76.6%) thought AI could be implemented to assist with operative reports, and 87.5% would be interested in using AI for their reports.

DISCUSSION AND CONCLUSION: The ability to distinguish AI-generated operative reports depends on level of training. Overall, they possess many similarities with human-written notes and there is interest among spine surgeons in using AI in the future for procedure documentation. In fields like healthcare and scientific research, where the results of AI-generated information can directly impact the lives of patients, extreme caution should be taken in the utilization of such technology. As such, it is critical that clinicians are able to distinguish human-generated information from information which is produced via AI. This is becoming increasingly difficult as programs like ChatGPT continue to train within the world of medicine and only improve with time. This study highlights the importance of awareness to AI generated text in healthcare and indicates that caution should be taken particularly during the early stages of a physician’s career when they are more likely to mistake AI writing for that of humans.

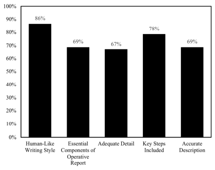

Figure 1. Qualities of Artificial Intelligence Generated Operative Notes

Table 1. Participant Characteristics and Response Accuracies

Respondent Characteristic	Number	Percent of Sample	Single Note Correct	Single Note Accuracy (%)	Comparison Correct	Comparison Accuracy (%)*
Total Respondents	52	(100%)	34	(65.4%)	34	(69.4%)
Level of Training						
Attending	19	(36.5%)	15	(78.9%)	17	(89.5%)
Fellow	4	(7.7%)	2	(50%)	2	(66.7%)
Resident	29	(55.8%)	17	(58.6%)	15	(55.6%)
PGY6	1	(1.9%)	0	(0%)	1	(100%)
PGY5	4	(7.7%)	2	(50%)	2	(50%)
PGY4	11	(21.2%)	8	(72.7%)	8	(72.7%)
PGY3	6	(11.5%)	3	(50%)	2	(33.3%)
PGY2	3	(5.8%)	2	(66.7%)	1	(50%)
PGY1	4	(7.7%)	2	(50%)	1	(33.3%)
Specialty						
Orthopaedics	50	(96.2%)	32	(64%)	32	(68.1%)
Neurosurgery	2	(3.8%)	2	(100%)	2	(100%)
Spine Surgeon	10	(19.2%)	9	(90%)	9	(100%)
Spine Surgery Rotation	52	(100%)				

*Percentages account for incomplete surveys; PGY, post-graduate year of training

Table 2. Looking at a Single Randomized Operative Report Alone

Operative Report Group	Respondents who thought the report was AI	Respondents who thought the report was Human	Total Respondents	Total % Correct	P-Value
ACDF					
Written by AI	9 (66% PRC)	2 (59% PRC)	26	9 (69.2%)	0.050
Written by Human	6 (60% PRC)	9 (63% PRC)			
Lumbar Microdiscectomy					
Written by AI	12 (55% PRC)	3 (50% PRC)	26	12 (61.5%)	0.239
Written by Human	7 (57% PRC)	4 (37% PRC)			

ACDF, anterior cervical discectomy and fusion

Table 3. Comparing Artificial Intelligence and Human Operative Reports Side-By-Side

Operative Report Group	Respondents who correctly distinguished reports	Respondents who incorrectly distinguished reports	Total Respondents	Total % Correct	P-Value
ACDF	19 (76% PRC)	5 (68% PRC)	24	19 (79.2%)	0.004
Expected (Null Hypothesis)	12	12			
Incomplete Survey Portion	2				
Participant-Reported Certainty					
Lumbar Microdiscectomy	15 (67% PRC)	10 (43% PRC)	25	15 (60%)	0.317
Expected (Null Hypothesis)	12.5	12.5	0	0	
Incomplete Survey Portion	1				
Participant-Reported Certainty					

ACDF, anterior cervical discectomy and fusion