Can artificial intelligence make elective hand clinic letters easier for patients to understand?

Abstract

We investigated whether ChatGPT was able to increase the Flesch reading ease and the Flesch–Kincaid reading level of elective clinic letters written by hand surgeons. ChatGPT could not reliably simplify the hand clinic letters any further.

Artificial intelligence (AI) has potentially enormous implications for healthcare as in other areas of life. Open AI's ChatGPT chatbot (https://chat.openai.com/, accessed November 2023) was launched in November 2022 and uses natural language processing (NLP) technology to generate human-like text. One area in which this new technology may help clinicians would be in preparing or improving their clinic letters. Although this is a very important way of communicating to patients, there are concerns that information may be presented in such a way that makes it difficult for the recipient to understand (e.g. the use of complex technical language, jargon, etc.) (Drury et al., 2021). The aim of the present study was to analyse clinic letters written by orthopaedic hand specialists in terms of their length and readability using the Flesch reading ease and the Flesch–Kincaid (FK) grade level scores (Flesch, 1948; Kincaid et al., 1975). The Flesch reading ease score is in the range of 0–100, with a higher score being easier to read. The FK reading level score has a range of 0–20, with a higher value being harder to read. The project was registered with the institutions' audit departments and was conducted in accordance with the Declaration of Helsinki and the guidelines for good clinical practice.

This prospective cohort study was conducted in two district general hospitals in the UK. Clinic letters dictated after new patient consultations in elective hand clinics were identified from electronic patient record systems. With the consent of the authors, the clinic letters were then selected for analysis. Letters were chosen sequentially over a 4-week period in November 2023. The grade of the assessing clinician was recorded along with the word count, Flesch reading ease and FK grade level. Any identifiable patient or clinician information was removed before being entered into ChatGPT version 3.5. Three stems were piloted using 10 clinic letters and the most promising was selected:

> *'Here is a letter written by a hospital specialist. Can you rewrite the letter including all the salient medical information but making it easier for a patient to understand?'*

The word count, Flesch reading ease and the FK grade level were then reassessed. Two authors (AS and LW) independently reviewed the content of the letters to check that they still contained all the medically relevant information.

Power calculations determined that 64 letters would have to be analysed to show a 10% change in readability score at the 5% probability level. The Student *t*-test was used to compare linear variables between groups after the Shapiro–Wilk test and Q-Q plots confirmed that the data were normally distributed. A *p*-value <0.05 was defined as being statistically significant.

A total of 70 new patient clinic letters were analysed (written by consultant, 27; fellow, 24; registrar, 12; other, 7). The human-authored letters had a mean word count of 250 (SD 95), a mean Flesch reading ease of 57.1 (SD 8.2) and a mean FK reading level of 9.2 (SD2.3). After ChatGPT editing, the mean word count increased to 272 ($p = 0.003$), Flesch

reading ease reduced to 42 ($p < 0.001$) and FK reading level increased to 11.3 ($p < 0.001$). Using the upgraded 4.0 version of ChatGPT did result in significantly easier Flesch reading ease and FK reading level than version 3.5 ($p < 0.001$ for both). However, there was no significance difference between Flesch reading ease and FK reading level between ChatGPT v4.0 and humans ($p = 0.348$ and 0.078, respectively). All letters were reviewed independently by the two lead authors (AS and LW), who found that the content remained unchanged with no significant deviations from the original text.

Many medical professionals are guilty of slipping into a particular pattern of rehearsed language or shorthand, such as 'please do not hesitate to get in touch', which ChatGPT translated to 'please reach out'. Overall, ChatGPT was good at interpreting jargon and replacing it with more everyday expressions, for example replacing 'neurovascularly intact' with 'nerves and blood vessels working'. ChatGPT was able to detect and correct simple typographic faults or transcription errors (duplicate words, misspellings etc.); however more technical errors were transcribed directly. For example, one letter referred to the 'plantar plate' instead of the 'palmar plate' owing to a mistranscription of the original dictation, and this was not picked up.

Surgeons are often tactful in letters, for example when mentioning patients' physical characteristics (e.g. weight); however, on occasion it was felt that the ChatGPT version of the letter lacked this tact. Furthermore, on several occasions the AI model inserted sentences that were not present in the original letter, such as 'please know that we are here to support you every step of the way'. Anecdotally ChatGPT v.4.0 was 'worse' about doing this. It also unnecessarily embellished certain sentences or inserted a flamboyant sign off to the patient's general practitioner, such as 'your expertise and insights are invaluable' (which, although this may be true, was not expressed in the original letter).

Based on the results of this study, ChatGPT could not reliably make clinic letters easier for patients to understand. Furthermore, although it made some useful editorial changes and converted medical jargon into plain English, it could not be considered reliable about issues of tact or the unnecessary embellishment of phrases. We tried giving different question stems to ChatGPT. Although it was possible to simplify letters further by re-wording the stem, this came at the expense of sacrificing clinically important information.

Although we found no benefit from using ChatGPT in the writing of clinical letters overall, the most junior clinicians (core surgical trainee level) produced the three most complex clinic letters and Chat GPT was able to reduce this complexity, which may indicate a potential role in teaching trainees how to construct simpler clinic letters. We would not currently recommend the routine use of AI to reformat clinic letters but there may be certain situations, for example in patients who have a low reading age, where it might improve comprehension. However, it would be necessary to check such a letter and so no time would be saved.

ORCID iD Lucy C. Walker https://orcid.org/0000-0003-0354-2342

References

Drury DJ, Kaur A, Dobbs T, Whitaker IS. The readability of outpatient plastic surgery clinic letters: are we adhering to plain English writing standards? Plast Surg Nurs. 2021, 41: 27–33.

Flesch R. A new readability yardstick. J Appl Psychol. 1948, 32: 221–3.

Kincaid JP, Fishburne RP, Rogers RL, Chissom BS. Derivation of new readability formulas for Navy enlisted personnel. Research branch report. Memphis, TN: Naval Air Station. 1975, 8–75.

Adam C. Stoneham[1], Lucy C. Walker[2],* , Michael J. Newman[1], Alex Nicholls[2] and Duncan Avis[2]
[1]Department of Trauma and Orthopaedics, University Hospitals Southampton, Southampton, UK
[2]Department of Trauma and Orthopaedics, Hampshire Hospitals Foundation Trust, Basingstoke and North Hampshire Hospital, Basingstoke, UK
*Corresponding author: Lcwalker86@gmail.com