Using ChatGPT to write patient clinic letters

Published Online
March 7, 2023
https://doi.org/10.1016/S2589-7500(23)00048-1

The appropriate recording and communication of clinical information between clinicians and patients are of paramount importance. Recently there has been a much-needed drive to improve the information that is shared with patients.[1] However, the preparation of clinical letters can be time consuming. Although there has been an increase in the use of letter templates and voice recognition systems, with the aim of improving efficiency, novel technologies such as natural language processing (NLP) and artificial intelligence (AI) have the power to revolutionise this area of practice. NLP algorithms are designed to recognise and understand the structure and meaning of human language, classify texts according to their content or purpose, and generate responses that are appropriate and coherent.[2] OpenAI's ChatGPT chatbot was launched in November, 2022, and uses NLP technology to generate human-like text. The generative pre-trained transformer (GPT) language model is based on a transformer architecture, which allows it to process large amounts of text data and generate coherent text outputs by learning the relationships between input and output sequences. The GPT language model has been trained on large datasets of human language, with several studies demonstrating that it is very good at generating high-quality and coherent text outputs.[3–5] As a result, AI, like ChatGPT, has the potential to produce high quality clinical letters that are comprehendible by patients while improving efficiency, consistency, accuracy, patient satisfaction, and deliver cost savings to a health-care system. In this Comment we describe the early adoption and evaluation of ChatGPT-generated clinical letters to patients with limited clinical input. The aim was to evaluate the readability, factual correctness, and humanness of ChatGPT-generated clinical letters to patients, using the example of skin cancer as the most common human cancer.

We created a series of different clinical communication scenarios that covered the remit of a clinicians' skin cancer practice. To simulate how clinicians might use ChatGPT in the clinical environment we created short-hand instructions to input into the chatbot, which we defined as limited clinical input, because the input is small compared with the relative amount of natural free text a clinician would otherwise be required to write or dictate to generate a clinical letter. In the USA, it is recommended that patient-facing health literature be written at or below a sixth grade level (age 11–12 years).[6] There are no specific guidelines on this in the UK. In view of this, all letters were instructed to be written at a reading age of 11–12 years. We sought to evaluate the capabilities of ChatGPT by presenting it with a series of instructions of increasing complexity (appendix p 1). These instructions ranged from simply following specific directions, to using national guidelines and data from these guidelines to provide clinical advice in the letter—eg, the management of anticoagulation peri-operatively. After submitting the instructions, ChatGPT then generated a response in the form of a clinical letter to be issued to the patient. The online tool readable was used to evaluate the readability of letters with commonly used formulae as described in many other studies.[7,8] Factual correctness and humanness of letters were assessed by two independent clinicians using a Likert scale ranging from 0 to 10, with 0 representing completely incorrect or inhuman and 10 representing completely correct and human. Error analysis was performed using linear regression. We used two separate generalised linear models (GLMs) to investigate the effect of the predictor variables of cancer type (basal cell carcinoma [BCC] as reference category), general commands, specific guidelines, and general guidelines on the outcome variables, and median humanness in the first GLM and median correctness in the second GLM. Statistical analysis was done in R (version 4.1.1) p<0·001 was deemed statistically significant.

See Online for appendix

For more on **readable** see https://readable.com

For more on **ChatGPT** see https://openai.com/blog/chatgpt/

38 hypothetical clinical scenarios were created, seven of which pertained to BCC, 11 to squamous cell carcinoma (SCC), and 20 to malignant melanoma (MM; appendix p 2). Overall, the readability scores suggest that the text might be suitable for a varying reading ability, and the mean readability age for the generated letters was at a USA ninth grade (aged 14–15 years) and considered by the US Department of Health and Human Services as average difficulty (appendix p 2). Overall median correctness of the clinical information contained in the letter was 7 (range 1–9). Overall median humanness of the writing style was 7 (5–9; appendix pp 3–4). The weighted κ for correctness was 0·80 (p<0·0001) and humanness was 0·77 (p<0·0001).

For median correctness, ANOVA howed a statistically significant difference among the groups ($F_{2,35}$=10·1, p=0·00035). The Tukey honestly significant difference (HSD) test showed that the mean difference between the MM and BCC groups was statistically significant (–2·71 [95% CI –4·32 to –1·11], p<0·0001). There was no statistically significant difference between the SCC and BCC groups (p=0·31) or the SCC and MM groups (p=0·016). For median humanness, ANOVA showed a statistically significant difference among the groups ($F_{2,35}$=27·76, p<0·0001). The Tukey HSD test showed that there was a statistically significant difference between the MM and BCC groups (–1·63 [–2·17 to –1·09], p<0·0001) and between the SCC and BCC groups (–1·43 [–2·03 to –0·83], p<0·0001). There was no statistically significant difference between the SCC and MM groups (p=0·55).

Results of the GLM for median humanness showed that cancer type was a significant predictor, with the MM coefficient being –1·45 (SE 0·30) and the SCC coefficient being –1·32 (0·28; both p<0·0001). The general commands, specific guidelines, and general guidelines variables were not significant predictors of median humanness. The multiple R^2 value of 0·622 indicated that the model explained 62·23% of the variance in median humanness, and the F-statistic of 10·55 and corresponding p<0·0001 indicated that the model was significant overall. In the GLM for median correctness, MM was found to be significantly associated with median correctness with a coefficient of –2·64 (SE 0·90; p<0·0001). The other predictors were not significantly associated with median correctness. The multiple R^2 value of 0·3729 suggested that the model explained approximately 37·29% of the variance in median correctness. The F-statistic and corresponding p value were used to test the overall significance of the model, and p<0·0001 indicated that the model is significantly different from a model with no predictors.

This pilot assessment shows that it is possible to generate clinic letters with a high overall correctness and humanness score with ChatGPT. Furthermore, these letters were written at a reading level that is broadly similar to current real-world human generated letters.[9] The ability of AI to generate clinical letters as an alternative to those written by clinicians raises important considerations for the quality and effectiveness of health-care communication. It is important that potential risks, such as omissions or errors, which might have serious consequences for patient care, are mitigated. The incorrect reporting of results or interpretation of treatment guidelines could affect patient morbidity and mortality. To mitigate these risks, it is important for the use of AI in health care, including the automated generation of clinical letters, be carefully regulated and monitored. In the early stages of adoption of such new technologies it is necessary to continue with a human-in-the-loop approach, whereby the outputs of such systems are carefully verified by health-care providers. To responsibly incorporate ChatGPT or similar generic AI systems into the clinical workflow, one approach could be to use voice-to-text recognition software with limited human input, followed by rapid clinician editing of the generated letter. This approach could be a feasible starting point for exploring the potential applications of this technology while also addressing any potential risks. Further studies are needed to assess the effectiveness of AI-generated clinical letters in real-world clinical settings and to compare the performance of various AI systems against one another. It is probable that an AI system with a greater medical focus will yield improved results. Additionally, the quality and writing style of input provided to any chatbot should be assessed in a range of settings, including different languages, resource levels, and cultural contexts.

Caution must be exercised and potential risks must be proactively addressed to ensure the safety and quality of patient care while introducing such an important technical advancement.

We declare no competing interests. SRA and TDD are funded by the Welsh Clinical Academic Training Fellowship. SRA received a Paton Masser grant from the British Association of Plastic, Reconstructive and Aesthetic Surgeons to support this work. ISW is the surgical Specialty Lead for Health and Care Research Wales; the Chief Investigator for the Scar Free Foundation & Health and Care Research Wales Programme of Reconstructive and Regenerative Surgery Research; an associate editor for the Annals of Plastic Surgery; an editorial board member of BMC Medicine and numerous other editorial board roles. Ethical committee approval has been obtained from Swansea University Medical School Research Ethics Subcommittee (2020-0025). The study was performed in accordance with the Declaration of Helsinki.

*Stephen R Ali, Thomas D Dobbs, Hayley A Hutchings, Iain S Whitaker
Stephen.ali@wales.nhs.uk

Reconstructive Surgery and Regenerative Medicine Research Centre, Institute of Life Sciences, Swansea University Medical School, Swansea SA2 8PP, UK (SRA, TDD, ISW); Welsh Centre for Burns and Plastic Surgery, Morriston Hospital, Swansea, UK (SRA, TDD, ISW); Swansea University Medical School, Faculty of Medicine, Health and Life Science, Swansea, UK (HAH)

1 General Medical council. Domain 3: communication partnership and teamwork https://www.gmc-uk.org/ethical-guidance/ethical-guidance-for-doctors/good-medical-practice/domain-3---communication-partnership-and-teamwork#paragraph-31 (accessed Dec 22, 2022).

2 Jurafsky D, Martin JH. Speech and language processing, 3rd edn. Prentice Hall, 2009.

3 Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *arXiv* 2020; published online July 22. https://doi.org/10.48550/arXiv.2005.14165 (preprint).

4 Radford A, Wu J, Child R, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv* 2019; published online July 28. https://doi.org/10.48550/arXiv.1910.10683 (preprint).

5 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. December 2017 (abstr 5998–6008).

6 US Department of Health & Human Services. National action plan to improve health literacy. 2010. https://health.gov/sites/default/files/2019-09/Health_Literacy_Action_Plan.pdf (accessed Dec 22, 2022).

7 Burke V, Greenberg D. Determining readability: how to select and apply easy-to-use readability formulas to assess the difficulty of adult literacy materials. *Adult Basic Educ Lit* 2010; **4:** 34–42.

8 Wang L-W, Miller MJ, Schmitt MR, Wen FK. Assessing readability formula differences with written health information materials: application, results, and recommendations. *Res Social Adm Pharm* 2013; **9:** 503–16.

9 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.