

Generative AI use in writing tests: a case study of a MedComms agency's first suspected submissions

Bernard Kerr and Valerie Moss

Prime, London, UK

Conclusions

- These cases highlight the value of human insight and understanding in the preparation of academic publications, as well as the importance of reviewing and fact checking AI-generated text.
- Tools to detect artificial intelligence (AI)-generated text do not appear sufficiently reliable for real-world use, given the potentially serious ramifications of incorrect results for both candidates and agencies.
 - Available data indicate that the current generation of such tools have non-negligible rates of false-negative classification of AI-generated text and false-positive classification of human-written text.¹
- The various issues identified in these writing tests highlight that it is possible for a human reviewer to identify AI-generated text in some scenarios, based on subjective quality and the reviewer's experience.
 - In particular, the combination of features we have described led to suspicion that the text was potentially AI-generated.
 - Identification of factitious citations in the bibliography was considered to provide confirmation that generative AI was used.
- Identification of multiple hallmark characteristics may suggest undisclosed use of generative AI and should prompt caution when assessing writing tests.
 - With the exception of factitious statements and citations, none of the hallmarks presented here can be considered diagnostic of AI-generated text.
 - Due care should be taken when drawing conclusions around the provenance of content.
- We have updated our writing test instructions to ask candidates not to use generative AI and alerted all writing test reviewers to be vigilant for potential AI-generated text.
- The implications of the identified hallmarks in other AI use cases may warrant further investigation.

Introduction

- Medical communications agencies employ writing tests during the recruitment process to assess medical writer candidates' proficiency.
 - Candidates are typically asked to prepare a sample of academic writing based on a standardised brief and source materials.
 - The purpose of the test is to confirm that candidates have the knowledge, skills, and experience to succeed in a medical writer role and complete work to a high standard.
- Use of generative AI to complete a writing test, either through modification of human-written text or by generating content *de novo*, may result in a misleading assessment of a candidate if undetected.
- Here, we present a case study of the first writing tests received by our agency that were suspected to have been completed using generative AI, and outline the characteristics that led to our suspicions.

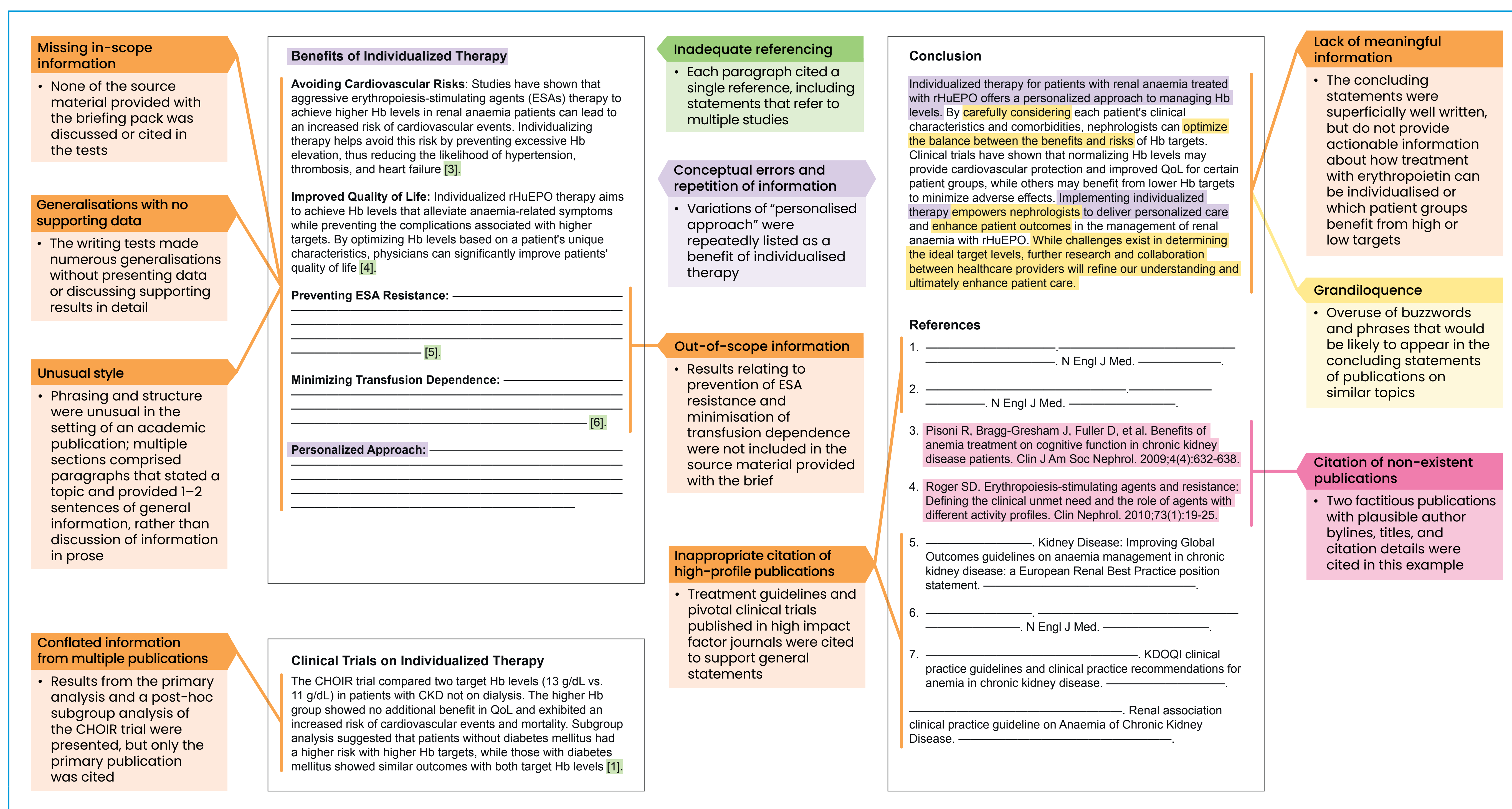
Methods

- Candidates were provided with source material and asked to prepare a 1000-word mini-review article on the benefits of individualising recombinant erythropoietin therapy in the management of renal anaemia.
- The source materials included a summary of data from four published articles, as well as a review article on a generally similar topic.
 - All source materials were published well before the cut-off date for the freely available ChatGPT 3.5 large language model, and so are presumed to be included in its training dataset.²
- Blinded review of tests was performed by experienced medical writers using a standardised scoring system to evaluate candidates' competency in terms of writing style, evaluation of the subject area, and technical accuracy.

Case report findings

- On initial review, the tests appeared to be written in a clear style, with an overall structure typical of a review article.
 - The tests did not have any notable spelling, grammar, or punctuation errors other than inconsistent use of UK and US English.
- The introduction sections were insightful, setting the scene and communicating key points that would typically be seen in tests likely to attain a high overall score.
- On further assessment, a variety of quality issues and unusual features were identified, leading to suspicion of undeclared use of generative AI (Figure 1).
 - The writing style varied across different parts of the tests, with some sections written in a style typical of academic writing, some in a "commercial" style unsuited to an academic publication, and some comprising multiple short paragraphs reminiscent of a bulleted list.
 - The tests relied heavily on non-specific general statements.
 - The concluding statements were repetitive, did not provide meaningful information and contained an unusually high number of "buzzwords".
 - The bibliographies were weighted towards publications of pivotal clinical trials from high impact factor journals and treatment guidelines, and some references listed in the bibliography sections were not cited in the text.
 - None of the source material provided with the brief was presented or cited.
 - The bibliographies included plausible but non-existent references that had author bylines and titles similar to real publications in the literature, consistent with reports of "AI hallucination".
- Use of generative AI was not disclosed by the candidates.

Figure 1: Proposed hallmarks of AI-generated text and example content from submitted writing tests



References

- Elkhatat AM, et al. Int J Educ Integr. 2023;19:17.
- OpenAI. Models. GPT-3.5. 2023. Available at: <https://platform.openai.com/docs/models/gpt-3-5>. Accessed December 18 2023.

Disclosures

Bernard Kerr and Valerie Moss are employees of Prime, London, UK. We disclose that AI was used to generate a plain language summary of this poster.

Acknowledgements

We would like to thank the Production and Editor teams for their support with developing this poster.

