J Korean Med Sci. 2023 Aug 7;38(31):e240
https://doi.org/10.3346/jkms.2023.38.e240
eISSN 1598-6357·pISSN 1011-8934

JKMS

Opinion
Editing, Writing & Publishing

The Acceptable Text Similarity Level in Manuscripts Submitted to Scientific Journals

Farrokh Habibzadeh

Past President, *World Association of Medical Editors* (WAME)
Editorial Consultant, *The Lancet*
Associate Editor, *Frontiers in Epidemiology*

Received: Jun 3, 2023
Accepted: Jun 16, 2023
Published online: Jul 13, 2023

Address for Correspondence:
Farrokh Habibzadeh, MD
R&D Center, Petroleum Industry Health Organization Polyclinic, Eram Blvd, Shiraz, 7143837877, Iran.
Email: Farrokh.Habibzadeh@gmail.com

ORCID iD
Farrokh Habibzadeh
https://orcid.org/0000-0001-5360-2900

Disclosure
The author has no potential conflicts of interest to disclose.

ABSTRACT

Plagiarism is among commonly identified scientific misconducts in submitted manuscripts. Some journals routinely check the level of text similarity in the submitted manuscripts at the time of submission and reject the submission on the fly if the text similarity score exceeds a set cut-off value (*e.g.*, 20%). Herein, I present a manuscript with 32% text similarity, yet without any instances of text plagiarism. This underlines the fact that text similarity is not necessarily tantamount to text plagiarism. Every instance of text similarity should be examined with scrutiny by a trained person in the editorial office. A high text similarity score does not always imply plagiarism; a low score, on the other hand, does not guarantee absence of plagiarism. There is no cut-off for text similarity to imply text plagiarism.

Keywords: Plagiarism; Verbatim; Journalism; Text Similarity; Publication Ethics; Scientific Writing

Introduction

The US Office of Research Integrity defines plagiarism as "the theft or misappropriation of intellectual property and the substantial unattributed textual copying of another's work."[1] Plagiarism can be divided into two broad categories—plagiarism of idea and plagiarism of text, also called "verbatim." Tantamount to "theft," plagiarism of idea is not acceptable at all.[2] However, depending on the study filed, degrees of verbatim may be tolerable. For instance, while plagiarism of text in literature (where the originality of a work essentially depends on its wordings) is not acceptable by any means, degrees of verbatim in scientific writing (where wordings are just a means to convey the scientific idea and not the foundation of the originality and scientific merits of the work) might be tolerable, as long as we can be sure that the authors could correctly construe the text piece they borrowed.[2,3]

Plagiarism has long been an important concern for science journal editors as it is among research misconducts commonly identified in manuscripts submitted to scientific journals.[4] And, that is why some journals have chosen to use plagiarism detection software programs such as iThenticate® to determine the extent of text similarity in the submitted manuscripts; some use it at the time of submission and reject the submission on the fly if the text

similarity exceeds a certain level set by the journal. Some scientific journals have chosen a stringent cut-off of 10–15%; others decided to be laxer and tolerate a text similarity of up to 25%.[5-8] Choosing the strategy of rejecting a submission merely based on a text similarity score reported by a software inclusively reflects that the journal officials consider every "text similarity" an instance of plagiarism, and that is exactly the root of a major problem in science publishing. Herein, I wish to present a piece of my own work which could not pass the text similarity test and has thus been rejected outright by a journal. I would like to emphasize that text similarity does not always mean plagiarism of text.

Case Study

After the COVID-19 pandemic, many researchers studied and published many articles on the disease. Our team also worked on this issue. In one of our original works, we drafted a manuscript consisting of 194 words in its title, abstract and keywords; 1,122 words in the body; 386 words in 12 references; and 47 words in the figure legend of a graph. The manuscript structure followed the IMRaD format.[9] It was submitted to a scientific journal. After a while, the journal rejected the manuscript without sending it for external peer review because of 32% text similarity, as reported by iThenticate® (see **Supplementary Data 1**). The manuscript's first page (containing the authors' names and affiliations) and references had not been checked. The 32% text similarity occurred in the title, abstract, keywords, body of the manuscript, and figure legend. Examining the results revealed that the manuscript had text in common with 29 articles; the highest percentage of text similarity (8%, 119 non-consecutive words) came from an article authored by my son and his colleagues on the same topic,[10] followed by another article we had published on seropositivity of patients who recovered from COVID-19 (3%, 39 non-consecutive words).[11] The extent of text similarity with the remaining 27 articles was 1% or less (see **Supplementary Data 1**).

Discussion

One of our manuscripts was rejected by a journal because almost one-third of its text was similar to previously published articles; yet, I believe no plagiarism of text (verbatim) has happened. The parts of text reported to be similar between the examined manuscript and the first two references with the highest text similarity score,[10,11] were non-specific, phrases such as "*potassium and magnesium levels above 4 mEq/L and 3 mg/dL*" or "*medications with a wide range of side effects.*" The references and the first page of the manuscript had correctly not been examined for text similarity; otherwise, the similarity score would have been higher. As pointed out by other researchers, text similarity does not necessarily mean plagiarism of text.[12-14] I do believe that it is naïve to just rely on a simple score to determine if an author has tried to intentionally deceive us and label a submitted manuscript plagiarized. As a principal of good editorial practice, I believe, it is a moral duty of the editorial staff to evaluate the suspicious similar text with scrutiny to determine if the similar text is really an instance of verbatim or not. The example presented here indicates that a high text similarity score, even as much as 32%, is not evidence of plagiarism. Who knows? One may find a manuscript with a higher text similarity score, but without real plagiarism. On the other hand, manuscripts with a trivial text similarity score may be found as a case of plagiarism if they are examined with enough scrutiny. Sometimes, similarity of a single sentence or phrase is enough to designate text plagiarism. There is no cut-off for text similarity to imply text plagiarism.

https://doi.org/10.3346/jkms.2023.38.e240

Given the current situation we have with chatbots and generative artificial intelligence (AI) and their role in scholarly publications and the risk of inadvertent plagiarism caused by the algorithms involving in the generation of the output of such AI-based units, editors should be aware of instances of verbatim more than ever.[15-18] Recently, the *World Association of Medical Editors* (WAME, https://www.wame.org/) and the *International Committee of Medical Journal Editors* (ICMJE, https://www.icmje.org/) have published their statements on use of AI-based technologies in writing scholarly manuscripts.[15,18] They suggest authors to exclusively mention use of AI-based technologies, such as ChatGPT,[19] in conducting their study and in writing their submitted manuscripts, and advise editors to consider this issue seriously as it might cause serious problems in terms of plagiarism. Both statements clearly mention that the responsibility of the text coming in the submitted manuscript should be solely shouldered by the [human] authors, not machines. WAME also recommends that editors should be equipped with necessary means to identify text generated by AI-based technologies.[15] A reasonable approach to identify AI-generated text would be use of AI-based technologies. GPTZero (https://gptzero.me/) is an example of such tools. Nonetheless, its performance in identifying AI-generated text is still dubious.[20] All these underline the need for training of editorial staff regarding the definition of plagiarism and the way to identify whether text similarity should be interpreted as plagiarism or not.

Although identification of verbatim in submitted manuscripts might look imperative at the first glance, I believe this is not that important in scientific writing. The essence of science is not in the wordings of a manuscript. The eloquence of the text of a submitted manuscript does not affect the scientific merits of a research article. Given the limitation in the number of scientific terms, there are only a few acceptable ways to satisfactorily express an idea in a manuscript.[2,3] Furthermore, I believe the contribution of AI-based units will substantially increase in research and scientific writing. There will be a time when a Universal AI (UniAI) will take over and researchers will no longer need to write their own manuscripts; they just need to fill in the blank fields in a template;[3,21] the UniAI will take care of the rest. At the heart of any misconduct is the "intention to deceive others." As the UniAI is going to write all the articles and it will have no reason to deceive others, any instance of text similarity, if exists, could be considered an instance of text similarity rather than plagiarism.

The analogy in the syntax and semantics rules between the human and computer programming languages makes it possible to better explore the situation. Human languages have a large pool of words; computer languages have few keywords. As an example, there are only 32 keywords (reserved words) in *C* programming language (just an example of a computer programming language). The limitation in the number of keywords in a language increases the likelihood that program developers write very similar pieces of codes. Sometimes, this similarity originates from a historical event; almost all program developers across the globe, regardless of the computer language they employ, use the letters '*i*' and '*j*' to designate their loop counter variables. Most of them do not realize that this habit dates back to old days when the FORTRAN programming language used by default the variables beginning with '*i*' and '*j*' to store integers. Sharing parts of the codes in the above example does not imply plagiarism in computer sciences. Human languages (*e.g.*, English) share similar linguistic properties, only with a larger pool of keywords. Likewise, use of similar pieces of text, all generated by UniAI,[21] will no longer be considered text plagiarism, as there will be no intention to deceive others, I believe.[3,21]

https://doi.org/10.3346/jkms.2023.38.e240

SUPPLEMENTARY MATERIAL

Supplementary Data 1
iThenticate report

Click here to view

REFERENCES

1. US Office of Research Integrity. ORI policy on plagiarism. https://ori.hhs.gov/ori-policy-plagiarism. Accessed May 30, 2023.
2. Vessal K, Habibzadeh F. Rules of the game of scientific writing: fair play and plagiarism. *Lancet* 2007;369(9562):641.
PUBMED | CROSSREF
3. Habibzadeh F. Plagiarism: What does the future hold for science writing? *Eur Sci Ed* 2014;40(4):91-3.
4. Gupta L, Tariq J, Yessirkepov M, Zimba O, Misra DP, Agarwal V, et al. Plagiarism in non-Anglophone countries: a cross-sectional survey of researchers and journal editors. *J Korean Med Sci* 2021;36(39):e247.
PUBMED | CROSSREF
5. Memon AR. Similarity and plagiarism in scholarly journal submissions: bringing clarity to the concept for authors, reviewers and editors. *J Korean Med Sci* 2020;35(27):e217.
PUBMED | CROSSREF
6. Memon AR, Mavrinac M. Knowledge, attitudes, and practices of plagiarism as reported by participants completing the authorAID MOOC on research writing. *Sci Eng Ethics* 2020;26(2):1067-88.
PUBMED | CROSSREF
7. Mahian O, Treutwein M, Estellé P, Wongwises S, Wen D, Lorenzini G, et al. Measurement of similarity in academic contexts. *Publications* 2017;5(3):18.
CROSSREF
8. Peh WC, Arokiasamy J. Plagiarism: a joint statement from the Singapore Medical Journal and the Medical Journal of Malaysia. *Med J Malaysia* 2008;63(5):354-5.
PUBMED
9. Sollaci LB, Pereira MG. The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *J Med Libr Assoc* 2004;92(3):364-7.
PUBMED
10. Habibzadeh P, Moghadami M, Lankarani KB. The effect of potential therapeutic agents on QT interval in patients with COVID-19 Infection: The importance of close monitoring and correction of electrolytes. *Med Hypotheses* 2020;143:109847.
PUBMED | CROSSREF
11. Habibzadeh P, Sajadi MM, Emami A, Karimi MH, Yadollahie M, Kucheki M, et al. Rate of re-positive RT-PCR test among patients recovered from COVID-19. *Biochem Med (Zagreb)* 2020;30(3):030401.
PUBMED | CROSSREF
12. Manley S. The use of text-matching software's similarity scores. *Account Res* 2023;30(4):219-45.
PUBMED | CROSSREF
13. Li Y. Text-based plagiarism in scientific publishing: issues, developments and education. *Sci Eng Ethics* 2013;19(3):1241-54.
PUBMED | CROSSREF
14. Shashok K. Plagiarism: Intention and diagnostic criteria. *Saudi J Anaesth* 2012;6(2):188.
PUBMED | CROSSREF
15. Zielinski C, Winker MA, Aggarwal R, Ferris LE, Heinemann M, Lapeña JF, et al. Chatbots, Generative AI, and Scholarly Manuscripts. WAME Recommendations on Chatbots and Generative Artificial Intelligence in Relation to Scholarly Publications. https://wame.org/page3.php?id=106. Updated 2023. Accessed June 2, 2023.
16. Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? *Crit Care* 2023;27(1):75.
PUBMED | CROSSREF
17. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)* 2023;11(6):887.
PUBMED | CROSSREF

18. International Committee of Medical Journal Editors (ICMJE). Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals. https://www.icmje.org/icmje-recommendations.pdf. Updated 2023. Accessed June 12, 2023.
19. Doskaliuk B, Zimba O. Beyond the keyboard: academic writing in the era of ChatGPT. *J Korean Med Sci* 2023;38(26):e207.
PUBMED | CROSSREF
20. Heumann M, Kraschewski T, Breitner MH. ChatGPT and GPTZero in research and social media: a sentiment- and topic-based analysis. https://ssrn.com/abstract=4467646. Updated 2023. Accessed June 12, 2023.
21. Habibzadeh F. The future of scientific journals: the rise of UniAI. *Learn Publ* 2023;36(2):326-30.
CROSSREF