

Detecting and handling suspected plagiarism in submitted manuscripts

Durga Prasanna Misra,¹ Vinod Ravindran²

Keywords: Similarity; duplicate; software; artificial intelligence; machine learning; editorial process; peer review; retraction

Financial and Competing Interests: DPM is an Associate Editor of the Journal of the Royal College of Physicians of Edinburgh, and serves as editor/editorial board member/reviewer for several other international journals. VR is the Editor in Chief of the Journal of the Royal College of Physicians of Edinburgh, and serves as editor/editorial board member/reviewer for several other international journals. This paper has undergone peer review in accordance with JRCPE's policies.

Correspondence to:

Vinod Ravindran
Centre for Rheumatology
Calicut, Kerala
India

Email:

drvinod12@gmail.com

Plagiarism is a great masquerade. Detecting plagiarism is daunting even for experienced reviewers and editors.^{1,2} Plagiarism constitutes scientific misconduct by violating the intellectual property rights of the creators of such content.³ Literature suggests that up to a sixth of manuscripts submitted to journals might be affected by plagiarism.⁴ Through this editorial, we share our perspectives on how to identify and manage instances of suspected plagiarism in manuscripts submitted to journals for publication.

Plagiarism – quite different from mere similarity!

It is imperative for all stakeholders (i.e. editors, reviewers, and authors) to understand that plagiarism is distinct from similarity.⁵ Plagiarism essentially refers to the reproduction of information without duly attributing the source of such information, i.e. passing it off as one's own. This information could be text (as in manuscripts submitted to journals), previously published figures and tables in manuscripts, or ideas/hypotheses for research. On the other hand, similarity of text simply refers to how similar or dissimilar the text in the manuscript in question is, when compared with the available literature, either in previous manuscripts, web pages or grey literature.⁵ Software like iThenticate and Turnitin are commonly used to identify the extent of similarity in submitted manuscripts by journals.⁵ Furthermore, the extent of similarity detected by such software depends upon filters employed by the user, such as the number of words that are required to be similar in succession, or whether reference lists are included. The latter inevitably makes a manuscript appear more similar than it actually is, since references are generally listed in databases or in previous manuscripts. Generally, a limit of 8-10 words in succession is considered as an indicator of significant degree of similarity. However, even this might have limitations. There might be long phrases such

as names of certain research tools or organisations similar with the published literature, however this does not amount to plagiarism.⁵ Organisations across the world responsible for regulating teaching and research, such as the University Grants Commission (UGC) in India, have set out limits for similarity in different parts of theses and manuscripts, dividing this further into areas like methods where similarity to some extent might be tolerable, and results and discussion where even minimal similarity might not be acceptable. Such guidelines appear to imply that thresholds of similarity are synonymous with plagiarism.⁶

However, plagiarism is not as simple as reproducing words. First, the mere reproduction of words need not indicate plagiarism at all, if this is done appropriately by placing the reproduced text within double quotes and referencing the source, however, such a practice should be minimised while writing manuscripts. Second, authors in the present world are aware of the ability of similarity checking software to identify such reproduced text, hence, might be able to partially bypass similarity detection by changing a few words here and there from the reproduced source.⁵ Third, such similarity checking software are unable to detect plagiarism of figures and tables. Fourth, the plagiarism of ideas or research hypotheses cannot be detected by checking similarity. Instead, this might be identifiable by cross-checking lists of references and their order in the source manuscript compared with the present manuscript, or based on evidence provided to the journal by the scientist who has claimed priority over the given idea. It is evident that the identification of most of these forms of plagiarism requires considerable manual input from editors and reviewers in the present day. Therefore, detecting plagiarism should not be simply thought of as an output of similarity checking software, which can only serve as a tool to support editors and reviewers in identifying instances of plagiarism. Ultimately, the identification of plagiarism (whether

¹Department of Clinical Immunology, Sanjay Gandhi Postgraduate Institute of Medical Sciences (SGPGIMS), Lucknow-226014, India;

²Centre for Rheumatology, Calicut, Kerala, India

Table 1 Plagiarism and similarity – comparison and contrast

	Plagiarism	Similarity
Theme	Duplication of content without appropriate attribution of its source	Similarity of a manuscript to other published literature, web pages or grey literature
Ease of identification	Difficult, particularly for plagiarised figures, tables and ideas	Relatively straightforward and automated. Output is generated by similarity checking software
Gold standard	Source from which plagiarism has occurred	Varies depending on the breadth of coverage of the similarity checking software
Judgement	Subjective based on editors' and reviewers' assessment of plagiarism	Objective as similarity is provided by the similarity checking software as percentage
Threshold for action	Based on the editorial policies driven by peer reviewers' and editors' oversight	Arbitrary, based on cut-offs for similarity decided by a journal which does not necessarily reflect the presence/absence of plagiarism or its extent
Possible actions	<p>If judged to be minor, content might be corrected by authors before publication, or subject to erratum or partial retraction after publication</p> <p>If judged to be significant, the manuscript might be rejected before publication or retracted if it has already been published. The authors' host institution might also need to be informed by the journal regarding the plagiarism</p>	Similar texts may be flagged up to the authors for revision

in the presence of any significant similarity or not) remains the purview of editors and reviewers. Table 1 summarises differences between plagiarism and similarity.

Practical considerations for journal editors and reviewers to identify plagiarism

It is preferable to use the output from similarity checking software as a starting point. Appropriate filters should have been used to exclude references as well as to set a word limit of 8-10 words before tagging a phrase as similar. Thereafter, such similarity reports should be manually reviewed in detail to identify portions that appear similar and span more than two or three lines. Editors should be careful that such stretches of words might have a few dissimilar words here and there which have been intentionally placed to avoid flagging of the text as unoriginal by similarity checking software.⁵ The source of such similar content as indicated by the software should be reviewed. Sometimes, the automated links to such sources do not function. In such a situation, those selections of text should be copied and pasted manually on to a search engine such as Google to identify the source, which should then be reviewed to confirm similarity. A common source of similarity is prior conference presentation abstracts, and this hardly indicates plagiarism. These can be easily identified by asking authors to declare prior conference presentations in the instructions to authors or during the process of peer review. Plagiarism of figures and tables is considerably difficult to identify. A search conducted on databases of figures such as Google images with the keywords related to the figure in question and manual checking of figures that appear similar with the submitted one is one way for editors and reviewers

to identify plagiarised figures. Similarly, a manual scrutiny of tables in previous articles, including review articles, remains the only way to identify plagiarised tables. Even in the absence of output from similarity checking software, experienced editors and reviewers might suspect plagiarism in manuscripts if there are abrupt changes in flow and tone between sections. This might suggest that such sections have been reproduced from written work by other authors.


Box 1 How can editors and reviewers detect plagiarism in manuscripts?

- Analyse outputs from similarity checking software as a screening tool. Appropriate filters should be used to exclude references and limit the sequence of identity of successive words to be considered similar.
- Beware intermittently changed words to avoid software-based plagiarism detection.
- Review reference lists of notable review articles and the order of such references, to detect potential plagiarism of ideas.
- Check images in articles through Google images.
- Compare tables in review articles to prior published reviews on the topic.
- Seek disclaimer from authors regarding absence of plagiarism.
- Seek declaration of prior conference presentations to avoid unnecessary flagging by similarity checking software.
- Beware abrupt changes in tone of manuscripts – were sections copied from elsewhere?

After completing such a review of the similarity report, authors might be asked to revise minor areas of similarity, or a decision might be taken to reject the manuscript if there are extensive areas of similarity across different sections, irrespective of the percentage of similarity. Furthermore, the instructions to authors (for new manuscripts) or revision comments (for revised manuscripts) might explicitly mention the need to seek permission from copyright holders along with appropriate referencing in instances where figures or tables have been reproduced with or without modifications from elsewhere. The authors might also be asked to add a disclaimer stating that the content and ideas expressed in the manuscript are original and have not been copied from elsewhere. Box 1 summarises such points for consideration by the journal editors and reviewers to detect plagiarism.

Future perspectives about plagiarism detection and prevention

The emergence of artificial intelligence and machine learning as a tool to automate diverse processes might also be

exploited to assist editors and reviewers to detect plagiarism, particularly of figures and tables. However, instances such as the difficulty of machine learning in distinguishing the image of a duck from a rabbit when rotated indicate that human oversight shall always be required no matter how sophisticated the automation might be.⁷ Since a significant proportion of retracted articles are due to plagiarism,^{8,9} databases of retractions such as the one provided by Retraction Watch might be linked to manuscript submission systems to flag authors whose manuscripts have been previously retracted due to research misconduct (including plagiarism). Such manuscripts might have a higher-than-average pre-test probability of misconduct, therefore, might be scrutinised more carefully by editors. Lack of education about what constitutes plagiarism remains an important determinant of such behaviour, especially in regions of the world such as Asia.⁶ Therefore, continuing attempts at educating authors worldwide regarding publication ethics and scientific writing should also include illustrative sessions on plagiarism.¹⁰ 

References

- 1 Ahmed S, Anirvan P. The true meaning of plagiarism. *Indian J Rheumatol* 2020; 15: 155-58.
- 2 Gasparyan AY, Nurmashev B, Seksenbayev B, Trukhachev VI, Kostyukova EI, Kitas GD. Plagiarism in the context of education and evolving detection strategies. *J Korean Med Sci* 2017; 32: 1220-27.
- 3 Misra DP, Ravindran V. Publication misconducts related to copyright: tread carefully to avoid falling. *J R Coll Physicians Edinb* 2020; 50: 3-5.
- 4 Higgins JR, Lin F-C, Evans JP. Plagiarism in submitted manuscripts: incidence, characteristics and optimization of screening – case study in a major specialty medical journal. *Res Integr Peer Rev* 2016; 1: 13.
- 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: e217.
- 6 Misra D, Ravindran V, Agarwal V. Plagiarism: software-based detection and the importance of (human) hardware. *Indian J Rheumatol* 2017; 12: 188-89.
- 7 Bhagyashree R. *Researchers input rabbit-duck illusion to Google Cloud Vision API and conclude it shows orientation-bias* [Internet]. March 2019 [Accessed 18 April 2021]. Available from: <https://hub.packtpub.com/researchers-input-rabbit-duck-illusion-to-google-cloud-vision-api-and-conclude-it-shows-orientation-bias/>
- 8 Misra DP, Agarwal V. Integrity of clinical research conduct, reporting, publishing, and post-publication promotion in rheumatology. *Clinical Rheumatol* 2020; 39: 1049-60.
- 9 Moylan EC, Kowalczyk MK. Why articles are retracted: a retrospective cross-sectional study of retraction notices at BioMed Central. *BMJ Open* 2016; 6: e012047.
- 10 Goyal M, Misra DP, Rajadhyaksha S et al. Effectiveness of a 1-day workshop on scientific writing conducted by the Indian Journal of Rheumatology. *Indian J Rheumatol* 2018; 13: 117-20.