Original article

Google versus other text similarity tools in detection of plagiarism: a pilot study in the *Journal of Clinical and Diagnostic Research*

Hemant Jain
Journal of Clinical and Diagnostic Research, Delhi, India; drhemantjain@jcdr.net
Sunanda Das
Journal of Clinical and Diagnostic Research, Delhi, India; drsunandadas24@gmail.com
Aarti Garg
Journal of Clinical and Diagnostic Research, Delhi, India; artigarg7@gmail.com; ORCID ID 0000-0003-3565-1025

DOI: 10.20316/ESE.2016.42.012

Abstract
Background: We practised using plagiarism detection software in the *Journal of Clinical and Diagnostic Research*, but after a few significant items were missed, we re-assessed our strategy and compared Google with three other text similarity programmes.

Method: 25 manuscripts (16 original articles and 9 case reports) were randomly selected, where the decision to publish had been greatly affected by plagiarism. These manuscripts were checked for plagiarism, searching each sentence using Google. The same manuscripts were run through three text similarity software programmes (iThenticate, Viper, and Plagiarism Checker X). For original research, we considered methodology, results, and discussion; and for case reports, we considered case details and discussion. Each report was checked by the investigators for scoring in addition to the percentage of plagiarism reported in the software.

Results: When checking original articles, Google performed the best, iThenticate missed a few minor sections, and Plagiarism Checker X had a lower number of hits, followed by Viper. On analysing the case reports, Google and iThenticate were found to be similar. Plagiarism Checker X missed minor sections, and Viper missed significant parts and was therefore considered less reliable.

Conclusions: Based on the study results, we suggest using two software programmes and manual verification of the manuscript.

Keywords: plagiarism, text similarity, detection, Google, iThenticate, journal.

Introduction

Plagiarism, defined as "appropriation of another person's ideas, processes, results or words without giving appropriate credit to the source or author",[1] is among the most grave types of misconduct and an unethical exploitation of scientific literature. Text similarity software has been developed to curb and check this misconduct.[2] The effect of plagiarism varies depending upon the type of article, ie research work, case reports, and reviews.[3]

In the *Journal of Clinical and Diagnostic Research (JCDR)*, we make a decision about plagiarism depending on which part of the manuscript is affected. This sectional assessment of plagiarism is also followed by certain other journals.[4–9] A study conducted by Baždarić and colleagues[4] in 2012 showed that 14% of manuscripts submitted to the *Croatian Medical Journal* were suspected of plagiarism. However, on manual verification, 11% of these submissions were considered plagiarised. Therefore, to avoid false positive results, a manual verification by an editor is required.

We practised using CrossCheck (iThenticate) in *JCDR*, to check for plagiarism, but after a few instances where we could identify misses, we re-evaluated our strategy. Later, we started to manually search for similar sentences using Google. We conceptualised a study to bring out any significant gain that we might have had by this change in our practice against the increased manual work. For the completion of the study we considered eight software programmes,[2] ran a feasibility test and found that three of them were easily assessable and usable. Our primary aim was to compare three text similarity programmes (iThenticate, Plagiarism Checker X, and Viper) against Google in detecting text similarity in our journal.

Methods

Since 2014, we have employed an in-house staff member to check for plagiarism using Google. The staff member was asked to provide all manuscripts, submitted from 1st January to 31st May, 2015, which were flagged during the initial screenings with markedly high text similarity.

JCDR received 1700 manuscripts in this time period. In 96 manuscripts, the decision was greatly affected due to the presence of text similarity. From these, 25 manuscripts (16 original manuscripts and 9 case reports) were randomly selected. We do not usually quantify the text similarity into percentages. The chunks of similar text are marked and the URL links of the source are tagged along the manuscript. In certain cases, the editor rejects a manuscript exclusively based on this so-called Google report and a review report where the reviewer advises rejection of the manuscript. In both situations, the text similarity is considered as an alarm for the decision on the manuscript—either major revision or rejection. The editorial policy is to remove the text similarities by rephrasing the sentences and citing the references, if the original article or case-report quality of data outweighs the draft originality.

Procedures

For the purpose of analysis, each original article was divided into three sections: methodology, results, and discussion; and case reports into two: case details and discussion. Thus, we had 48 sections in original articles and 18 in case reports, making a total of 66 sections to be analysed.

The presence of text similarity in an Introduction is ignored during the initial screening, hence this was not considered for analysis. In cases of plagiarism in Methodology, the journal allows limited referencing or redrafting, as in certain laboratory procedures. The laboratory procedures are standard and specific, thus an author cannot modify them for the sake of originality. As an example, the test for serum proteins or method for ELISA, staining procedures in histopathology.

Because it was a retrospective data analysis, all the manuscripts already had plagiarism reports. To avoid bias, the manuscripts were re-evaluated using Google and three plagiarism detection programmes. For Google, the process was to take each sentence and search for its match. The sentences which were word-by-word similar to the sources were highlighted and a comment box was added to the right side of the word document with the link to the source in the box. The comments were of the following types:

- Completely similar and referenced (the text was verbatim and the source has been cited as a reference)
- Completely similar and not referenced (the text was verbatim but with no reference citation of the source)
- Completely similar but not from the reference that was cited
- Sentence matched but the statistics/numerical values were different.

The text similarity was checked by two members of staff. One used Google and the other used the three text similarity programmes. For iThenticate and Plagiarism Checker X, membership was obtained and Viper, a free software, was downloaded in a computer system. Both of the staff were blinded to each other's findings. Each report was checked by the three investigators for scoring.

The findings of Google were considered as standard in this study. The 'percentage' marked in the software reports were not considered for the scoring. Since every sentence marked using Google was read by the investigators, the quantification was not done. Rather, the decision was based on the significance of the sentence with respect to the research or case.

The gradings for each section were:

- A (no plagiarism);
- B (suggestive of plagiarism, the amount of hits same as Google)—comparison between the software reports and Google;
- C (suggestive of plagiarism, the amount of hits less than Google)—the inter-group comparison among the three software programmes

The grades for each section were added and a final score given (ie total of each grade earned by a manuscript).

Results

The results of the comparison of Google with three text similarity programmes, based on scores are presented in Table 1. The final score for Google was 48 (72%) of 66, iThenticate 47 (71%), followed by Plagiarism Checker X, 44 (67%) and Viper, 33 (50%).

Original articles

Google marked 35 (73%) out of 48 sections which had text similarity. When these manuscripts were checked using iThenticate, it marked 34 (71%) sections of which 33 sections were marked similarly by Google (grade B). The remaining 1 section was marked less than Google (grade C). One section marked by Google was completely missed by iThenticate. Plagiarism Checker X marked a total of 31 (65%) sections. Of these, 26 were found to be similar as marked by Google (grade B). Five sections had fewer hits than Google (grade C). Therefore, it missed 4 sections.

The results from Viper showed that 23 (48%) of 31 sections were plagiarised. Of these 23, 18 were similar to Google (grade B), and 5 had fewer hits than Google (grade C). Thus, it failed to detect the text match in 12 sections.

On combining the total scores for Original articles, we found that iThenticate scored the best (34 of 35) and its hits were the most similar to that of Google, followed by Plagiarism Checker X (31 of 35), and then Viper (23 of 35).

Case reports

Google marked 13 out of 18 sections having text similarity. iThenticate had the same number of hits. 8 sections had the same texts marked as that in Google (grade B) and 5 sections had fewer (grade C). Similarly, Plagiarism Checker X marked 6 sections grade B and 7 grade C.

On combining the total scores for case reports, iThenticate and Plagiarism Checker X were the same but iThenticate reported more grade Bs, and hence it performed better than Plagiarism Checker X.

Discussion

Based on the results of our study, Google and iThenticate detected a similar amount of plagiarism. Google has an added advantage of using a large number of databases of images, figures, and tables. Plagiarism Checker X comes next and is the most cost-effective. On analysis of the study results, we noted that the misses by Plagiarism Checker X were greater than with iThenticate. But the misses can be nullified to some extent by using a second software or Google. Plagiarism Checker X can be subscribed for a nominal amount and has lifetime validity. Cross-checking multiple documents together and a side-by-side comparison of two documents can also be done using Plagiarism Checker X. The versions of this programme are Basic (free), Pro, and Business.[10] Being cost-effective, it can be helpful in resource-strained journals. After completion of the study, JCDR has started using Plagiarism Checker X, complemented by Google.

Table 1. Comparison of Google with three software programmes

	Original articles (maximum score 16×3=48)			**Case reports** (maximum score 9×2=18)			Final score (maximum score 48+18 =66)
	Same as Google (B)	Less than Google (C)	Total score B+C	Same as Google (B)	Less than Google (C)	Total score B+C	
Google	35	NA	35	13	NA	13	48
iThenticate (CrossCheck)	33	1	34	8	5	13	47
Plagiarism Checker X	26	5	31	6	7	13	44
Viper	18	5	23	6	5	11	33
NA=not applicable (as Google was considered the standard for comparison)							

When we started using Google to check for text similarities, we came across quite a few manuscripts which were found to have much less or no plagiarism according to the software, but with Google search, a significant part was found to be verbatim match with the sources. The rechecks were at times done on a hunch or due to certain valid reasons associated with manuscript drafting or author's reputation (past submissions to *JCDR*). With CrossCheck, the database of iThenticate has increased hugely. However, those manuscripts which are published in less established journals or are university publications might not be webbed by iThenticate. They might be listed only after a time period, when the software is updated. However, with Google, all publications can be searched, irrespective of the platform. This feature came to light while working on a manuscript submitted to *JCDR*. iThenticate had marked the manuscript as clean, but the topic was a very common research question and the concerned editor checked for similar publications using Google. Not only was the hunch validated, but the same manuscript was found to be published in another journal. This journal has a low impact and is published by a university.

From our experience with Viper, we could infer that the software takes a longer period of time to fetch results. Furthermore, the number of manuscripts that can be run in a day is limited. However, the advantage is that it does not require any charges for its usage.

The foremost advantage of Google is that it is easy to use. It can also search the largest number of databases. All the authors in a manuscript can be searched for against the title of that particular manuscript along with their other publications. This rules out cases of duplicate publication where an author (or others in the author list) is the first author in one and second (or later) in others.

Although the process of checking each sentence in the database and evaluating them manually is time-consuming, a person gains a faster hand through practice. Hence, we recommend that the practice of manual evaluation is also required. Gradually, the cost-benefit ratio improves.

The software's misses might be due to the frequency at which the databases are updated. Google might use databases that update most frequently. Although this feature was not completely analysed by the investigators, it has to be pointed out. None of the software has a provision for running a check on tables and figures. This is a major limitation of text similarity software.[11] iThenticate marks if the rows or columns have similar legends but the data is not verified. However, Google considers all tables and graphs as images and makes them available for search in Google images.

The *JCDR* has a full-time employee whose job description includes manual checking for plagiarism using Google. He starts by checking the manuscipt title along with the entire authors' list and then searches each sentence using Google. In a single day (8-hour shift), about 15 manuscripts can be searched. The process of checking each sentence in the database and evaluating them manually is time-consuming and the journal has to spend resources in terms of finance and space to employ a dedicated staff. On the other hand, iThenticate charges a nominal annual fee and a fee for each manuscript. For Plagiarism Checker X, the subscription can be obtained for a lifetime with a certain fee. The editorial board should be judicious and perform the cost-benefit analysis to utilise a programme or recruit an employee. The strength of the editorial team in numbers, the total remuneration for staff, past author behaviour as experienced by the journal etc should also be considered when formulating the plagiarism shunning policies of a journal.

The issue of plagiarism in scientific literature has deepened its roots. The reasons might be deliberate, for authors who do not give importance to originality of draft or are unaware of scientific methodology and research integrity policies. At times, plagiarism might arise in conditions where an author does not have a good command of English and ends up plagiarising the manuscript.[4,5]

The impact of plagiarism varies, depending on the type of manuscript, ie research work, case reports, or reviews. If text similarity is present in certain sections of manuscripts, where it can be fixed or it does not invalidate the work done, we believe a second opportunity should be provided, especially in the case of inexperienced authors whose writing skills are not scientifically appreciable.[3,5] Hence, we propose that decisions should be made after manual re-evaluation rather than based solely on the automated plagiarism report.

A total of 25 manuscripts might not be enough for a definitive inference. We were able to use only 3 software programmes as the others could not be taken up due to technical difficulties faced by the investigators. The secondary outcome regarding the timeline for software upgrade could not be further analysed as there were too few manuscripts. For a timeline check, a large-scale study is required with more manuscripts, and in which each manuscript is followed from submission to its final fate. Further study with more manuscripts and plagiarism detection tools is encouraged.

Every journal functions in a specific way. Editors improvise or innovate techniques for easy, faster, and fool-proof functioning. The results of this study, albeit on few manuscripts, are promising. If larger studies replicate these findings, it will be a breakthrough in the field of publication. Google can be used as a complementary tool to reduce the effect of the inherent shortcomings of text similarity software. Our manuscript assessment procedure has changed based on the results of this study. We now screen the manuscripts for text similarity using Plagiarism Checker X, right after the manuscript is submitted. If the text match is more than 35%, the manuscript is sent to the staff assigned for using Google, even before sending it for peer review. If both the reports mark a significant text match, the manuscript is rejected after being evaluated by an editor. This helps us to decide upon the plagiarised manuscripts faster and also saves on resources. For manuscripts which clear the text similarity checks, the Google report is sent to the authors along with the review report. Manuscipts that take a long time to reach the stage of acceptance for publication pass through the process of plagiarism checking again, using Google to detect multiple submissions.

We recommend use of any two tools for plagiarism check, either a software and Google or two different software programmes and at two points of time, for example first during the initial screening when the suitability of the manuscript is judged and again just before acceptance. Whatever results are obtained, they should always be re-evaluated by an editor.

Contributions

HJ conceived and designed the study, and proofread the article. SD and AG were responsible for data collection and analysis, and drafting.

Conflict of interests

The investigators/authors declare that they have no affiliation to any of the text similarity software.

Acknowledgements

We would like to thank Bijesh Mishra for the statistical analysis and running the manuscript through the plagiarism checking software programmes. We are also thankful to Aashay Saxena for manually checking the manuscripts for plagiarism using Google.

References

1 ORI. Office of the President, Office of science and Technology Policy: Federal Research Misconduct Policy. *Federal register* 2000; 65: 76260–4.

2 Rathore FA, Farooq F. Plagiarism detection softwares: Useful tools for medical writers and editors. *Journal of Pakistan Medical Association* 2014; 64: 1329–30.Available at: http://jpma.org.pk/PdfDownload/7088.pdf

3 Rosen RM. Plagiarism in the Medical/Scientific Literature. *Journal of Cardiovascular Pharmacology* 2010; **56:** 709.

4 Baždarić K, Bilić-Zulle L, Brumini G, Petrovečki M. Prevalence of plagiarism in recent submissions to *Croatian Medical Journal*. *SciEng Ethics* 2012; 18: 223–9. DOI: 10.1007/s11948-011-9347-2

5 Baždarić K. Plagiarism detection- quality management tool for all scientific journals. *Croatian Medical Journal* 2012; 53: 1–3. DOI: 10.3325/cmj.2012.53.1

6 Masic I. Plagiarism in scientific publishing. *Acta Inform Med* 2012; 20: 208–13.

7 Castillo M, Halm K. Cross-Checking for plagiarism. *AJNR* 2008; **29:** 1035.

8 Recommendations for the Conduct, Reporting, Editing and Publication of Scholarly Work in Medical Journals–Updated December 2015. ICMJE; 2015. Available at: http://www.icmje.org/icmje-recommendations.pdf [accessed April 4, 2016]

9 COPE. COPE flowchart. 2006. Available at: http://publicationethics.org/files/u7140/Full%20set%20of%20flowcharts.pdf [accessed April 4, 2016]

10 Plagiarism Checker X. http://plagiarismcheckerx.com/online-plagiarism. (accessed May 20, 2015).

11 Elsevier. Plagiarism detection. Available at: https://www.elsevier.com/editors/publishing-ethics/perk/plagiarism-complaints/plagiarism-detection (accessed August 30, 2016).