Bulletin of the National
Research Centre

RESEARCH

Open Access

A plagiarism paperdemic: determining plagiarism among COVID-19 articles in infectious disease journals between 2020 and 2021

Rahma Menshawey[1*], Esraa Menshawey[2], Ahmed Mitkees[2] and Bilal A. Mahamud[1]

Abstract

Background The COVID-19 pandemic has caused drastic changes in the publishing framework which allowed for the quick review and rapid publication of manuscripts in order to quickly share vital information about this new viral pandemic to the general public and scientists. Alarms have been raised for the potential for misconduct in COVID-19 research. The purpose of this study is to determine the presence of plagiarism in COVID-19 papers across infectious disease journals.

Methods COVID-19 related research and review articles published in infectious disease journals were collected. Each manuscript was optimized and uploaded to Turnitin, which is a similarity checking tool. Similarity reports were manually checked for events of true plagiarism using an 80% threshold, performed via human judgment.

Results In this cross-sectional study, 41.61% (*n* = 129) of manuscripts were deemed plagiarized out of a total of 310 papers that were analyzed. Plagiarism was identified in 35.07% of reviews (*n* = 47), and 46.6% of original research (*n* = 82). Among the plagiarized papers, the median number of copied sentences was 3 IQR 4. The highest recorded similarity report was 60%, and the highest number of copied sentences was 85. The discussion section of these articles was the most problematic area, with the average number of copied sentences in that section being 6.25 ± 10.16. The average time to judge all manuscripts was 2.45 ± 3.09 min. Among all the plagiarized papers, 72.09% belonged to papers where the similarity report was ≤ 15% (*n* = 93). No significant differences were found with regards to plagiarism events among the quartiles.

Conclusions Plagiarism is prevalent in COVID-19 publications. All similarity reports should be supplemented with human judgment.

Keywords COVID-19, Ethics, Plagiarism, Similarity report, Turnitin

Background

"It is better to fail in originality, than to succeed in imitation." Herman Melville (Melville 1850).

The Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV-2) was first discovered in China (Wuhan city) in 2019, and declared a global pandemic by the WHO in March of 2020 (He et al. 2020). Since then, research on the virus has taken the scientific world by storm. A massive number of articles has been published on the topic,

*Correspondence:
Rahma Menshawey
rahma.menshawey.94@gmail.com
[1] Kasr Al Ainy Hospital, Faculty of Medicine, Kasr Al Ainy, Cairo University, Geziret Elroda, Manial, Cairo 11562, Egypt
[2] Kasr Al Ainy School of Medicine, Cairo University, Cairo, Egypt

as evidenced by a Google scholar search with the key word "COVID-19" from 1/1/2020 to 31/12/2021 which yields a total of 1,310,000 articles.

In an effort to overcome scientific publishing barriers during this public health emergency, some journals have adopted a fast-tracked means to publish (Carvalho et al. 2021). Changes in publication models and priorities may have inadvertently compromised the integrity and quality of research on this topic. Evidence of compromised academic integrity can be seen in the alarming number of retracted COVID-19 papers listed by Retractionwatch.com (354 retracted articles as of 2023–08-01). The retraction record for COVID-19 papers has greatly surpassed other viral related topics, whose basal record of retractions is 4 out of every 10,000 papers (Yeo-Teh and Tang 2021). There have been several alarms made about events of fraudulent science and misconduct regarding research on the COVID-19 topic. Scientific misconduct is not without consequences and has resulted in major financial loss, court trials, and retractions (Dinis-Oliveira 2020).

One type of misconduct that has yet to be examined in COVID-19 publications is plagiarism. Plagiarism is the appropriation of another person's work (Aronson 2007). In publications, plagiarism may be observed as the direct copying of text from another paper (verbatim plagiarism), or excessive copying of one's own previously published work (self-plagiarism) (Burdine et al. 2018). Other types of plagiarism include; Mosaic plagiarism (the mixing of one's words with another), Paraphrasing (restating someone else's words or ideas by changing very few words), and image plagiarism (the use of pictures or figures without permission of the owners) (Masic 2014; Dhammi and Ul Haq 2016).

Most journals have policies denouncing plagiarism and claim to use programs to check for similarity, such as Turnitin and iThenticate (Meo and Talha 2019). Similarity checking tools such as iThenticate or Turnitin provide a similarity report, which is not a definitive indicator of plagiarism (Meo and Talha 2019). A high percentage similarity report may likely denote plagiarism, but reports with a lower percentage may be misinterpreted if considered independently. Furthermore, journals and publishers have a threshold for the maximum allowable similarity report result, which is typically 15% (Polyanin and Shingareva 2022). A blind spot to plagiarism appears here; a low level of similarity does not automatically exclude plagiarism, as plagiarism can be directly copied text within the threshold selected by the journal or publisher. Another blind spot can occur in the peer review process, as it is unclear if journals employ another similarity check after revisions have been suggested to the authors. At that moment, changes to the manuscript may have introduced added similarity, and even plagiarism to the text. Ultimately, similarity reports can be misleading (Polyanin and Shingareva 2022).

It is for this reason that a manual check via human judge is needed to supplement the similarity report to determine if plagiarism exists within the text. The aim of this study was to evaluate the presence of plagiarism in COVID-19 related papers within infectious disease journals, using the similarity report provided by Turnitin.

Methods

We surveyed papers published in infectious disease journals from four quartiles, as listed on SCOPUS journal rankings, for original research and reviews relating to the COVID-19 topic from Jan 1 2020 to Dec 31 2021. The first 5 journals from each quartile were selected after an initial screen of the journal title and scope to ensure that the journal focused on the general topic of infectious diseases (Q4 required eight journals due to the small number of articles published in this quartile) (see Table 1).

Inclusion and exclusion criteria of articles

Papers were selected through each journal's search engine in their order of publication using keywords including; "SARS-COV-2," and "COVID-19" (both these terms were developed in the year 2020, and therefore acted as a type of temporal limit in the search—this was useful where search engines were not advanced enough to allow search by dates).

We excluded from our search the following: commentary, case reports, case reviews, editorials, letters to the editors, highlights, non-English articles, etc. Papers met our initial screening via title, and abstract reading to ensure relevance to the COVID-19 topic. We only included papers for which the full text was available. The full text was downloaded, and an optimized manuscript was developed for each.

Optimized manuscripts

Optimized manuscripts were developed for each of the papers that met our inclusion criteria. This idea was inspired by the methods of Higgins et al. (Higgins et al. 2016). Our optimized manuscripts included the abstract, introduction, results, discussion and conclusion sections. The methods section was removed as this section can contain high similarity between papers of a similar topic as the technical language is often reused (Sun et al. 2010). Additionally, the truer indicator of plagiarism is if the copying comes from the results section (Meo and Talha 2019). Based on this logic, we removed the methods section from all the manuscripts we analyzed. These optimized manuscripts were then uploaded to Turnitin and analyzed.

Table 1 Comparison Table—Quartile, Plagiarism detected in each Quartile, Journal, Papers Analyzed, Plagiarism Policy, Cite Score, and Publisher

Quartile	Plagiarism detected in each Quartile	Journal	Number of Papers analyzed	Presence of Plagiarism policy or warning in Author Guidelines	Cite score	Publisher
Q1	36.7% (*n* = 29)	Immunity	20	Yes	46	Elsevier
		Nature Reviews Microbiology	16	Yes	48.4	Springer Nature
		Clinical Microbiology Reviews	15	Yes	47.5	American Society for Microbiology
		Lancet—Infectious Diseases	20	Yes	50.3	Elsevier
		Trends in Microbiology	8	No	23.9	Elsevier
Q2	52.5% (*n* = 52)	European Journal of Clinical Microbiology and Infectious Disease	19	Yes	3.27	Springer
		Virus Research	20	Yes	7.2	Elsevier
		Travel Medicine and Infective Disease	20	Yes	14.8	Elsevier
		Journal of Microbiology, Immunology, and Infection	20	Yes	12.0	Elsevier
		Expert Review of Anti Infective Therapy	20	No	6.5	Taylor and Francis
Q3	20% (*n* = 13)	Current Clinical Microbiology Reports	6	No	6.3	Springer Nature
		Infection Disease and Health	20	Yes	3.7	Elsevier
		Canadian Journal of Infectious Diseases and Medical Microbiology	14	Yes	3.7	Hindawi
		Germs	8	No	2.1	European Academy of HIV/AIDS and Infectious Diseases
		The Brazilian Journal of Infectious Diseases	17	Yes	3.5	Elsevier
Q4	52.2% (*n* = 35)	Clinical Microbiology Newsletter	5	Yes	1.5	Elsevier
		Biosafety and Health	22	Yes	4.8	Elsevier
		The Ethiopian Journal of Health Development	4	No	1.0	Ethiopian Public Health Association
		Molecular Genetics, Microbiology and Virology	2	Yes	0.6	Pleiades Publishing
		Journal of Communicable Disease	8	Yes	0.2	Indian Society for Malaria and Other Communicable Diseases
		Global Epidemiology	6	Yes	2.0	Elsevier
		Infectious Diseases In Clinical Practice	13	No	0.4	Wolters Kluwer Health
		JAMMI—Official Journal of the Association of Medical Microbiology and Infectious Disease in Canada	7	No	1.8	University of Toronto Press

Turnitin program

Turnitin is an internet-based tool developed by iParadigms LLC for the purpose of recognizing similarity among electronically submitted documents, and is used by institutions as a tool to detect plagiarism. Its database provides a repository of over 70 billion webpages, 1 billion student papers, and scholarly content of over 1700 publishers (Meo and Talha 2019).

The following settings were applied to Turnitin for each uploaded optimized manuscript:

- Only periodicals, journals, publications, were included
- No student papers were considered in our analysis

The following was excluded from the results of the similarity report:

- Under filter and settings: exclude quotes, exclude bibliography, exclude sources that are less than 10 words.
- The original source being analyzed as well as any external link, i.e., a website or repository, that was hosting the exact manuscript. If the publication was hosted on a preprint server, the preprint source was excluded.
- Any publication that was published after the date of the analyzed article was excluded. Due to the fast-tracked peer reviews that was adopted for COVID-19 research, we excluded papers that were matched up to 1 month before the publication date of the analyzed paper.

The final report was downloaded and analyzed for results.

Outcomes

Our outcomes included:

- The percentage result of the similarity report
- Country of origin (corresponding authors first listed institution)
- Language of the country of origin
- Number of authors
- Whether or not there was plagiarism
- Number of plagiarized sentences and their location in the text
- The presence of self-plagiarism. This was determined by if the majority of the similarity was coming from any other published paper belonging to any of the authors, that was listed as the top match in the similarity report
- Time in minutes to analyze the similarity report

Determination of plagiarism

Each similarity report was analyzed for plagiarism manually.

Our criteria for plagiarism were the following:

- If 80% of a sentence was found to be identical to previously published sources, the sentence was counted as plagiarized (Higgins et al. 2016). This was manually calculated by counting each highlighted word in the report and dividing it by the total number of words in that sentence.
- If a single sentence was determined to be plagiarized, then the whole manuscript was scored as plagiarized.

Exclusion criteria for plagiarism

We excluded the following sentences from being tallied as part of plagiarized text:

- Standard sentences – descriptive sentences, or definitions were excluded. Sentences determined by the judges (the authors) that could not be worded any other way, (provided it was appropriately referenced) were also excluded.

  For example "The COVID-19 pandemic is caused by the virus SARS-Cov-2 that was first discovered in Wuhan China…."

- If the use of a conjunctive adverb was the only part of the sentence not highlighted, we did not include that word(s) among the total number of words in each sentence. Examples of these terms were: however, furthermore, moreover, additionally, as well as statements such as "in our study" and "in their study." We justified the removal of these words from the total word count in a sentence as this can be done to conceal verbatim text copying.

Finally, an initial pilot study was conducted using n=3 papers from each quartile, or a total of 12 optimized reports, and examined by each of the 4 authors independently to confirm inter-rater agreement with the prescribed methods for identifying plagiarism; no differences were found between the judges.

Statistics

Descriptive statistics were reported as percentages, frequencies, averages and standard deviations. Shapiro–Wilk test was used to determine the normality of the data. When the distribution of variables was not Normal, the degree of relationship between the variables was determined using Rank correlation (Spearman's Rho). The optimal cutoff of the Turnitin Similarity score was explored using ROC analysis. All statistical analysis was performed using MedCalc for Windows version 19.1 ((MedCalc Software, Ostend, Belgium). A P value <0.05 were considered statistically significant.

Results

We examined a total number of 310 articles ($n=176$ original research, $n=134$ reviews) (see Fig. 1). 183 were published in the year 2021 and 127 were published in 2020 (see Table 2). The presence of plagiarism was observed in 41.61% ($n = 129$) of manuscripts. Shapiro Wilk Test revealed non-normal distribution of data ($W = 0.0042$). Similarity reports revealed a median similarity of 6.5 (IQR 8) *among all analyzed papers*. The lowest value of any report was 0%, and the highest was 60%. Plagiarism was identified in 35.07% of reviews ($n = 47$), and 46.6% of original research ($n = 82$).

Among the plagiarized papers, the median number of copied sentences was 3, IQR 4 sentences (lowest value = 1, highest value = 85) (Shapiro–Wilk test showed non-normal distribution, $W = 0.4734$). A significant negative correlation was found between the year of publication and the presence of plagiarism, rho = − 0.0135 ($p = 0.0173$). Among the papers published in 2020, 48.81% had plagiarism ($n = 62$), while in 2021 publications plagiarism was seen in 36.61% ($n = 67$) of papers. A significant negative correlation was found between the number of authors and the presence of plagiarism (Spearman rho = − 0.119, $P = 0.0364$,

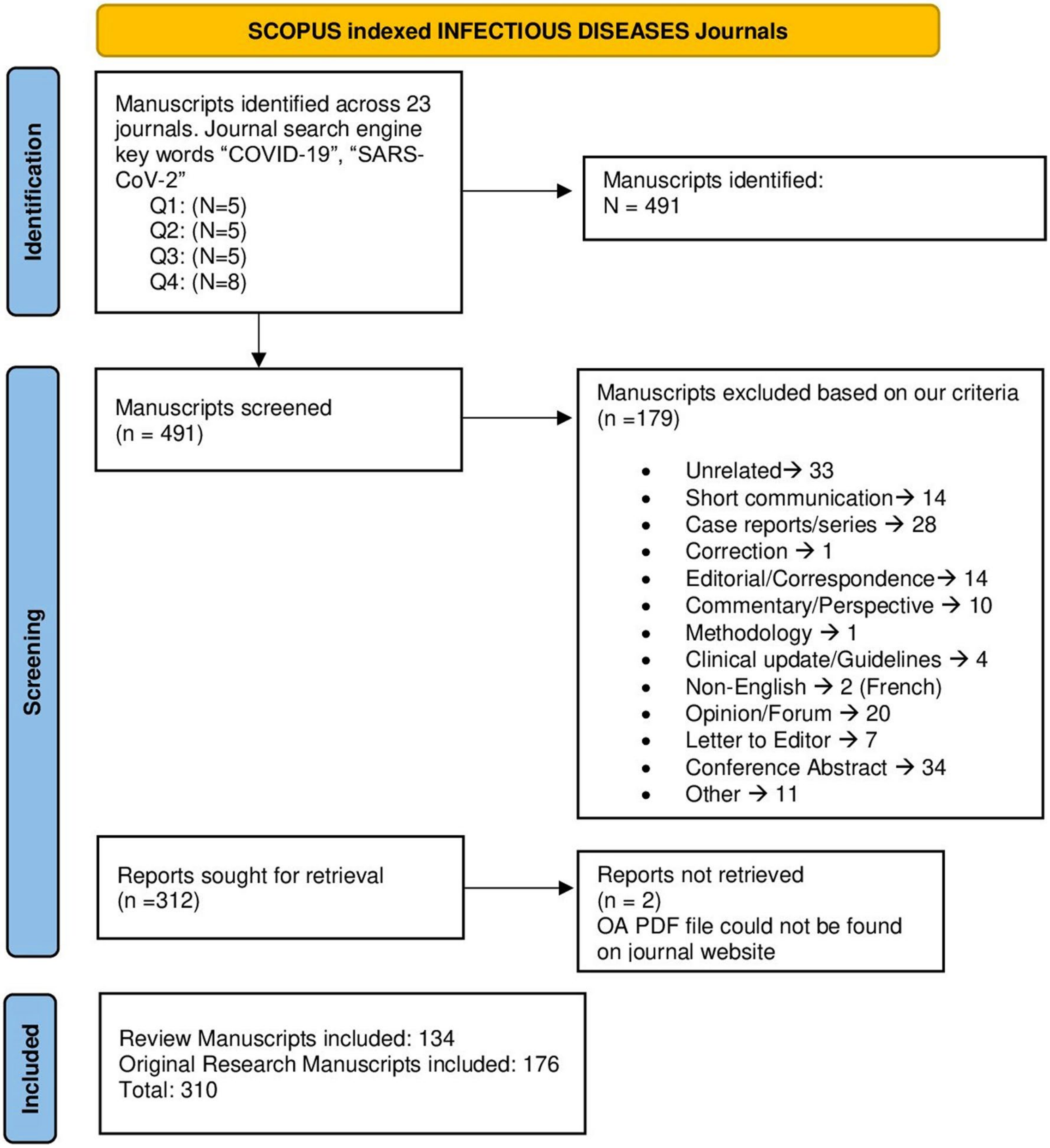
SCOPUS indexed INFECTIOUS DISEASES Journals
Identification
Manuscripts identified across 23 journals. Journal search engine key words "COVID-19", "SARS-CoV-2"
Q1: (N=5)
Q2: (N=5)
Q3: (N=5)
Q4: (N=8)
Manuscripts identified: N = 491
Screening
Manuscripts screened (n = 491)
Manuscripts excluded based on our criteria (n =179)
Unrelated→ 33
Short communication→ 14
Case reports/series → 28
Correction → 1
Editorial/Correspondence→ 14
Commentary/Perspective → 10
Methodology → 1
Clinical update/Guidelines → 4
Non-English → 2 (French)
Opinion/Forum → 20
Letter to Editor → 7
Conference Abstract → 34
Other → 11
Reports sought for retrieval (n =312)
Reports not retrieved (n = 2) OA PDF file could not be found on journal website
Included
Review Manuscripts included: 134
Original Research Manuscripts included: 176
Total: 310

Fig. 1 Flow diagram depicting the search strategy. Manuscripts were acquired across 23 journals, and a total of 310 manuscript met the inclusion criteria for further analysis

Table 2 Summary Statistics for Turnitin Similarity Report

	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Number of papers examined	79	99	65	67
Average Number of Authors	14.82 ± 12.38	7.82 ± 4.76	7.0 ± 4.59	4.99 ± 3.4
Median Number of Authors	10	7	6	4
Similarity report (avg ± std)	6.9% ± 5.66	9.89% ± 8.81	7.31% ± 8.28	8.76% ± 7.81
Similarity report (Median, IQR)	6.5 (IQR, 8)			
Presence of plagiarism	36.7% (*n* = 29)	52.5% (*n* = 52)	20% (*n* = 13)	52.2% (*n* = 35)
Average # of copied sentences	0.95 ± 2.08	2.89 ± 5.71	2.32 ± 10.92	2.79 ± 5.0
Median # of copied sentences	3 (IQR4)			
Redundancy	3.8% (*n* = 3)	19.2% (*n* = 19)	4.6% (*n* = 3)	8.9% (*n* = 6)
Average Time Spent in analysis (min)	2.44 ± 2.09	2.55 ± 2.84	2.25 ± 3.8	2.51 ± 3.67
Far Out Reports (based on Tukey Test)	1 32%	4 32%, 34%, 36%, 61%	3 33%, 33%, 40%	0

$R^2 = 0.9728$). The median number of authors was six (95% CI 6.0–7.0).

Based on the country of origin of the corresponding author, a publication was further determined as "English" if belonging to a core Anglosphere country (USA, UK, Canada, Australia, Ireland, English speaking Caribbeans). Based on this definition 27.74% of the publications came out of core anglosphere countries ($n = 86$). A significant negative correlation was found between papers published in core anglosphere countries and the presence of plagiarism, rho = − 0.172 ($p = 0.0023$) (see Fig. 2). This suggests that Anglophone speaking countries are not associated with plagiarism. This may be due to the difficulty of writing English faced by non-native users which may present as difficulty in phrasing original statements (Husain et al. 2017). Additionally, the concept of plagiarism and copyright appears to differ between English and non-English speakers (Maxwell et al. 2008).

The presence of self-plagiarism was identified in 31 publications overall. The average time spent analyzing the plagiarized reports was 4.53 ± 3.80 min, while the average time spent on the non-plagiarized papers was 0.96 ± 0.88

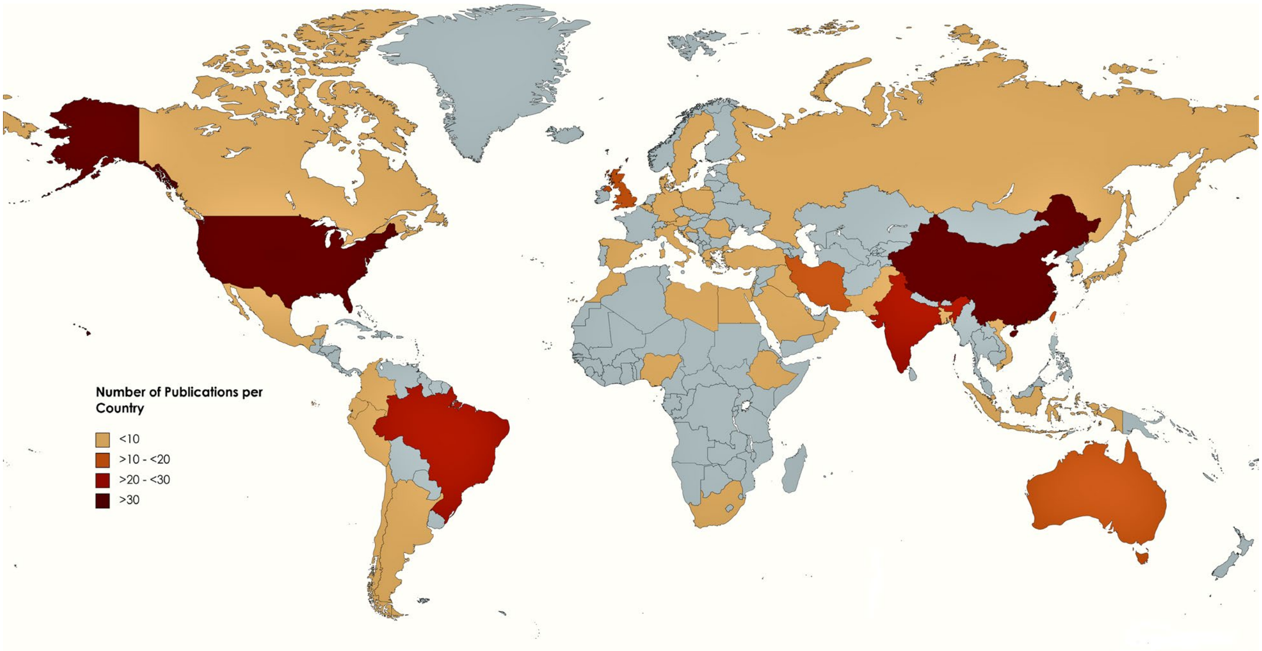

Fig. 2 Map depicting the country of origin of the examined papers. Country of origin was based on the country of the affiliation of the corresponding author of each paper. USA and China had the most publications

min. The area with the most copying was the discussion section (*n* = 85), with the average number of copied sentences in that section being 6.25 ± 10.16. The area with the second most number of copied sentences was the introduction with 1.6 ± 1.47 sentences (see Fig. 3).

ROC analysis revealed an area under the curve (AUC) 0.828 (95% CI 0.781 to 0.868, P < 0.0001) with an optimal criterion of 6% which maximized sensitivity and specificity (specificity being 72.38 (95% CI 65.3 to 78.7) and sensitivity being 81.40 (95 CI 73.6 to 87.7)) (see Fig. 4). Based on the "accepted" similarity levels for journals of 15% (Polyanin and Shingareva 2022) –sensitivity becomes 27.91 (95% CI 20.4 to 36.5) and specificity increased to 97.24. Increasing the threshold decreases the sensitivity while increasing the specificity. The sensitivity in this scenario represents the ability of the similarity report to identify a plagiarized article, while specificity represents how well the similarity report identified articles without plagiarism. Only at a criterion of 4%, sensitivity reaches 90.7% (see Table 3). While a similarity report of 6% or less is not impossible (*n* = 269, 86.77%), it can difficult given contested areas of similarity such as the methods section which uses recycled language, bearing in mind that we did not include any methods section in our analysis. Therefore, the Turnitin similarity report when used alone, is not reliable at 15% threshold to identify those articles that contain plagiarism.

Plagiarism was identified in 72.09% of the plagiarized papers where the similarity report was 15% or less (*n* = 93, 93/129). Therefore, nearly one-third of the total papers we examined contained plagiarism that would not have otherwise been captured without human judgment. This supports the position that the majority of plagiarism occurs below the 15% cutoff for similarity used by journals and publishers (Polyanin and Shingareva 2022).

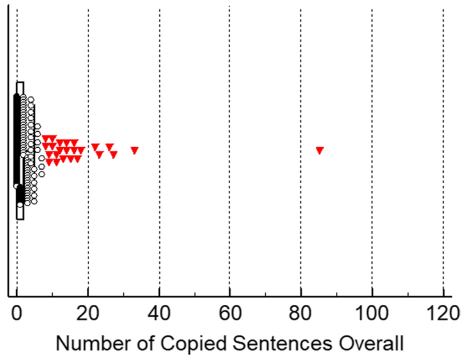

Fig. 3 Figure depicts the number of copied sentences overall among the articles deemed plagiarized. We deemed an article to be plagiarized if it contained at least 1 sentence that matched 80% to another source on the similarity report. The highest number of sentences copied identified in a single manuscript was 85

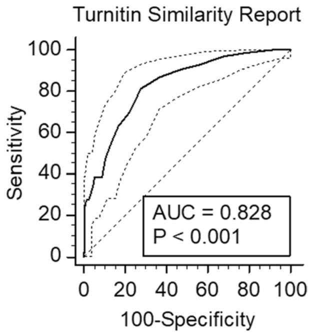

Fig. 4 ROC curve and AUC. Specificity and Sensitivity were maxed at a 6% cutoff. This suggests that the similarity reports are unreliable when used alone to determine the presence of plagiarism in a manuscript. Sensitivity in at this cutoff was only 81.40

We examined papers from all four quartiles (Q1 = 79, Q2 = 99, Q3 = 65, Q4 = 67). Using Shapiro–Wilk test normal distributions between the number of papers analyzed among the quartiles was observed (*P* = 0.318, accept normality). No significant difference was found with regards to the presence of plagiarism among the quartiles (high impact = Q&Q2, versus low impact = Q3&Q4), *P* = 0.1071 (see Table 2).

Similarity reports were highest in Quartile 2, and the least in Quartile 1; 9.89% ± 8.81 and 6.9% ± 5.66,

Table 3 Range of Specificity and Sensitivity at different Criterion for Turnitin based on Our Results

Criterion	Sensitivity	95% CI	Specificity	95% CI
≥ 0	100.00	97.2–100.0	0.00	0.0–2.0
> 0	100.00	97.2–100.0	8.29	4.7–13.3
> 1	98.45	94.5–99.8	22.10	16.3–28.9
> 2	96.90	92.3–99.1	32.04	25.3–39.4
> 3	93.02	87.2–96.8	43.65	36.3–51.2
> 4	90.70	84.3–95.1	53.04	45.5–60.5
> 5	86.82	79.7–92.1	62.98	55.5–70.0
> 6	81.40	73.6–87.7	72.38	65.3–78.7
> 7	69.77	61.1–77.5	77.90	71.1–83.7
> 8	63.57	54.6–71.9	82.87	76.6–88.1
> 9	55.04	46.0–63.8	86.74	80.9–91.3
> 10	47.29	38.4–56.3	89.50	84.1–93.6
> 11	38.76	30.3–47.7	91.16	86.0–94.9
> 12	38.76	30.3–47.7	94.48	90.1–97.3
> 13	34.11	26.0–43.0	95.58	91.5–98.1
> 14	30.23	22.5–38.9	96.69	92.9–98.8
> 15	27.91	20.4–36.5	97.24	93.7–99.1
> 16	27.91	20.4–36.5	98.34	95.2–99.7
> 17	26.36	19.0–34.8	98.90	96.1–99.9
> 18	23.26	16.3–31.5	99.45	97.0–100.0

respectively (see Fig. 5). The number of far out similarity reports were the most in Quartile 2 (with the highest recorded similarity percentage being 61%). Quartile 4 had no publications of major concern. Based on the results of Tukey test for far out variables, we used this to determine which of these papers required the attention of the Editor in Chief of the respective journals and informed them of our findings through email. Quartile 2 showed 52.52% plagiarized papers ($n=52$ out of 99 articles in that quartile), while Quartile 3 had the least with 30%. Self-plagiarism was highest in Quartile 2 (19.2%), and the least in Quartile 1 (3.8%).

Combined, quartiles 1 and 2 had the most articles with plagiarism despite seemingly acceptable similarity reports. High quartile journals may be more stringent in their screening policies and peer review, in order to catch misconduct pre-publication (Elango 2021). On the other hand, post-publication retractions are more common in high quartile journals which reflects their interest in correcting the scientific literature, with plagiarism being one of the leading reasons.

Lastly, most journals, 69.56% ($n=16$) had addressed their stance on plagiarism or their use of similarity checking tools, within the author guidelines.

Discussion

Plagiarism is defined as adopting another person's work or ideas. It is considered a severe academic offense and is a type of misconduct (other types if misconduct include fabrication and falsification). There are many factors which can explain but do not excuse plagiarism. These include academic pressure to publish, the number of authors, inexperience of the authors, length of the manuscript, and poor citation skills. (Debnath 2016). There even exists cultural differences in the perception of concepts such as plagiarism and copyright (Maxwell et al. 2008). Plagiarism is one of the leading causes of retraction of academic publications (Campos-Varela and Ruano-Raviña 2019).

We examined 310 consecutively occurring COVID-19 papers in infectious disease journals among 4 quartiles to determine the presence of plagiarism. We found evidence of plagiarism in 41.61% of the examined papers ($n=129$). There are limitations to using similarity checking tools to also detect plagiarism. While high similarity between texts can likely suggest plagiarism is at hand, a low value does not exclude it. In our study, at a criterion of 15% similarity, the tools sensitivity was only 27.91.

Several blind spots within the publishing process may miss the presence of plagiarism. Within the submission

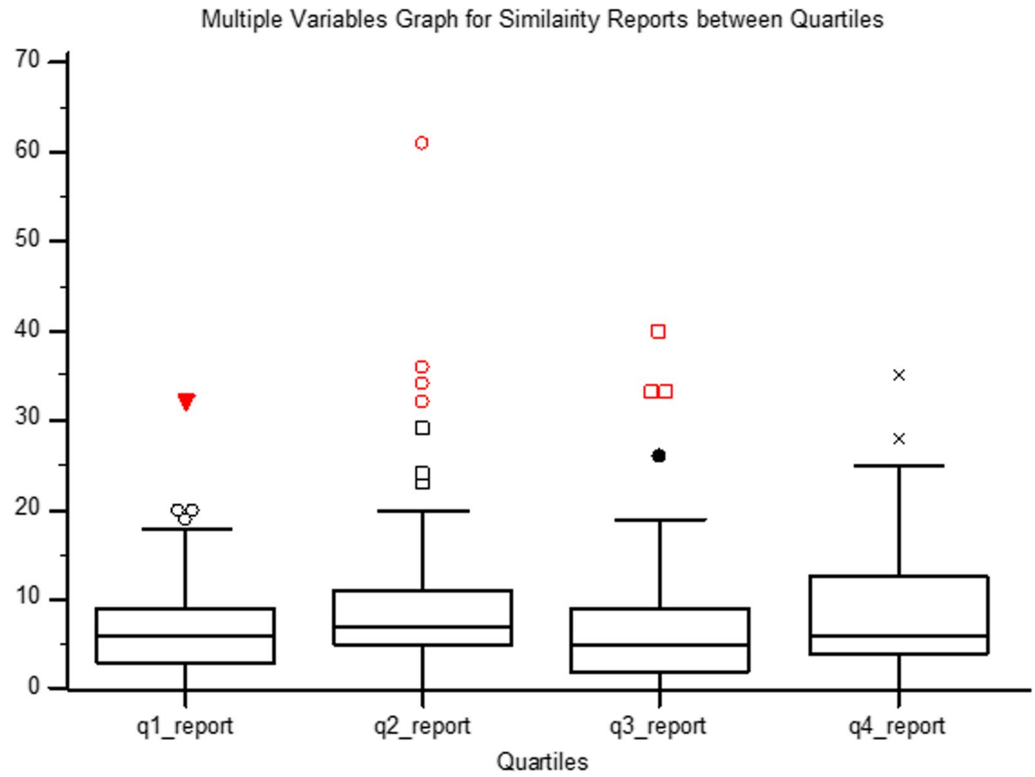

Fig. 5 Similarity reports for all quartiles. No significant differences were found with regards to the similarity reports between the high and low quartile journals ($P=0.1071$). Quartile 2 journals contained the most plagiarism

process, journals often employ text-matching programs on a manuscript before it reaches reviewers or editors. If a high similarity is observed, an immediate rejection is often sent to the authors or a request to address the issue and then to resubmit. The threshold that denotes whether a manuscript can progress in the publication process is typically 15% (Polyanin and Shingareva 2022). The blind spot is that within this 15%, the manuscript can still be copied text if the report is not supplemented with human judgment. Another area of concern lies in any decision for revisions, where the manuscript is returned to the authors for further adjustment. This may inadvertently introduce added similarity or even plagiarism to the text. It is unclear if journals employ a second re-check for the manuscripts after a response to the reviewers, and if not this could be a useful consideration.

In the context of COVID-19, drastic changes were made to the publishing framework to swiftly share vital information with the public and scientists, such as fast-tracked peer review and prioritization of COVID-19 research. These changes call for intensified scrutiny of all publications (Dinis-Oliveira 2020). Our study is the first study to examine the presence of plagiarism in COVID-19 related papers in a cross-sectional way among varying quartiles. Our findings may inspire further scrutiny in other areas or specialities that saw an influx in COVID-19 research, such as pulmonology or intensive care.

Few other studies have examined plagiarism among papers within speciality journals. In a study by Higgins et al. which examined plagiarism in 399 submitted manuscripts in a major speciality genetics journal, plagiarism was found in 17% of the analyzed articles (Higgins et al. 2016). Similar to our findings, they found that plagiarism was highest in countries where English was not the official language. They spent an average time of 5.9 min to analyze plagiarized reports, while the non-plagiarized papers took 1 min to assess. Similarly, we discovered that plagiarized papers took an average time of 4.53 ± 3.80 min to analyze, while non-plagiarized papers took an average of 0.96 ± 0.88 min. Higgins et al. identified self-plagiarism in 53% of plagiarized reports, while in our study we observed a lower percentage of 25.58% and this may be attributed to the relative novelty of the COVID-19 research at this time.

A study by Baskaran examined the events of plagiarism in a major speciality andrology journal using two tools; Turnitin and iThenticate (Baskaran et al. 2019). The Turnitin reports revealed an average similarity of 8.66% ± 8.62. Turnitin gave a higher mean similarity report, and this may be due to the larger database against which it compares submitted work (>70 billion publications) (Baskaran et al. 2019). Unlike Basakaran et al. who observed higher similarity report for reviews, our study found that similarity reports were higher in original papers. We hypothesize that the relative novelty of the COVID-19 topic may have caused this difference, especially given that reviews were likely being extrapolated and published based on previous models of disease before original research was available.

In our study, the most problematic area of manuscript was the discussion section, with an average number of copied sentences being 6.25 ± 10.16. A study by Rohwer examined events of plagiarism among African Medical journals using Turnitin and discovered 17% of papers had evidence of plagiarism (Rohwer et al. 2018). Likewise, the most problematic areas they identified were the introduction and discussion sections. Another study investigated plagiarism in 110 manuscripts that were submitted to the American Journal of Roentgenology; plagiarism was identified in 10.9% of the manuscripts, while the methods and discussion sections as areas contained the most plagiarism (Taylor 2017). In this study it was concluded that the current methods to identify plagiarism are suboptimal, and our findings are in agreement with this. They proposed that an improved method to identify plagiarism would involve screening of manuscripts with reports ≥ 15% where sensitivity reached 100%. In our study however, sensitivity reached 100% only when similarity reports were 0%. Reports reached 90% sensitivity only at 4% similarity reports, and this stresses the need for manual interpretation of the reports.

Despite some of the limitations of similarity reports and their potential for misinterpretation, they remain a useful indicator of misconduct and retractions. A study that analyzed the similarity index of 131 retracted anesthesia articles found an extensive degree of plagiarism (> 35% score) in the articles irrespective of the cause of retraction. The similarity index was a reasonable indicator of plagiarism and fabrication (El-Tahan 2019). One study analyzed the costs of iThenticate and staff wages and found that the annual cost to use the tool and analyze the reports (assuming 10 min was spent on each manuscript) was $6804.48. The enormous cost of dealing with misconduct, retractions, and damage to a journals reputation as a result, justify the cost of the broader and more extensive use of similarity checking tools.

One of the major factors believed to have influenced the rise in plagiarism seen in the past decade is the mounting academic pressure to publish (Baskaran et al. 2019). The novelty and the worldwide focus on COVID-19 has exacerbated plagiarism (Dinis-Oliveira 2020), and this is likely why we observed a larger number of plagiarized texts when compared to other studies on non-COVID-19 articles. The COVID-19 legacy on academic integrity has yet to be scoped to its full extent and is an important area for future research. Plagiarism is just the tip of the iceberg, and similarity checking tools are limited in several regards as they can typically only identify verbatim text copying which is only one of many types of plagiarism. Retractions for COVID-19 papers are at an all-time high—more attention and action is needed against all forms of misconduct (Yeo-Teh and Tang 2021; Anderson et al. 2021; Shimray 2022; Peterson et al. 2022).

Plagiarism is considered a serious offense in academia and amounts to dishonesty and a breach of ethics. A Spanish study examined plagiarism in medical theses by authors who had yet to publish in a scientific journal and found plagiarism in 37.3% of introduction sections (Saldaña-Gastulo et al. 2010). This study suggests that the root causes of plagiarism are at the academic level, before publishing. Another study observed that having taken courses on medical ethics was significantly associated with a negative attitude toward plagiarism (Alhadlaq et al. 2020).

Methods to prevent plagiarism include the proper citation of sources, avoiding direct quotation of large parts of copied text without explicit permission from publishers, use of quotations as needed but not excessively, as well as restating the text of interest in one's own words with proper citing of the referenced materials (Wiwanitkit 2013; Dhammi and Ul Haq 2016). It has also been suggested that continued education on the topic of plagiarism is a key tool to combat this type of misconduct (Min 2020), including lectures during freshman orientation, continued education throughout undergraduate and post graduate levels, as well as official notices posted on university websites (Issrani et al. 2021). Once plagiarism is detected in published work, editors may take action by directly addressing the authors and their institutions to investigate further or to take punitive actions. Penalties for plagiarism can include disciplinary action, retraction of the published work, and even criminal prosecution (Kumar et al. 2014). Plagiarism can be addressed at the journal level by stressing their stance against it in policy statements or submission guidelines (Debnath and Cariappa 2018), with links to plagiarism detection tools as well as using these tools at at least two points in time—upon submission and right before publication to catch added similarity during the review process. Most importantly, any similarity report must be interpreted by a human.

Conclusions

Plagiarism is common in COVID-19 related publications. Our results revealed that the majority of plagiarism was happening in papers with $\leq$ 15% similarity report. Therefore, relying solely on text-matching results is not enough due to blind spots and the potential for misinterpretation.

We recommend that authors employ ethical writing standards in their work and use resources available to them to identify areas of concern in their manuscript. Journals should utilize plagiarism checking tools along with human judgment, to ensure that a truly original paper (at least devoid of verbatim plagiarism) has entered the scientific literature.

Limitations

There are some limitations to our study. First, we only examined papers indexed on SCOPUS infection journals list, therefore our results cannot be extrapolated to represent the status of all infectious disease journals or articles. Secondly, we set the Turnitin settings to examine only journals, periodicals, and publications. We did not include other repositories such as student and institution papers. Inclusion of these repositories may have increased the overall similarity reports, events of plagiarism or number of copied sentences. Lastly, we did not include the methods section of any manuscript in our analysis, similarity results considering the methods section will likely yield higher results. Future research can focus on examining all the articles of specific journals using this method, especially during times where unprecedented interest in publishing is occurring, such as during any pandemic.

Abbreviations

IQR	Interquartile Range
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
COVID-19	Coronavirus disease of 2019

Acknowledgments
Not applicable.

Author contributions
RM and EM devised the study question and idea. RM, EM, AM, BM participated in data collection, analysis and writing of the manuscript. All authors have contributed equally and fairly to this research, and meet the requirements of authorship according to ICJME. All authors have read and approve this manuscript.

Funding
No funding was received for this research.

Availability of data and materials
Available upon reasonable request to the corresponding author.

Declarations

Ethics approval and consent to participate
This study was conducted on publicly available information using a commercially available program. It is therefore exempt from review and approval.

Consent for publication
Not applicable.

Competing interests
The authors have no competing interests to declare.

Received: 21 June 2023 Accepted: 9 October 2023
Published online: 18 October 2023

References

Alhadlaq AS, Bin Dahmash A, Alshomer F (2020) Plagiarism perceptions and attitudes among medical students in Saudi Arabia. Sultan Qaboos Univ Med J [SQUMJ] 20:77. https://doi.org/10.18295/squmj.2020.20.01.011

Anderson C, Nugent K, Peterson C (2021) Academic journal retractions and the COVID-19 pandemic. J Prim Care Community Health 12:215013272110155. https://doi.org/10.1177/21501327211015592

Aronson JK (2007) Plagiarism - please don't copy. Br J Clin Pharmacol 64:403–405. https://doi.org/10.1111/j.1365-2125.2007.03042.x

Baskaran S, Agarwal A, Panner Selvam MK et al (2019) Is there plagiarism in the most influential publications in the field of andrology? Andrologia 51:e13405. https://doi.org/10.1111/and.13405

Burdine LK, de Castro Maymone MB, Vashi NA (2018) Text recycling: Self-plagiarism in scientific writing. Int J Womens Dermatol 5:134–136. https://doi.org/10.1016/j.ijwd.2018.10.002

Campos-Varela I, Ruano-Raviña A (2019) Misconduct as the main cause for retraction a descriptive study of retracted publications and their authors. Gac Sanit 33:356–360. https://doi.org/10.1016/j.gaceta.2018.01.009

Carvalho CJ, Fuller MP, Quaidoo EA et al (2021) A review of COVID-19-related publications and lag times during the first six months of the year 2020. West J Emerg Med 22:958–962. https://doi.org/10.5811/westjem.2021.3.51737

Debnath J (2016) Plagiarism: a silent epidemic in scientific writing – Reasons recognition and remedies. Med J Armed Forces India 72:164–167. https://doi.org/10.1016/j.mjafi.2016.03.010

Debnath J, Cariappa MP (2018) Wishing away Plagiarism in scientific publications! Will it work? a situational analysis of Plagiarism policy of journals in PubMed. Med J Armed Forces India 74:143–147. https://doi.org/10.1016/j.mjafi.2017.09.003

Dhammi IK, Ul Haq R (2016) What is plagiarism and how to avoid it? Indian J Orthop 50:581–583. https://doi.org/10.4103/0019-5413.193485

Dinis-Oliveira RJ (2020) COVID-19 research: pandemic versus "paperdemic", integrity, values and risks of the "speed science." Forensic Sci Res 5:174–187. https://doi.org/10.1080/20961790.2020.1767754

Elango B (2021) Retracted articles in the biomedical literature from Indian authors. Scientometrics 126:3965–3981. https://doi.org/10.1007/s11192-021-03895-1

El-Tahan M (2019) Can the similarity index predict the causes of retractions in high-impact anesthesia journals? A Bibliometr Anal Saudi J Anaesth 13:2. https://doi.org/10.4103/sja.SJA_709_18

He F, Deng Y, Li W (2020) Coronavirus disease 2019: What we know? J Med Virol 92:719–725. https://doi.org/10.1002/jmv.25766

Higgins JR, Lin F-C, Evans JP (2016) Plagiarism in submitted manuscripts: incidence, characteristics and optimization of screening—case study in a major specialty medical journal. Res Integr Peer Rev 1:13. https://doi.org/10.1186/s41073-016-0021-8

Husain FM, Al-Shaibani GKS, Mahfoodh OHA (2017) Perceptions of and attitudes toward plagiarism and factors contributing to plagiarism: a review of studies. J Acad Ethics 15:167–195. https://doi.org/10.1007/s10805-017-9274-1

Issrani R, Alduraywish A, Prabhu N et al (2021) Knowledge and attitude of Saudi students towards Plagiarism—a cross-sectional survey study. Int J Environ Res Public Health 18:12303. https://doi.org/10.3390/ijerph182312303

Kumar PM, Priya NS, Musalaiah S, Nagasree M (2014) Knowing and avoiding plagiarism during scientific writing. Ann Med Health Sci Res 4:S193-198. https://doi.org/10.4103/2141-9248.141957

Masic I (2014) Plagiarism in scientific research and publications and how to prevent it. Materia Socio Medica 26:141. https://doi.org/10.5455/msm.2014.26.141-146

Maxwell A, Curtis GJ, Vardanega L (2008) Does culture influence understanding and perceived seriousness of plagiarism? Int J Educ Integr 4:25–40. https://doi.org/10.1348/014466604X23464

Melville H (1850) The Literary World. Osgood & Company, Hawthorne and His Mosses by A Virginian Spending July in Vermont

Meo S, Talha M (2019) Turnitin: is it a text matching or plagiarism detection tool? Saudi J Anaesth 13:48. https://doi.org/10.4103/sja.SJA_772_18

Min S-K (2020) Plagiarism in medical scientific research: can continuing education and alarming prevent this misconduct? Vasc Specialist Int 36:53–56. https://doi.org/10.5758/vsi.203621

Peterson CJ, Alexander R, Nugent K (2022) COVID-19 article retractions in journals indexed in PubMed. Am J Med Sci 364:127–128. https://doi.org/10.1016/j.amjms.2022.01.014

Polyanin AD, Shingareva IK (2022) The Similarity Index of Scientific Publications with Mathematical Equations and Formulas. Publ Res Q 38:180–188. https://doi.org/10.1007/s12109-022-09869-2

Rohwer A, Wager E, Young T, Garner P (2018) Plagiarism in research: a survey of African medical journals. BMJ Open 8:e024777. https://doi.org/10.1136/bmjopen-2018-024777

Saldaña-Gastulo JJC, Quezada-Osoria CC, Peña-Oscuvilca A, Mayta-Tristán P (2010) Alta frecuencia de plagio en tesis de medicina de una universidad pública Peruana. Rev Peru Med Exp Salud Publica 27:63–67. https://doi.org/10.1590/S1726-46342010000100011

Shimray SR (2022) Research done wrong: a comprehensive investigation of retracted publications in COVID-19. Account Res. https://doi.org/10.1080/08989621.2021.2014327

Sun Z, Errami M, Long T et al (2010) Systematic characterizations of text similarity in full text biomedical publications. PLoS ONE 5:e12704

Taylor DB (2017) JOURNAL CLUB: plagiarism in manuscripts submitted to the ajr: development of an optimal screening algorithm and management pathways. Am J Roentgenol 208:712–720. https://doi.org/10.2214/AJR.16.17208

Wiwanitkit V (2013) How to avoid plagiarism. Ann Biomed Eng 41:3. https://doi.org/10.1007/s10439-012-0683-4

Yeo-Teh NSL, Tang BL (2021) An alarming retraction rate for scientific publications on Coronavirus Disease 2019 (COVID-19). Account Res 28:47–53. https://doi.org/10.1080/08989621.2020.1782203

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.