


Similarity analysis of articles published in the Turkish Journal of Sports Medicine

Spor Hekimliği Dergisi'nde yayımlanan makalelerin benzerlik analizi

Şerife Şeyma Torgutalp¹ , Bülent Ülkar² 

¹Sports Medicine Section, Gaziler Physical Therapy and Rehabilitation, Training and Research Hospital, Ankara, Turkey

²Department of Sports Medicine, Faculty of Medicine, Ankara University, Ankara, Turkey

ABSTRACT

Objective: Turkish Journal of Sports Medicine (TJSM) has used a web-based software program to define the phrasal similarity of submitted articles since 2017. The aims of this study were; to determine the similarity scores obtained in the preliminary evaluation of the articles published in TJSM since 2017, and to evaluate the relationship between the similarity scores and the article type, article language, and publication year.

Materials and Methods: A total of 125 articles published in TJSM from 2017 to 2020 were retrospectively reviewed from the digital archive. Research articles, review articles, and case reports were included in the analysis. Similarity scores, including total similarity score and highest match scores in All Sources mode and Match Overview mode, were obtained from Similarity Reports acquired using iThenticate plagiarism checker software. Data were recorded regarding the type and language of the article, the year of publication, whether the corresponding author was from Turkey or any other country, and similarity scores.

Results: Of the 119 analyzed manuscripts, 76.5% (n=91) were research articles. The majority of the articles (95%, n=113) were submitted by authors from Turkey, and most of the articles (62.2%, n=74) were in Turkish. The median similarity score for all articles was 9.0% (Q1: 4.0 - Q3: 17.0), and the median highest matching scores were 2.0% (Q1: 1.0 - Q3: 3.0) and 3.0% (Q1: 2.0 - Q3: 6.0) for the Match Overview mode and All Sources mode, respectively. The median total similarity score and the median highest matching score from All Sources mode were significantly higher in research articles (p = 0.004 and p = 0.017, respectively) and in articles written in English (p < 0.001 for both).

Conclusion: Total similarity and highest matching scores among articles published in TJSM between 2017 and 2020 were significantly higher in research articles and articles written in English.

Keywords: *Bibliometrics, similarity index, research misconduct, editorial policies*

ÖZ

Amaç: Spor Hekimliği Dergisi (TJSM), 2017 yılından bu yana gönderilen makalelerin metinsel benzerliğini tanımlamak için web tabanlı bir yazılım programı kullanmaktadır. Bu çalışmanın amaçları; 2017 yılından itibaren TJSM'de yayınlanan makalelerin ön değerlendirilmesinde elde edilen benzerlik puanlarını belirlenmesi ve benzerlik puanları ile makale türü, makale dili ve yayın yılı arasındaki ilişkinin incelenmesidir.

Gereç ve Yöntemler: 2017-2020 yılları arasında TJSM'de yayınlanan toplam 125 makale, dijital arşiv kullanılarak geriye dönük olarak incelendi. Analize araştırma makaleleri, derlemeler ve olgu sunumları dahil edildi. Toplam benzerlik skoru ile Tüm Kaynaklar Modu ve Eşleşme Modu'ndan elde edilen en yüksek benzerlik skorlarını içeren benzerlik puanları, iThenticate intihal kontrol programı kullanılarak elde edilen Benzerlik Raporlarından elde edildi. Makalenin türü, dili, yayın yılı, sorumlu yazarının Türkiye'den mi yoksa başka bir ülkeden mi olduğu ve benzerlik puanlarına ilişkin veriler kaydedildi.

Bulgular: Analiz edilen 119 makalenin %76.5'i (n=91) araştırma makalesiydi. Makalelerin çoğu (%95, n=113) Türkiye'deki yazarlar tarafından gönderilmişti ve makalelerin çoğunun (%62.2, n=74) dili Türkçe idi. Tüm makaleler için toplam ortalama benzerlik skoru %9.0 (Q1: 4.0 - Q3: 17.0), ortalama en yüksek benzerlik skorları Eşleşme Modu'nda %2.0 (Q1: 1.0 - Q3: 3.0), Tüm Kaynaklar Modu'nda %3.0 (Q1: 2.0 - Q3: 6.0) bulundu. Ortanca toplam benzerlik skoru ve Tüm Kaynaklar Modu'ndan elde edilen en yüksek benzerlik skoru araştırma makalelerinde (sırasıyla p = 0.004 ve p = 0.017) ve İngilizce makalelerde (her ikisi için p < 0.001) daha yüksek saptandı.

Sonuç: 2017 ve 2020 yılları arasında TJSM'de yayınlanan makalelerin toplam benzerlik ve en yüksek eşleşme puanları, araştırma makaleleri ve İngilizce makaleler için anlamlı ölçüde daha yüksek saptandı.

Anahtar Sözcükler: *Bibliyometri, benzerlik indeksi, araştırma suistimali, editöryal politikalar*

INTRODUCTION

Plagiarism is defined as the appropriation of another person's ideas, processes, results, or words without giving proper reference (1). Phrasal similarity defines a percentage of matching text and has become one of the most baffling topics not only for authors but also reviewers and editors. Ho-

wever, whether high phrasal similarity is a violation of research ethics is a controversial issue, and web-based software programs have been widely used by editorial teams for the estimation of similarity in the pre-evaluation of submitted articles to prevent ethical and research misconduct.

Published / Yayın Tarihi: 01.03.2021

Correspondence / Yazışma: Şerife Şeyma Torgutalp · Spor Hekimliği Bölümü, Gaziler Fizik Tedavi ve Rehabilitasyon Eğitim ve Araştırma Hastanesi, Ankara, Turkey · seymatorgutalp@gmail.com

Cite this article as: Torgutalp SS, Ulkar B. Similarity analysis of articles published in the Turkish Journal of Sports Medicine. *Turk J Sports Med.* 2021;56(1):1-4
http://dx.doi.org/10.47447/tjism.0001

Turkish Journal of Sports Medicine (TJSM), the official journal of the Turkish Sports Medicine Association, has used a web-based software program to define the phrasal similarity of submitted articles since 2017. In order to prevent ethical misconduct, after the similarity analysis, the relevant texts are evaluated by an experienced editorial team in terms of frequency of use, place of use, and appropriate referencing in suspicious cases.

The aims of this study were; to determine the similarity scores obtained in the preliminary evaluation of the articles published in TJSM since 2017, and to evaluate the relationship between the similarity scores and the article type, article language, and publication year.

MATERIAL AND METHODS

Articles published in TJSM from 2017 to 2020 were retrospectively reviewed from the digital archive (<https://journalofsportsmedicine.org/eng/archive>) in December 2020. During this 4-year period, a total of 125 articles were published, including research paper (n=91), review article (n=16), case report (n=12), editorial (n=1), letter to the editor (n=3) and expert opinion (n=2). Research articles, review articles, and case reports were included in the study, while editorial, letter to the editor, and expert opinion articles and supplement issues were excluded.

Articles were anonymized and numbered so that the article and author identities were not revealed. Data were recorded regarding the type and language of the article, the year of publication, whether the corresponding author was from Turkey or any other country and similarity scores.

Similarity scores were recorded from Similarity Reports obtained using the plagiarism checking software (iThenticate, California, United States of America). Similarity Reports were obtained for all articles following the online submission and the bibliography was excluded from the similarity analysis. The similarity score was defined as the percentage obtained by dividing the total number of matching words found in an article by the total word count. Accordingly, the total similarity score, the highest matching scores from All Sources mode and Match Overview mode were obtained from the similarity report.

Statistical analysis

Statistical analyses were performed using the SPSS software (IBM SPSS Statistics for Mac, Armonk, New York, USA). The variables were investigated using visual (histograms and probability plots) and analytical methods (Kolmogorov-Smirnov test) to determine normal or non-normal distributions. Descriptive analyses were presented using as median, first quartile (Q1), third quartile (Q3) for continuous variables and using as frequency count and percentage for categorical variables. The Mann-Whitney U test was used to compare two independent groups. The Kruskal Wallis test was performed to compare three independent groups, and the Bonferroni-corrected Mann-Whitney U test was used to evaluate the parameters with significant differences. Statistical tests were two-sided, and a 5% type-I error level was used to infer statistical significance.

RESULTS

Of the 119 analyzed manuscripts, 76.5% (n=91) were research articles, 13.4% (n=16) were review articles, and 10.1% (n=12) were case reports. Distribution of these articles by years were 16.0% (n=19) in 2017, 21.8% (n=26) in 2018, 29.4% (n=35) in 2019, and 32.8% (n=39) in 2020, respectively. The majority of the articles (95%, n=113) were submitted by authors from Turkey, while 5% (n=6) were submitted by international authors. Most of the articles (62.2%, n=74) were in Turkish.

The median similarity score for all articles was 9.0% (Q1: 4.0 - Q3: 17.0), and the median highest matching scores were 2.0% (Q1: 1.0 - Q3: 3.0) and 3.0% (Q1: 2.0 - Q3: 6.0) for the Match Overview mode and All Sources mode, respectively. The total similarity and highest matching scores according to article types are given in Table 1. The median total similarity score and the median highest matching score from All Sources mode were significantly higher in the research articles (p=0.004 and p=0.017, respectively). It was observed that the difference in the total similarity score was more pronounced between types of articles written in Turkish (p=0.011) (Figure 1).

Table 1. The total similarity and highest matching scores according to article types.

	Research (n = 91)	Review (n = 16)	Case Report (n = 12)	P value
Total %	11.0 (5.8 - 18.0)	4.0 (2.0 - 8.5)	5.0 (1.0 - 10.0)	0.004*
Match Overview %	2.0 (1.0 - 3.0)	1.5 (1.0 - 3.0)	2.0 (1.0 - 2.0)	0.7
All Sources %	4.0 (2.0 - 6.0)	2.5 (1.0 - 4.5)	2.0 (1.0 - 3.0)	0.017*

Data are presented as median (Q1 - Q3)

* Significant difference between original and review articles and between original articles and case reports.

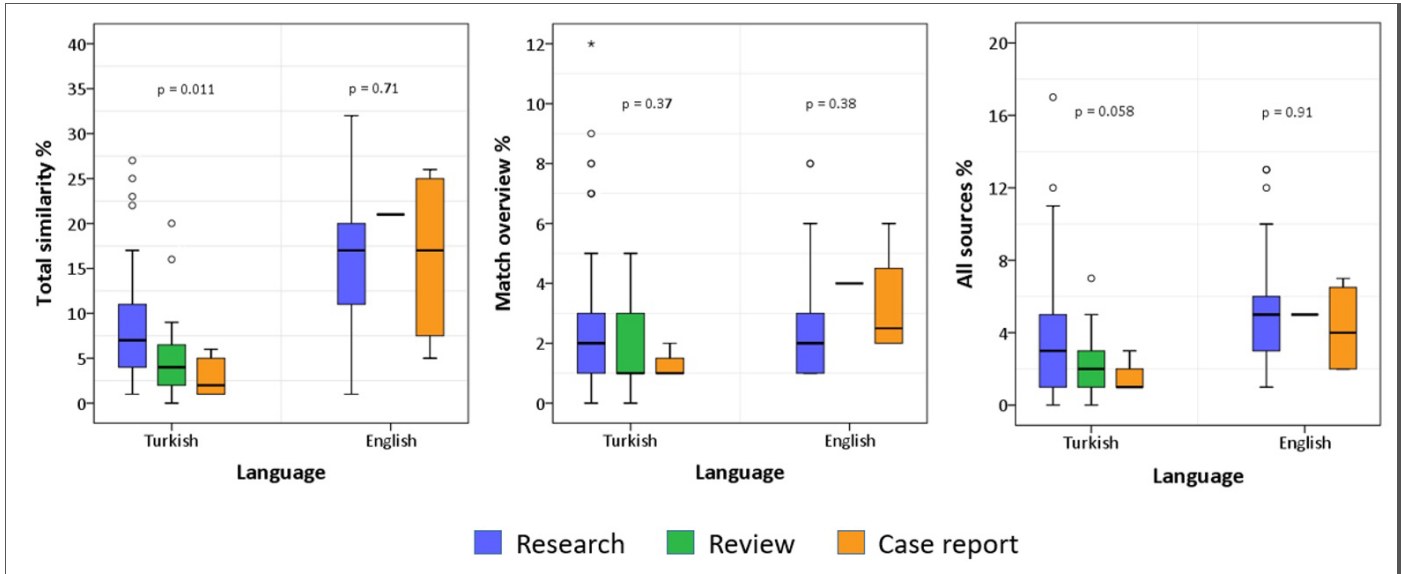


Figure 1. Comparison of total similarity score, highest matching score obtained from Match Overview mode, and highest matching score obtained from All Sources mode according to article type and language.

The total similarity and highest matching scores according to article language are presented in Table 2. The median total similarity score and the median highest matching score from All Sources mode were significantly higher in papers written in English ($p < 0.001$ for both).

Table 3. gives information about the total similarity and highest matching scores by years between 2017-2020. There were no significant differences in similarity scores between

years.

Table 3. The total similarity and highest matching scores by years.

	2017 (n = 19)	2018 (n = 26)	2019 (n = 35)	2020 (n = 39)	P value
Total %	7.0 (1.8 - 17.8)	9.0 (3.5 - 15.3)	9.0 (3.8 - 18.3)	11.0 (5.0 - 18.0)	0.5
Match %	2.0 (1.0 - 5.3)	1.5 (1.0 - 3.0)	2.0 (1.0 - 3.0)	2.0 (1.0 - 3.0)	0.7
All Sources %	3.0 (1.0 - 5.5)	3.0 (1.0 - 5.0)	3.0 (1.8 - 6.0)	4.0 (2.0 - 6.0)	0.6

Data are presented as median (Q1 - Q3).

DISCUSSION

In this study, we evaluated the results of the plagiarism detection software program used for the similarity analysis of the articles published in TJSM. We found significantly higher total similarity and the highest matching scores in research articles and articles written in English.

Similarity indices might differ according to article types, especially for original articles and image articles (2). It is thought that similar experimental methods among research articles might lead to such a result, even for articles with original content (2). Additionally, an increase in the similarity score can be observed in English articles written by authors whose mother tongue is not English, especially when information is given from other studies on a similar subject (3). Similarly, in a study, the similarity index of articles submitted from English-dominant countries was found lower than that for authors from countries that primarily speak another language (2).

Table 2. The total similarity and highest matching scores according to article language.

	Turkish(n = 74)	English(n = 45)	P value
Total %	6.0 (3.0 - 10.0)	17.5 (10.8 - 21.3)	< 0.001
Match %	1.0 (1.0 - 3.0)	2.0 (1.0 - 4.0)	0.066
All Sources %	3.0 (1.0 - 5.0)	5.0 (3.0 - 6.3)	< 0.001

Data are presented as median (Q1 - Q3).

Similarity check software programs detect ethical and research misconduct by using text-matching methods to determine the amount of textual overlap between submitted manuscripts and source publications in a wide range of databases (4). Nevertheless, the ability of these programs to explore research misconduct alone is rather limited. Four important tools are suggested for evaluating the article regarding ethical misconduct: peer reviewers, a software program, authors who recognize their research without proper citation to the original source, and shortening the content and referencing the original source that includes the full details (5). As can be seen from our results, the number of articles published in TJSM has increased over the years. In parallel with this, it would not be wrong in thinking that the number of articles submitted has increased over the years. For this purpose, using the similarity check software program in the editorial evaluation of the articles might provide a quick and easy preliminary assessment in terms of ethical and research misconduct. However, it

should be kept in mind that this alone is not a sign of ethical or research misconduct. As mentioned above, peer review and correct citation of authors to the references from which they obtain the information used in their articles are also important.

In a journal's editorial report, it was observed that accepted manuscripts had a lower similarity index than non-accepted manuscripts (2). Accordingly, the limitation of our study is that we only used similarity analysis reports obtained from already published articles, and data from unpublished manuscripts were not included. On the other hand, the evaluation of all articles published since the web-based software program started to be used in 2017 is the strength of our study.

Although institutions such as the Institute of Electrical and Electronics Engineers (IEEE) have set similarity percentage categories for providing a guide for editors when reviewing submitted manuscripts, there is no uniform and standardized approach to detect ethical misconduct in academic publishing (6). In future studies, it might be suggested to develop a more detailed and standardized algorithm based on all factors that are thought to cause ethical misconduct, such as the degree of similarity, relevant content areas, and correct citation.

In conclusion, we determined the total similarity and highest matching scores of the articles published in TJSM since 2017 and higher total similarity and highest matching scores were observed in research articles and case-reports written in English.

Conflict of Interest / Çıkar Çatışması

The authors declared no conflicts of interest with respect to authorship and/or publication of the article.

Financial Disclosure / Finansal Destek

The authors received no financial support for the research and/or publication of this article.

REFERENCES

1. Federal Research Misconduct Policy | ORI - The Office of Research Integrity; c2021 [cited 2021 Jan 22]. (Available from: <https://ori.hhs.gov/federal-research-misconduct-policy>)
2. Lee JH. Analysis of crosscheck data on two years' worth of papers submitted to Archives of Plastic Surgery. *Arch Plast Surg*. 2014;41(5):449–51.
3. Yilmaz I. Plagiarism? No, we're just borrowing better English. *Nature*. 2007;449(7163):658.
4. Park S, Yang SH, Jung E, Kim YM, Baek HS, Koo YM. Similarity analysis of Korean medical literature and its association with efforts to improve research and publication ethics. *J Korean Med Sci* 2017;32(6):887–92.
5. Hinds PS. How Many Recipes for Chocolate Cake Do We Need?" or When Does Similarity Become Self-plagiarism? *Cancer Nurs*. 2019;42(1):1–2.
6. User's Guide for the IEEE CrossCheck Portal and Prohibited Authors List Database;c2021 [cited 2021 Jan 22]. (Available from: <https://www.ieee.org/content/dam/ieee-org/ieee/web/org/pubs/IEEE%20CrossCheck%20Portal%20Guide.pdf>)