ORIGINAL ARTICLE
Editing, Writing & Publishing

JKMS

https://doi.org/10.3346/jkms.2017.32.6.887 • J Korean Med Sci 2017; 32: 887-892

Similarity Analysis of Korean Medical Literature and Its Association with Efforts to Improve Research and Publication Ethics

Soyoung Park,[1] Seung Ho Yang,[1] Eugene Jung,[1] Yeon Mi Kim,[1] Hyun Sung Baek,[2] and Young-Mo Koo[2]

[1]Asan Medical Library, University of Ulsan College of Medicine, Seoul, Korea; [2]Department of Medical Humanities and Social Sciences, University of Ulsan College of Medicine, Seoul, Korea

Received: 26 January 2017
Accepted: 4 March 2017

Address for Correspondence:
Young-Mo Koo, PhD
Department of Medical Humanities and Social Sciences, University of Ulsan College of Medicine, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Korea
E-mail: ymkoo@amc.seoul.kr

In the present study, the frequency of research misconduct in Korean medical papers was analyzed using the similarity check software iThenticate®. All Korean papers written in English that were published in 2009 and 2014 in KoreaMed Synapse were identified. In total, 23,848 papers were extracted. 4,050 original articles of them were randomly selected for similarity analysis. The average Similarity Index of the 4,050 papers decreased over time, particularly in 2013: in 2009 and 2014, it was 10.15% and 5.62%, respectively. And 357 (8.8%) had a Similarity Index of ≥ 20%. Authors considered a Similarity Index of ≥ 20% as suspected research misconduct. It was found that iThenticate® cannot functionally process citations without double quotation marks. Papers with a Similarity Index of ≥ 20% were thus individually checked for detecting such text-matching errors to accurately identify papers with suspected research misconduct. After correcting text-matching errors, 142 (3.5% of the 4,050 papers) were suspected of research misconduct. The annual frequency of these papers decreased over time, particularly in 2013: in 2009 and 2014, it was 5.2% and 1.7%, respectively. The decrease was associated with the introduction of CrossCheck by KoreaMed and the frequent use of similarity check software. The majority (81%) had Similarity Indices between 20% and 40%. The fact suggested that low Similarity index does not necessarily mean low possibility of research misconduct. It should be noted that, although iThenticate® provides a fundamental basis for detecting research misconduct, the final judgment should be made by experts.

Keywords: Plagiarism; Duplicate Publication as Topic; Editorial Policies; Scientific Misconduct; Periodicals as Topic; Software

INTRODUCTION

To encourage their registration in international bibliographic and citation databases, Korean medical journals are increasingly being equipped with international publishing standards, including online paper submission systems, publication of the papers in English, conversion to Open Access, and digital object identifier (DOI) registration. These elements increase the access of researchers abroad to Korean medical papers. These changes to the publishing environment have led to concerns about the compliance of Korean medical papers in terms of research and publishing ethics: in 2007, a survey of 165 academic journal member organizations in the Korean Association of Medical Journal Editors (KAMJE) showed that the editors considered research ethics to be the second most important issue (37.3%) next to the Science Citation Index (SCI) journal registration procedure (60.8%) (1).

Ethical misconduct in research can largely be divided into research ethics and publishing ethics. Research ethics refer to data fabrication and falsification and plagiarism, while publishing ethics refer to issues that can arise during the process of research result publication, including authorship, conflict of interest declarations, and duplicate publication (2).

Plagiarism is the use of other people's ideas or copyrighted works without disclosing the source. It has been defined in various ways by research ethics-related domestic and foreign institutions (2-5). Source description errors are generally errors where the description of the source is either omitted or inappropriate (6). Another source description error is self-plagiarism, where one uses one's own published works without disclosing the original publication. However, the term "text recycling" is more widely used than "self-plagiarism," and whether it is an infringement of research ethics should be determined on the basis of the frequency of use and the circumstances of the author(s) (2).

The rates of duplicate publication in the Korean medical articles in KoreaMed were 5.9%, 6.0%, and 7.2%, respectively in 2004, 2005, and 2006. Since then they declined steadily: in 2007, 2008, and 2009, the rates were 4.5%, 2.8%, and 1.2%, respectively. In

© 2017 The Korean Academy of Medical Sciences.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

pISSN 1011-8934
eISSN 1598-6357

terms of the duplicate publication patterns during 2004 and 2009, copies were most common (53.4%), followed by salami slicing (27.8%) and aggregation (18.8%) (7). KAMJE launched the nationwide ethics campaign starting in 2006, which led to a general increase in domestic awareness regarding medical research and publishing ethics. Moreover, in 2008, Google Scholar® began to cover the data in the KoreaMed database, which increased the ease of duplicate publication detection. The decline of duplicate publication could be attributed to these activities.

Text-matching software that efficiently detects research misconduct has been developed. This software detects overlapped texts numerically and is often used particularly by journal editors to determine whether submitted papers bear texts that are already present in published papers or materials on the internet. One of such software is CrossCheck, which is Crossref Similarity Check powered by iThenticate® and as of May 2015 was being used by over 500 Similarity Check Members including the most influential publishers (8). As of June 2015, 98 of the 215 medical journals registered in KoreaMed were using CrossCheck to detect ethical misconduct (9,10).

In 2014, Lee used CrossCheck to determine the similarity of papers that were submitted to *Archives of Plastic Surgery* in 2012 and 2013 (11). The Similarity Index was also determined for the accepted and rejected papers, papers from English- and non-English-language countries, different types of publications, and clinical and experimental papers. The papers that were submitted in 2013 had a lower Similarity Index than those that were submitted in 2012. The accepted papers had a lower Similarity Index than the rejected papers, the papers from English-language countries had a lower Similarity Index than those from non-English-language countries, and original articles and image articles had higher Similarity Indices than the other article categories (case reports, review articles, and letters), however, the difference was not statistically significant. Moreover, of the original articles, the experimental papers had a higher Similarity Index than the clinical articles; in large part, this was because of the use of similar experimental methods. Lee (11) recommended that papers with Similarity Indices higher than 40% should be sent back to the authors. However, he also recommended that experimental papers should be judged more flexibly.

A similar study was published by Zhang in 2010 (12). CrossCheck was used to assess 662 papers that were submitted between October 2008 and May 2009 to *Journal of Zhejiang University – Science (A & B)*, a Chinese academic journal. Zhang found that 22.8% (151 papers) contained unreasonable copying or self-plagiarism. Moreover, of these 151 papers, 39 (25.8%) were seriously suspected of plagiarism and copyright infringement. In 2012, Zhang and her co-author Jia (13) published the results of a survey of editors all over the world. They found that 42% had used CrossCheck. These editors also reported that this tool was very useful for screening for research misconduct. Zhang and Jia also observed that, while editors have clear editorial standards regarding plagiarism, there were small variations between different disciplines and countries.

In the present study, original medical research papers that were published between 2009 and 2014 in Korean medical journals indexed in KoreaMed were analyzed using the similarity check software iThenticate®. The Similarity Index of each paper was determined. Those with a Similarity Index exceeding 20% were suspected of research misconduct. The relationship between the frequency of the verified suspicious papers and the prevalence of similarity check software was assessed.

MATERIALS AND METHODS

All English-language full-text papers that were published between January 2009 and December 2014 in Korean medical journals indexed in KoreaMed Synapse were extracted. The year range was chosen so that the change in similarity rates before and after the introduction of CrossCheck at KAMJE in June 2011 could be assessed. Due to the large number of papers that were extracted, approximately 30% were randomly selected for further analysis. For this, equivalent numbers of papers from each journal per year were selected using a random number generation program (Microsoft C# Random function). In addition, reviews, case reports, letters, and editorials were excluded so that only original articles were included.

All papers were then uploaded on iThenticate®. The iThenticate® options were set to "Exclude quotes," "Exclude bibliography," and "Exclude section (Abstract, Methods and Materials)" before the papers were uploaded (14). "Exclude word matches that are less than 20 words" was also added to the options. The text length of 20 words was arbitrarily applied because there were no academically established standards for the number of words that best detects plagiarism (6,15). iThenticate® provides an overall similarity index for each submitted paper before publication by matching with already published sources. Because this similarity check was conducted even after papers had been published, the matching sources included even the papers which normally quoted them as well as the uploaded same papers. So the authors personally reviewed the Similarity Reports and excluded the same ones as the uploaded papers and also the published papers after the year they had been published.

iThenticate® functionally excludes the citations that have double quotation marks from the similarity calculation through "Exclude quotes." However, it considers citations even with other normal quotation marks than double quotation marks such as reference numbers as textual overlap. Therefore, to accurately detect papers with suspected research misconduct, we had to personally review the papers with textual overlap marked with normal quotation marks and adjust their Similarity Indices. We

https://doi.org/10.3346/jkms.2017.32.6.887

applied these adjustments to all papers that had a Similarity Index higher than 20%. The papers that still had Similarity Indices higher than 20% after adjustment were finally regarded as papers with potential research misconduct.

RESULTS

Fig. 1 shows the flow chart of paper extraction, random paper selection, adjustment of Similarity Index through excluding authors' own papers and published papers after them, and correction of text-matching errors. Thus, 23,848 papers that were published between January 2009 and December 2014 were extracted in December 2014 from 145 English-language journals that provided full-text files in KoreaMed Synapse. Of these, 7,802 papers (32.7%) were randomly selected. Of those, 4,050 original articles were uploaded on iThenticate® to check for similarity with other publications. 357 of these 4,050 papers had Similarity Indices that exceeded 20% (Fig. 2).

The papers with Similarity Indices higher than 20% were accounted for 8.8% of all papers (Fig. 2). Indeed, when the papers were categorized according to whether their Similarity Index was 0%–20%, 20%–40%, 40%–60%, or > 60%, there was a marked increase over time in the papers with Similarity Indices less than

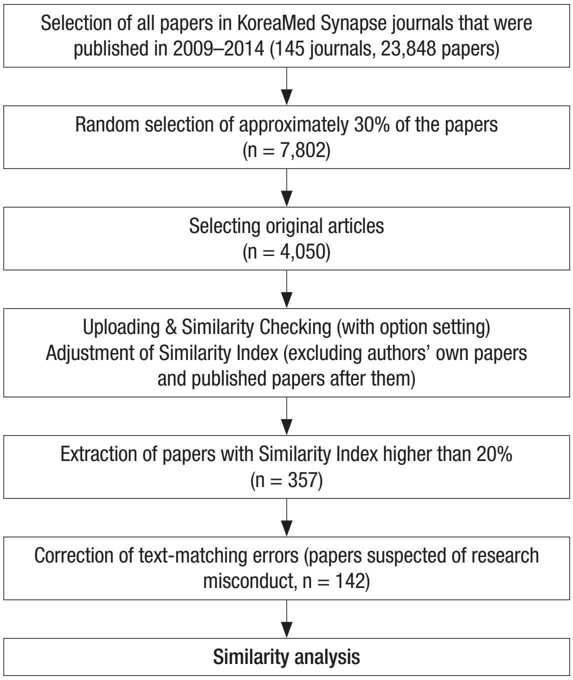
Selection of all papers in KoreaMed Synapse journals that were published in 2009–2014 (145 journals, 23,848 papers)
Random selection of approximately 30% of the papers (n = 7,802)
Selecting original articles (n = 4,050)
Uploading & Similarity Checking (with option setting) Adjustment of Similarity Index (excluding authors' own papers and published papers after them)
Extraction of papers with Similarity Index higher than 20% (n = 357)
Correction of text-matching errors (papers suspected of research misconduct, n = 142)
Similarity analysis

Fig. 1. Flow chart showing paper extraction, random paper selection, exclusion of authors' own papers, and published papers after them, correction of text-matching errors, and identification of papers suspected of research misconduct.

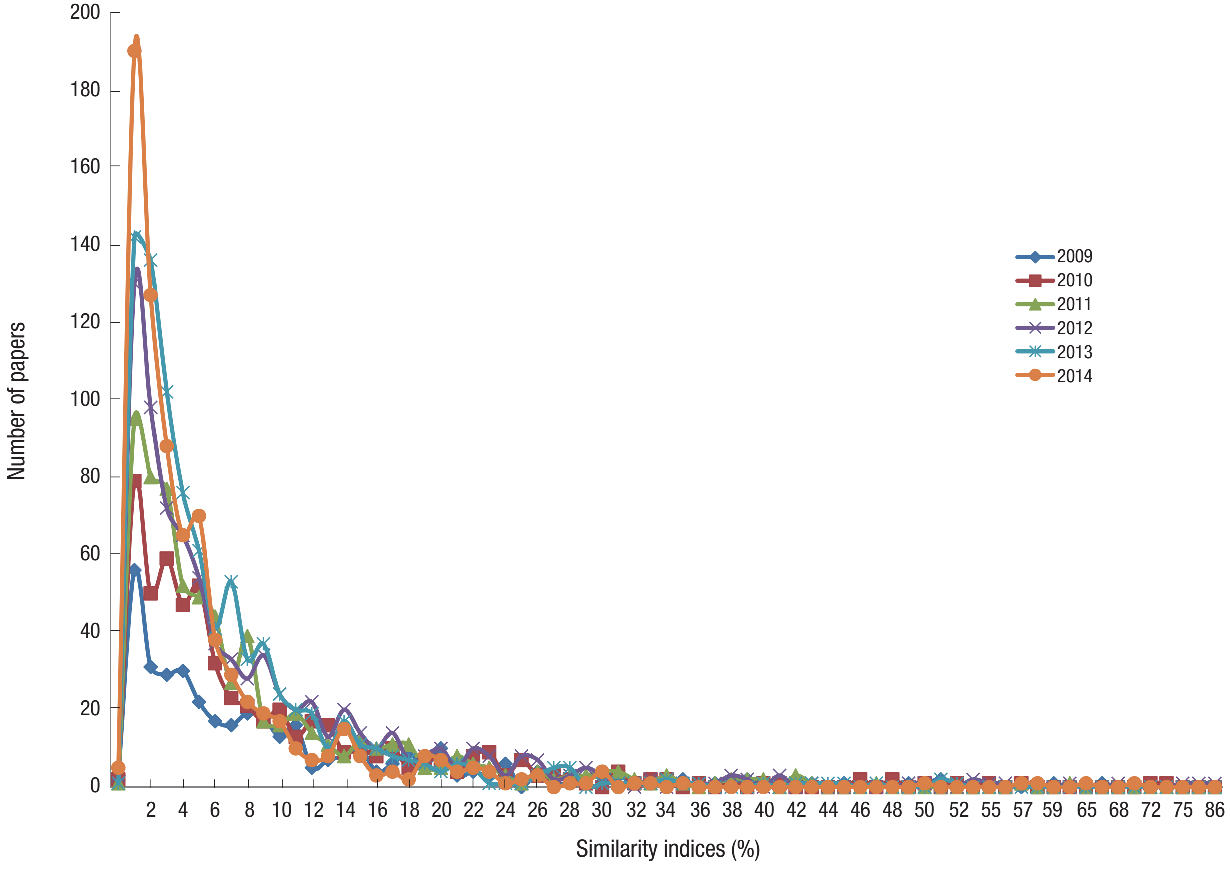

Fig. 2. Distribution of Similarity Indices in 4,050 Korean medical papers in 2009–2014.

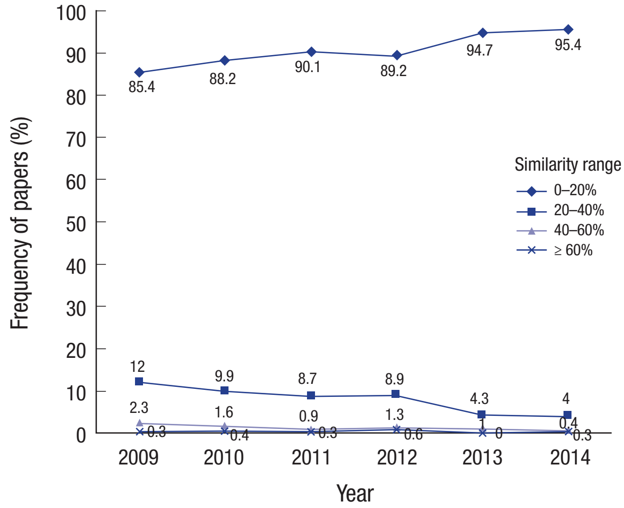

Fig. 3. Change over time (2009–2014) in the frequency of Korean medical papers that fell into specific Similarity Index categories.

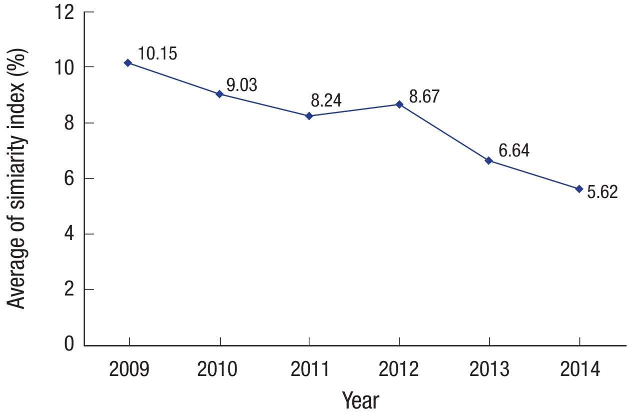

Fig. 4. Average Similarity Index per year (2009–2014) of Korean medical papers.

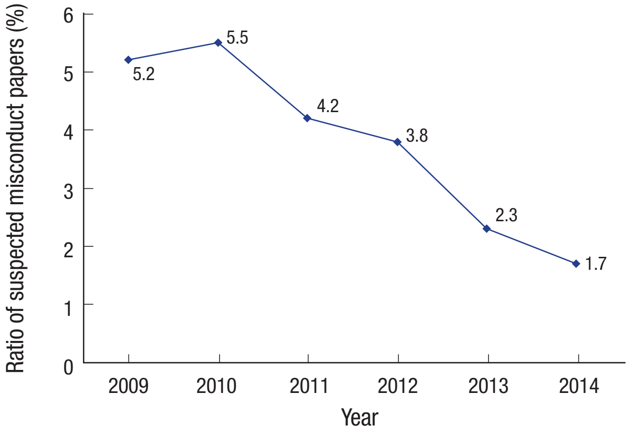

Fig. 5. Frequency over time (2009–2014) of papers with suspected research misconduct.

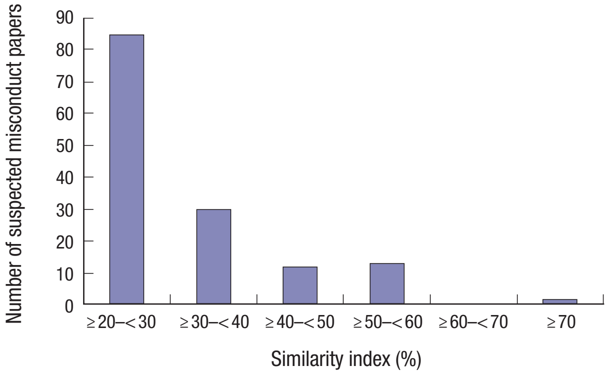

Fig. 6. Distribution of the Similarity Indices of the 142 papers with suspected research misconduct.

20%, while the number of papers with Similarity Indices 20%–40% decreased over time, especially in 2013 (Fig. 3).

Similarly, the average Similarity Index decreased as time progressed: in 2009, 2010, 2011, 2012, 2013, and 2014, it was 10.15%, 9.03%, 8.24%, 8.67%, 6.64%, and 5.62%, respectively. A particularly sharp decrease in average Similarity Index was observed in 2013 (Fig. 4).

After correction of text-matching errors of iThenticate®, 142 papers (3.5% of the 4,050 papers) still had Similarity Indices higher than 20% and were regarded as papers with suspected research misconduct. Analysis of these 142 papers showed that their frequency decreased over time: in 2009, 2010, 2011, 2012, 2013, and 2014, the frequencies were 5.2%, 5.5%, 4.2%, 3.8%, 2.3%, and 1.7%, respectively. A particularly sharp decrease was observed in 2013 (Fig. 5).

Of the papers that were suspected of research misconduct, the majority had a Similarity Index in the range of 20%–40% (115 cases, 81%). Twenty-five papers (17.6%) had a Similarity Index of 40%–60%, and 2 papers (1.4%) had a Similarity Index of 60%–80%. None of the suspected papers had a Similarity Index higher than 80% (Fig. 6).

DISCUSSION

Our initial similarity check of original Korean medical papers that were published between 2009 and 2014 showed that 8.8% of the papers had a Similarity Index that exceeded 20%. However, over time, the frequency of papers with a Similarity Index that exceeded 20%, and the annual average Similarity Index gradually decreased. After correcting for text-matching errors of iThenticate®, 142 papers (3.5% of the total) were suspected of research misconduct. The frequencies of these papers also decreased gradually. Moreover, all of these variables showed a particularly sharp decrease in 2013. This suggests that overall research misconduct decreased since 2012. One of the main reasons is probably the prevalence of similarity check software and the use of CrossCheck by KoreaMed Synapse journals since 2012.

Our analysis showed that the vast majority (81%) of papers that were suspected of research misconduct predominantly had Similarity Indices of 20%–40%. This shows that, even in papers with low Similarity Indices, research misconduct cannot be ruled out.

In this study, Similarity Checking was based on papers that had already been published, not on unpublished manuscripts. To reduce this limitation, the same papers as the uploaded papers and the published papers after the year they had been published were excluded. However, the Similarity Index may not exactly match the results conducted prior to publication. One of this study strength was that several options were set in iThenticate® to improve the accuracy of the similarity check. The other was that we personally corrected text-matching errors of iThenticate® that misreads citations with other normal marks than double quotation marks as textual overlap.

Similarity indices can vary according to article types (11). However, we considered original articles as the most representative type of articles for our similarity analysis. Another possible study limitation was that a Similarity Index of ≥ 20% was set arbitrarily to detect papers with suspected research misconduct in consideration of the practical range of personal review. Consequently, the possibility of research misconduct in the papers with a Similarity Index less than 20% was ruled out. It should be noted that the 142 suspicious papers were not subjected to a full review to verify research misconduct. Thus, it cannot be stated conclusively that these papers were actually committing research misconduct. A separate study in which experts fully review all papers that are found by similarity checking to be suspicious of research misconduct would indicate the accurate rates of research misconduct. Furthermore, additional data regarding more specific category of plagiarism or duplication for each suspicious article and the sections of manuscripts with excessive textual overlaps can be obtained through the separate review.

iThenticate® technology essentially detects research misconduct by using text-matching methods to identify the degree of textual overlap expressed as Similarity Index between submitted papers and source publications in a vast array of databases. However, the ability of this software to detect research misconduct is obviously limited. It has a fatal functional weakness, that is, it is unable to recognize normal quotation marks other than double quotation marks. In addition, even with double quotation marks, more precise judgments are required for papers with broader text matching over several paragraphs.

Institutions such as Institute of Electrical and Electronics Engineers (IEEE) have set the following Similarity Index alert levels to provide a guide for editors during the review process: papers with Similarity Indices below 10% are not likely to be issues (disregard), papers with Similarity Indices of 10%–50% may have possible issues (review briefly), and those with Similarity Indices above 50% probably have probable issue (review carefully) (16). However, the judgment regarding a paper's Similarity Index depends on the editors or experts (15,17). A more detailed and standardized consensus among expert groups can minimize the errors in Similarity Indices that can arise from arbitrary use of similarity check software and establish standardized research misconduct detection methodology as well.

ACKNOWLEDGMENT

All authors would like to thank Professor Sun Huh (Department of Parasitology, College of Medicine Hallym University, Chuncheon, Korea) for his critical comments on the draft of this paper.

DISCLOSURE

The authors have no potential conflicts of interest to disclose.

AUTHOR CONTRIBUTION

Conceptualization: Park S, Koo YM. Data curation: Yang SH, Jung E, Kim YM, Baek HS. Investigation: Park S, Yang SH, Jung E, Kim YM, Baek HS, Koo YM. Writing - review & editing: Park S, Jung E, Koo YM.

ORCID

Soyoung Park http://orcid.org/0000-0002-7130-8382
Seung Ho Yang http://orcid.org/0000-0002-3479-3497
Eugene Jung http://orcid.org/0000-0001-7256-3548
Yeon Mi Kim http://orcid.org/0000-0002-3067-3957
Hyun Sung Baek http://orcid.org/0000-0003-1743-9020
Young-Mo Koo http://orcid.org/0000-0002-3736-5797

REFERENCES

1. Korean Association of Medical Journal Editors. Ten-year History of the Korean Association of Medical Journal Editors. Seoul, Korean Association of Medical Journal Editors, 2008.
2. Korean Association of Medical Journal Editors. Good Publication Practice Guidelines for Medical Journals. 2nd ed. Seoul, Korean Association of Medical Journal Ediotrs, 2013.
3. Korean Association of Academic Societies. Guideline for Research Ethics. Seoul, Korean Association of Academic Societies, 2010.
4. Hwang ES, Cho EH, Kim YM, Park GB, Son HC, Yoon TW, Lim JM. Manual for research and publication ethics in science and engineering [Internet]. Available at http://www.cre.or.kr/board/?board=textbook&no=1384677 [accessed on 10 June 2015].
5. Seoul National University (KR). Seoul National University's guideline for research ethics [Internet]. Available at http://www.snu.ac.kr/research/images/down/research_08.pdf [accessed on 10 June 2015].
6. Lee IJ. The understanding of research ethics for a desirable academic writing. *Endocrinol Metab* 2011; 26: 12-24.

7. Kim SY, Bae CW, Hahm CK, Cho HM. Duplicate publication rate decline in Korean medical journals. *J Korean Med Sci* 2014; 29: 172-5.
8. Crossref (US). CrossCheck members [Internet]. Available at http://www.crossref.org/crosscheck_members.html [accessed on 4 June 2015].
9. Korean Association of Medical Journal Editors. CrossCheck journals [Internet]. Available at http://www.kamje.or.kr/intro.php?body=crosscheck [accessed on 1 November 2014].
10. Kwon OH. What is CrossCheck? [Internet]. Available at http://www.kamje.or.kr/workshop/2012/0216/11.pdf [accessed on 28 November 2014].
11. Lee JH. Analysis of CrossCheck data on two years' worth of papers submitted to archives of plastic surgery. *Arch Plast Surg* 2014; 41: 449-51.
12. Zhang Y. CrossCheck: an effective tool for detecting plagiarism. *Learn Publ* 2010; 23: 9-14.
13. Zhang Y, Jia X. A survey on the use of CrossCheck for detecting plagiarism in journal articles. *Learn Publ* 2012; 25: 292-307.
14. Huh S. What happened when CrossCheck was not used for a month in Journal of Neurogastroenterology and Motility? *J Neurogastroenterol Motil* 2014; 20: 417-8.
15. Wager L. How should editors respond to plagiarism? COPE discussion paper [Internet]. Available at http://publicationethics.org/files/COPE_plagiarism_disc%20doc_26%20Apr%2011.pdf [accessed on 1 June 2015].
16. Institute of Electrical and Electronics Engineers (US). User's guide for the IEEE CrossCheck portal and prohibited authors list databases [Internet]. Available at https://www.ieee.org/publications_standards/publications/rights/crosscheck_portal_users_guide.pdf [accessed on 1 June 2015].
17. Bae CW, Kim SY, Huh S, Hahm CK. Sample Cases of Duplicate Publication. Seoul, Korean Association of Medical Journal Editors, 2011.