BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 2 2008, pages 243–249
doi:10.1093/bioinformatics/btm574

Data and text mining

Déjà vu—A study of duplicate citations in Medline

Mounir Errami[1], Justin M. Hicks[1], Wayne Fisher[1], David Trusty[1], Jonathan D. Wren[2], Tara C. Long[1] and Harold R. Garner[1,*]

[1]UT Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas TX 75390-9185 and [2]Arthritis and Immunology Research Program, Oklahoma Medical Research Foundation, 825 N.E. 13th Street, Oklahoma City OK 73104

Received on September 20, 2007; revised on November 14, 2007; accepted on November 15, 2007

Advance Access publication December 1, 2007

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Duplicate publication impacts the quality of the scientific corpus, has been difficult to detect, and studies this far have been limited in scope and size. Using text similarity searches, we were able to identify signatures of duplicate citations among a body of abstracts.

Results: A sample of 62 213 Medline citations was examined and a database of manually verified duplicate citations was created to study author publication behavior. We found that 0.04% of the citations with no shared authors were highly similar and are thus potential cases of plagiarism. 1.35% with shared authors were sufficiently similar to be considered a duplicate. Extrapolating, this would correspond to 3500 and 117 500 duplicate citations in total, respectively.

Availability: eTBLAST, an automated citation matching tool, and Déjà vu, the duplicate citation database, are freely available at http://invention.swmed.edu/and http://spore.swmed.edu/dejavu

Contact: harold.garner@utsouthwestern.edu

1 INTRODUCTION

Scientific misconduct comes in many forms. The Office of Science and Technology Policy defines research misconduct as 'fabrication, falsification or plagiarism in proposing, performing or reviewing research, or in reporting research results'. In the scientific research community, plagiarism and repeated publication of the same data are considered unacceptable practices, can result in misunderstanding, and are a waste of time and energy for authors, reviewers and readers. These practices seek to give a false impression of one's scientific productivity (Lehmann *et al*., 2006). The National Library of Medicine (NLM) defines a duplicate publication as one that 'substantially duplicates another article without acknowledgement'. (http://www.nlm.nih.gov/pubs/factsheets/errata.html) As of July 2006, NLM annotated 607 records in Medline with the publication type 'Duplicate Publication'. We manually inspected these 607 records and found that 409 included abstracts, enabling us to classify 171 (42%) as true duplicate publications. The remainder were errata, updates or comments. While some duplications may be justified, arguably to promote wider dissemination or to provide important updates to clinical trials, surreptitious duplications that are covert and do not properly acknowledge the original work are unethical. In contrast to the NLM annotations, in a 2005 study of 3234 respondents to an anonymous survey directed at NIH funded researchers, Martinson *et al*. compiled a table of frequencies of violations admitted to by those scientists (Martinson *et al*., 2005). In that survey 1.4% of the respondents admitted to plagiarism and 4.7% to multiple publications of the same data. Although the finer data interpretation nuances of the Martinson study are in debate, the magnitude of the problem is striking (Grinnell, 2005). There is a wide discrepancy between the rate of annotated duplicates found within the NLM tagged set (0.0011% = 171 duplicates/16 000 000 Medline records) and what one would anticipate based on the survey by Martinson *et al*. Other studies in which the literature of a narrow biomedical discipline was studied in depth have arrived at similar conclusions (Bailey, 2002; Barnard and Overbeke, 1993; Blancett *et al*., 1995; Bloemenkamp *et al*., 1999; Chennagiri *et al*., 2004; Durani, 2006; Gotzsche, 1989; Kostoff *et al*., 2006; Mojon-Azzi *et al*., 2004; Roig, 2005; Rosenthal *et al*., 2003; Schein and Paladugu, 2001). For example, Schein and Paladugu noted that, 'Almost 1 in every 6 original articles published in leading surgical journals represents some form of redundancy' (Schein and Paladugu, 2001). And in a comprehensive study of the full text of systematic reviews within the perioperative medicine field, von Elm found that 8.3% of papers cited within the reviews were duplicates and that the duplicates were cited as often as the original article and in journals with similar impact factors as the genuine original (Yank and Barnes, 2003).

Here we report a statistically significant sampling of Medline citations using a text similarity algorithm to compile a set of highly similar citations. The results are recorded in a web accessible database so that we and others may study individual cases or the collection as a whole to gauge the overall frequency, trends and characteristics of duplicate scientific publication. Because Medline citation records do not contain full text articles, detection of duplicate text is limited to titles and abstracts, which are only small fractions of the complete documents. Identifying potential duplicate publications, however, is frequently possible using only Medline citations since

*To whom correspondence should be addressed.

abstracts and titles conceptually distill the contents of the full text and it is difficult to describe that essence using significantly different terms. Duplicate citations identified using Medline citations and our text-similarity search tool eTBLAST are publicly available in our web-accessible database Déjà vu, at http://spore.swmed.edu/dejavu. However, due to copyright issues, we cannot post the full-text of most duplicate articles, which users may want to consult to judge how representative the citation text is of the manuscript as a whole.

2 METHODS

2.1 Identification of highly similar citations

The text similarity-based information retrieval and search engine eTBLAST was originally developed as a web tool to enable users to input the partial or full text of a document as the query for comparisons to electronic literature databases such as Medline (Lewis *et al*., 2006). It is freely available at http://invention.swmed.edu/etblast/index.shtml. For this study, eTBLAST was used to find the most similar citations using a selection of Medline titles and abstracts as the queries. Each query is formed by a title and an abstract, from which eTBLAST removes the stopwords. eTBLAST computes a quantitative similarity score between each title and abstract query and every Medline citation and returns a list of the most similar citations ranked by their similarity score. The citation with the highest similarity score is always the self-identity and is referred to as Rank 1. The most similar non-identical citations are listed in order and referred to as Rank 2, Rank 3, etc. eTBLAST uses the ratio of the similarity score of a citation to the query score to classify two citations as 'highly similar' or not. The score calculated by eTBLAST reflects the similarity between the documents, i.e. the higher the score, the more similar the citations and this score has no upper bound.

2.2 Training and experimental data sets

Four non-overlapping sets of queries were prepared for submission to the similarity engine and comparison to all Medline records: (1) a benchmarking dataset from the 171 known and visually-verified Medline duplicate pairs, (2) a set of 5313 randomly-selected Medline citations, all of which included both a title and abstract (only about 50% of Medline citations contain an abstract), (3) twelve sets of 5000 Medline records, 60 000 total, that included both titles and abstracts, selected randomly from each of the last 12 years and (4) a set of 5465 Medline records that also have full text available in PubMed Central (PMC). The random selection was performed with a random number generator to obtain PMIDs or PMCIDs which are unique integers identifying articles within PubMed and PubMed Central.

2.3 Manual classification of highly similar citations

Each highly similar duplicate pair identified by eTBLAST was manually verified by at least two authors of this study, who read each duplicate citation pair and classified the putative duplicates into one of the categories Duplicate/Different Authors (DA), Duplicate/Same Authors (SA), Duplicate/Update/Same Journal (SJ), Duplicate/Update/Different Journal (DJ), Duplicate Medline Issue (MI), Duplicate/Other, errata, false positive or no abstract. These categories, whose definitions are detailed within the descriptions of the database on the web, were developed to support the full range of highly similar documents that occur in Medline, and they enable us to distinguish among duplicates reflecting different behaviors by authors. In particular we sought to distinguish between duplicates resulting from appropriate versus inappropriate behaviors. For example, the presence of shared authors in a pair of duplicates was used to distinguish between cases of potential plagiarism and multiple publication of the same study by the same authors. Similarly, it was important to distinguish those duplicate citations representing updates to clinical trials or survey type research where duplication is not considered inappropriate. We further sub-categorize these duplications according to when and in which journals they appear as indication of a 'covert duplicate publication', (Yank and Barnes, 2003) which may be instances of authors publishing substantially the same results in multiple journals with the intent to get more citations from a single piece of work. Citations categorized as false positive frequently describe different sets of experiments within a research line that are distinct but reported by the same authors using similar phraseology. Such pairs are verifiable only by manual inspection. Errata, which may or may not be tagged as such in Medline, are generally very similar to the initial query (Rank 2/Rank 1 score ratio ~1), and often involve only a typographical correction.

In the course of this study we manually read and classified nearly 5000 citations and approximately 250 of their associated electronically available full text articles that had been categorized as highly similar by eTBLAST.

2.4 The Déjà vu results database

The Déjà vu interface was designed using python (http://www.python.org) and the Django web framework (http://djangoproject.com). Data is stored in an embedded SQLite Database (http://www.sqlite.org). On the Déjà vu website, users can

(1) browse Déjà vu entries with no specific search method. Each entry links to the scientific citation along with full text whenever freely available;

(2) search Déjà vu content by authors, title word, abstract word, year and comment word;

(3) view Déjà vu results in a particular category or identified by a particular 'discovery method' (eTBLAST or manual);

(4) provide comments in order to contest a record or submit a potential duplication that will be reviewed by authors of this manuscript.

3 IMPLEMENTATION

3.1 Duplicate pair identification using quantitative text similarity measures

Results from eTBLAST searches using the training dataset of known duplicate citations were compared with similar searches using the randomly-selected, and thus mostly non-duplicate, citations to identify a signature that could differentiate duplicates from non-duplicates. Histograms of the frequency distributions of the Rank 1 and Rank 2 scores for the 171 known duplicates are plotted together in Figure 1B, demonstrating a substantial overlap in the scores of the most similar (Rank 1, self) and second most similar (Rank 2) scores. In contrast, histograms of the Rank 1 and Rank 2 scores for the set of 5313 randomly-selected Medline citations show much less overlap (Fig. 1A). This figure suggests that the Rank 2/Rank 1 score ratio may distinguish duplicate and non-duplicate pairs. Further, a high Rank 2 score indicates that the Rank 2 citation is more similar to the original query than if the Rank 2 score is low, and so may also discriminate between duplicate and non-duplicate pairs.

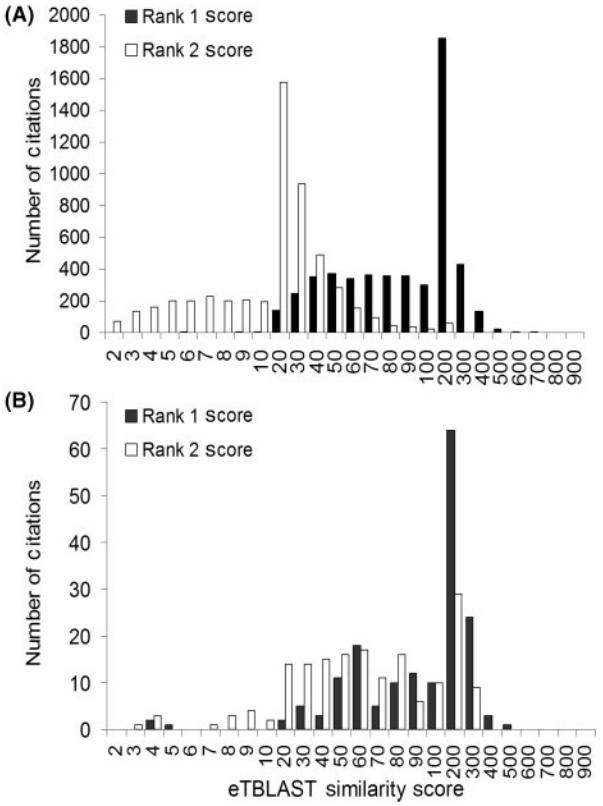

Fig. 1. (**A**) eTBLAST similarity scores for the two highest-ranked hits found by eTBLAST as a result of searching 5313 random Medline citations. (**B**) Overlapping distribution of the top ranked (identity) and second ranked (most similar non-identity) scores found by eTBLAST searches of the 171 citations annotated by NLM in Medline with a Publication Type 'Duplicate Publication', after removing errata.

Using these two observations together we plotted Rank 2 score as a function of Rank 2/Rank 1 score ratio to determine what thresholds would separate duplicate from non-duplicate pairs. The data points for the 5313 randomly-selected Medline citations, Figure 2, lay mostly at low Rank 2 scores and low Rank 2/Rank 1 score ratios. As shown in Figure 3, a much higher proportion of the 171 manually-verified, NLM annotated duplicate publication data points lay at high Rank 2 score and high Rank 2/Rank 1 score ratios. Therefore we looked for quantitative thresholds to separate duplicate and non-duplicate citations.

A threshold for Rank 2 score was established first. Bins were created along the x axis, in each bin a normal distribution of Rank 2 scores was assumed, and the mean and SD of the scores in each bin were calculated. A series of Z-score curves was constructed by a least squares fit of the mean plus SD points in the bins to a parabolic function. We chose to set the threshold using the curve determined using a conservative $Z = 3$ and found that data points for 272 citation pairs fell above the line. Manual inspection of all of these pairs confirmed 37, or 13.6%, were true duplicates. This high rate of false positives validated the need for a second threshold based on the Rank 2/Rank 1 score ratio.

To establish a threshold for the Rank 2/Rank 1 score ratio, we calculated the numbers of true positives, true negatives, false positives and false negatives using the data from the set of 5313 randomly-selected citations. We state that sensitivity and specificity are only estimates, for we did not inspect all 5313 citations manually to obtain true and false negative rates. The true and false negative rates can be estimated by adjusting the random search data in proportion to the known duplicates that lie on either side of the Z-score = 3 curve, and assuming that the overwhelming number of publications are novel. Using the distribution of the citations of 171 known Medline duplicate publications in the Rank 2 score versus Rank2/Rank1 ratio space (Fig. 2), we can compute the sensitivity (or recall),

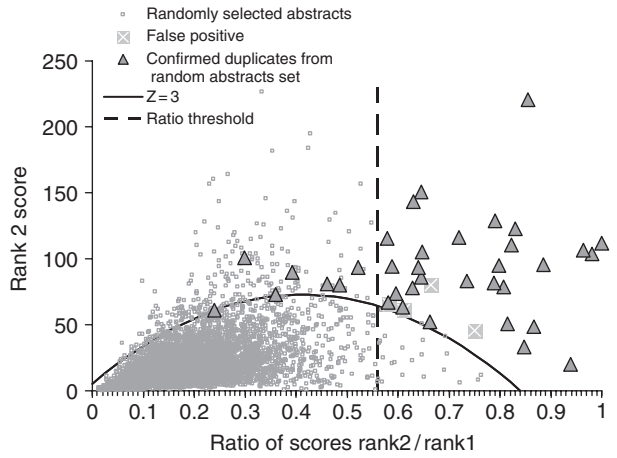

Fig. 2. The results of searching Medline with 5313 random citations as queries (small squares) are shown on a graph with Rank 2/Rank 1 score ratio (putative duplicate score/query) versus the raw Rank 2 (putative duplicate) score, two measures used to classify a citation. Inspection of the 272 citation pairs that were above the Z-score = 3 curve confirmed 37 as true duplicates (grey triangles). When the Rank 2/Rank 1 cutoff was included, 8 duplicates were erroneously classified as negative (left of 0.56 threshold) and 4 false positives (large squares) were identified. A family of Z-score curves was computed, but only the Z-score = 3 curve is shown here. The vertical dotted line represents the optimal threshold for specificity (see Fig. 4).

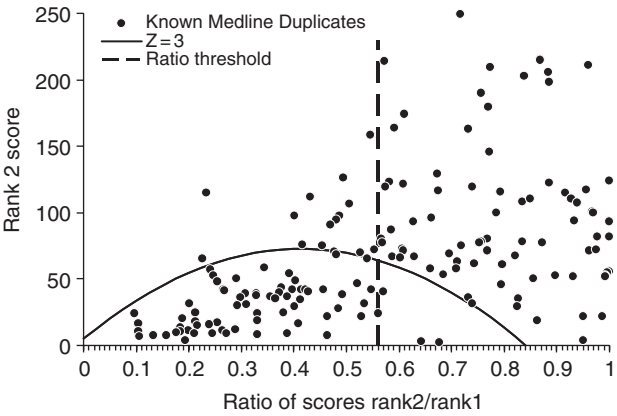

Fig. 3. The 171 citations in Medline with a Publication Type 'Duplicate Publication' after removing errata, are plotted as in Fig. 2.

the specificity, the positive predictive value (or precision) and the negative predictive value. See Tables 1 and 2. We chose to maximize the F measure, the harmonic mean of the precision and recall, as the best compromise between sensitivity and specificity, and found this selection gave a value of 0.56 for the Rank 2/Rank 1 score ratio (Fig. 4).

Fortunately, the maximum *F*-measure corresponds to parameter settings that assure a high specificity and a reasonable sensitivity not far below the maximum possible sensitivity of 62%. The selection of the Rank 2/Rank 1 value of 0.56 was further validated by extending the study to 60 000 searches, 5000 for each of the past 12 years (Table 1). Table 2 shows the details of the calculation for the set of 5313 citations.

The extended study also allowed estimates of the variation in the sensitivity and specificity. The sensitivity was calculated to be $50.3 \pm 4.0\%$, thus in any sample of Medline citations we expect to find 50.3% of the true duplicates using the $Z = 3$ and Rank 2/Rank 1 score ratio = 0.56 thresholds. The specificity was $99.8 \pm 0.1\%$, thus we expect only 0.2% of any sample of Medline citations to be falsely identified as duplicates. Applying these thresholds to the randomly-selected sample of 5313 citations, we find 33 data points above the two thresholds. Manual inspection confirmed 28 of these were duplicates with the same authors (Duplicate/SA), one was a duplicate with different authors (Duplicate/DA), and were false positives that described very similar studies or updates, duplication that we do not consider inappropriate. An additional 50 citation pairs not satisfying at least one of these thresholds were randomly selected and inspected and no duplication was found.

Table 1. Duplicate algorithm statistics averaged on the 12 year time series (60 000 searches)

Characteristics	Mean ± SD (%)
Sensitivity (or Recall)	50.3 ± 4.0
Specificity	99.8 ± 0.1
Positive predictive value (or Precision)	87.8 ± 10.9
Negative predictive value	99.3 ± 0.4

3.2 Characteristics of duplicate citations with different authors through inspection of full text

Thirteen duplicate citation pairs with non-overlapping author sets, designated Duplicate/DA in the Déjà vu database, were

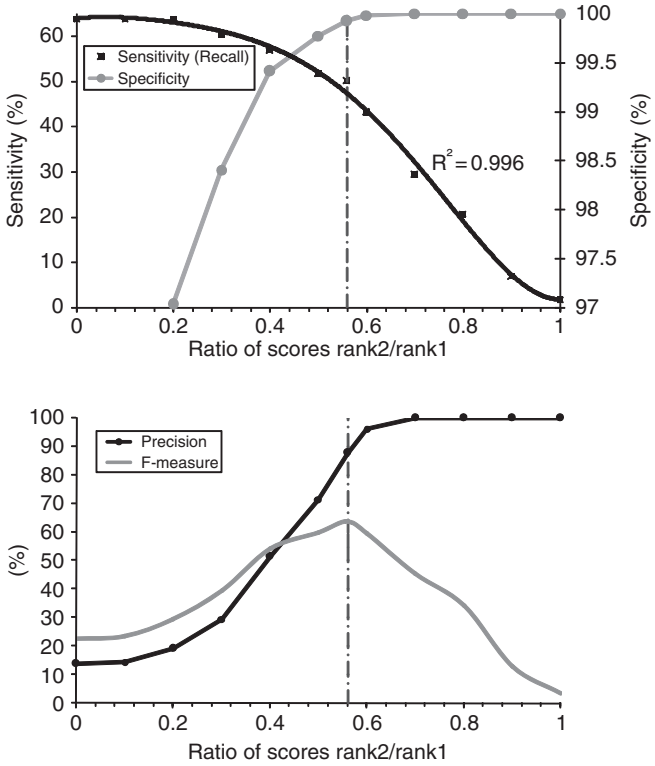

Fig. 4. The Rank 2/Rank 1 score ratio threshold was determined from inspecting the sensitivity and specificity curves. A ratio of 0.56 corresponds to the highest *F*-measure as the best compromise between precision and recall.

Table 2. Duplicate algorithm statistics based on the 5313 dataset

Characteristics	Formula	Value (%)
Sensitivity (or Recall)	TP/(TP + FN) = 29/(29 + 29)	50.0
Specificity	TN/(FP + TN) = 5251/(5251 + 4)	99.9
Positive predictive value (or Precision)	TP/(TP + FP) = 29/(29 + 4)	87.8
Negative predictive value	TN/(TN + FN) = 5251/(5251 + 29)	99.4
F-measure	2 × precision × recall/(precision + recall)	63.7

Determination of TP: TP are pairs of articles that meet both requirements: *Z*-score > 3 and Rank2/Rank1 ratio >0.56 These pairs were manually checked and found to be classified as Duplicate/SA, Duplicate/DA, Duplicate/Update/SJ, Duplicate/Update/DJ or Duplicate/Other. Determination of FP: FP are pairs of articles that like TP meet both requirements. However, upon manual verification these pairs were not found to be classified as Duplicate/SA, Duplicate/DA, Duplicate/Update/SJ, Duplicate/Update/DJ or Duplicate/Other. Estimation of FN: Note that FN is an estimation based on an extrapolation using the 171 known Medline citations of the Publication Type 'Duplicate Publication', after removing errata. Of these, 71 fell below the *Z*-score = 3 curve and 100 were above. Applying this proportion to our set knowing that we detected 29 duplicates, then we can estimate the number of false negatives missed (lying under the *Z*-score = 3 curve) as $29 \times 71/100 \sim 21$. We also confirmed 8 false negatives that are above the *Z*-score = 3 curve (but are below the 0.56 ratio threshold). Therefore, the total number of False Negatives is estimated to be: 21 + 8 = 29 Determination of TN: TN is obtained so that TN = 5313 − TP − FP − FN.

identified by similarity searching. Because of the particular ethical implications of manuscripts in this category, the full text of the corresponding articles was manually inspected. Each of these 13 pairs is analyzed in detail in the comments available in the Déjà vu database. Two of the duplicate pairs with PMIDs 25197 and 312889 and PMIDs 11320875 and 15305241 were found to result from errors in the Medline citations (see discussion below), and so these two pairs were not included in the subsequent full text analysis. Within the other true Duplicate/DA pairs in this category, the major observations include the following.

(1) On average, within the 11 Duplicate/DA pairs, 59% (SD = 27%) of the cited references are shared. This rate is significantly ≥9.3% (SD = 7.5%) of shared references for 1000 randomly selected pairs of related articles from PubMed Central.

(2) Only 50% of the later articles in the Duplicate/DA pairs cited the earlier highly similar article. This proportion is in agreement with previous studies that found that over 60% of duplicates did not have a cross-reference to their corresponding original article (Bailey, 2002; von Elm *et al.*, 2004).

(3) The duplicate publications for this category were significantly different from the duplicate publications in all other categories in two respects. First, they were more frequently published in journals with no available impact factors (82% versus 23% for all other duplicate categories, from Thomson ISI Web of Knowledge V3, http://isiwebofknowledge.com), and second, they were cited only 1/4 as often as their original counterparts, compared to the other categories where duplicates were cited almost as often, if not more often, than the original.

(4) For each of the authors of the later publications in this Duplicate/DA category all other citations available by them in Medline were inspected. Our initial search for duplicates identified two studies from two different locations reporting on survival following a surgical procedure. The later paper (PMID 9372373) did not cite the earlier one (PMID 8604907), but had identical wording in >95% of the paper, virtually identical figures and tables and an identical set of references. The number of patients in the second study was exactly double the number in the first, and the patient clinical parameters, including male/female ratio, were identical. Given the highly suspect nature of the later publication, we investigated the publication history of the author of this article, by further searching Medline and the web (Fig. 5). In total, of his eight publications, five are highly similar to articles published previously by other authors. Of these five, he published three at least twice. He also published one other of his eight articles twice.

3.3 The Déjà vu results database

All data collected in this study have been consolidated into a free, web accessible database available at http://spore.swmed.edu/dejavu. This database captures the putative duplicate citation pairs we have identified and presents them side-by-side with similarities and differences highlighted, providing a user-friendly interface to search and browse the results. This interface proved particularly useful to rapidly and efficiently classify cases of duplication. It should be noted that this is an ongoing study, and the results reported in the Déjà vu database change almost daily, so the number of duplicates in the online database may differ significantly from those reported herein.

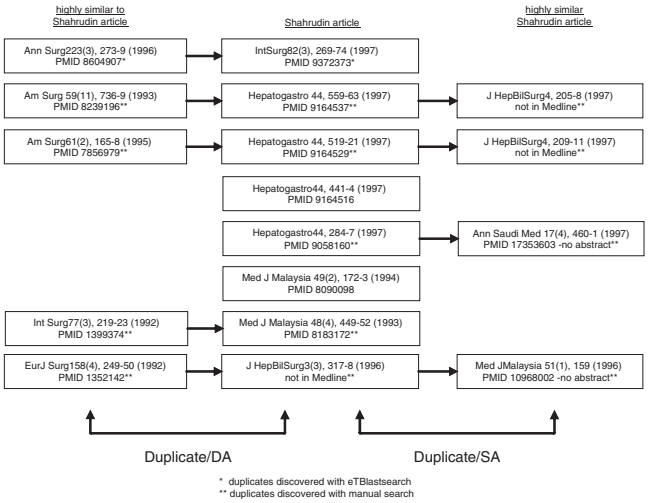

Fig. 5. Multiple duplicate publications by Shahrudin, Mohd-Dun.

4 DISCUSSION

Among the important findings in this study is that in all categories but Duplicate/DA, duplications are as visible as their original counterpart as defined by impact factor and citation index (data not shown). In the Duplicate/DA category, however, we observed that duplications were predominantly in journals with no impact factor and that these articles were rarely cited. If the primary value of a publication is to disseminate scientific findings and knowledge, it is not accomplished by publications in this category, so one must question the intent of the author of a Duplicate/DA publication. Perhaps the selection of low visibility journals is motivated by the desire to escape detection.

While on average only 10% of the references in related articles are shared, we noted that a much higher fraction of the references in the duplicate, ~60%, also appeared in the original article. This fact could be a marker of high similarity between publications within full text databases (e.g. PubMed Central), and could in the future be used to enhance the sensitivity of duplicate detection.

Although the major goal of this project was to identify typical signatures for duplicate publications and to study any general trends, much can be learned by close examination of particular duplicate pairs. These can reveal problems in the Medline database, authors with a tendency to repeat duplicate publication, and that duplicate publication may even be an early indicator of later ethical difficulties. Inspection of individual cases revealed a spectrum of text and figure

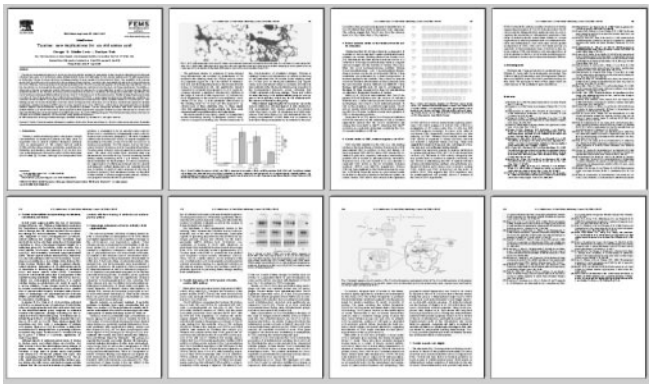

Fig. 6. Original and Duplicate/SA-classified articles share many elements. In this example of duplication with shared authors the text, figures and references that are identical (as verified by full text inspection) between PMIDs 14553911 and 14992270 are highlighted in yellow (PMID 14992270 shown here). The original paper (G Schuller-Levis and E Park, '*Taurine: new implications for an old amino acid*'. PMID 14553911) was accepted for publication five days after the submission of the duplicate (G Schuller-Levis and E Park, '*Taurine and its chloramine: modulators of immunity, a mini-review*', PMID 14992270). Note the high number of shared references.

borrowing was used by authors of duplicate publications, from replication of style with differing word usage to virtually verbatim reproduction. We also noted a frequent tendency for a review of an area following a significant manuscript by the same authors who then duplicate much of the text from their previous work, resulting in an unbalanced review. For example, in one pair of an original research article and a review article by the same authors, approximately 75% of the text, two of the five figures, and 90% of the references were identical (Fig. 6). This degree of duplicate text would represent a violation of most journal copyright agreements. Additionally, neither paper acknowledged the other.

Following the inspection of full text for the 13 citations in the category Duplicate/DA and their corresponding original, it was determined that two of the citations contained errors resulting in their being mis-categorized as Duplicate/DA, underscoring the need for manual verification. The source of the error in the citation cannot be determined (perhaps improperly submitted to Medline by the journal, improperly submitted to the journal, or a parsing error in preparation for inclusion in the Medline database), nor can that error be rectified by our group, so notes describing the suspected error are entered into the comments field for the duplicate pair in the Déjà vu database, and they were re-assigned to the appropriate category prior to statistical analysis presented herein. These errors are also being prepared for submission to the National Library of Medicine group that curates Medline. Following corrections to the Medline database, these Duplicate/DA pairs will be reassigned or eliminated from the Déjà vu database. One of the pairs in this category was PMIDs 25197 and 312889. This pair was identified and annotated because the author name on the full text of the manuscript corresponding to PMID 312889 was Marguerite MB Kay, but in the Medline citation it was Marguerite MB. This resulted in an improper assignment to the category Duplicate/DA (duplicate different author) initially based on its citation, however, it should have been assigned to the category, Duplicate/SA (duplicate same author), for although the author name was in error, the abstract was highly similar. Indeed, further investigation revealed that this author had published three highly similar articles (PMIDs 25197, 313889 and 155761) based on inspection of the full text of each. Professor Kay at the University of Arizona was later involved in another ethical controversy that resulted in her being the first tenured professor at the university to be released from her faculty appointment based on events spanning 1992–1997. (PMID 10655084, 10230727, 10712142) It is unclear if these duplicate publications that occurred in 1978 and 1979 were discussed in Dr Kay's dismissal, but these questionable actions may have been an early indication of ethical problems to come. The possibility exists that had a resource, such as eTBLAST, for identifying relatively minor ethical transgressions been available as a deterrent, Prof. Kay may not have progressed to more serious misconduct. The other pair, PMIDs 11320875 and 15305241, was identified and annotated because inspection of the English translation revealed that the English abstract that appeared in the Medline citation did not correspond to the German abstract or the general content of the full text manuscript.

A rough estimate of the total number of duplicate citations can be extrapolated from the rates observed in this study. At the end of 2006, Medline had approximately 16 million citations. Using the 1.35% sensitivity-corrected rate for duplicate citations, we can estimate that there are approximately $117\,400 \pm 40\,000$ duplicate citations in Medline. Of these, one can estimate there to be 3500 ± 1600 duplicate citations by differing authors. And Medline is growing at a rate of ~500 000 new citations per year, most of which now have abstracts (http://www.nlm.nih.gov/bsd/bsd_key.html). Therefore, approximately 6700 ± 2300 citations with overlapping authorship, of which 200 ± 100 are duplicate citations with differing authors, are being added annually. The ultimate rates for these behaviors are likely higher if, for example, only portions of full text articles are inappropriately borrowed by scientists, as these might not be reflected in a citation and thus would be missed by our approach. This condition could partially explain the discrepancy between the number of scientists admitting to duplicate publication (4.7%) and our observed rate (Budinger and Budinger, 2006). However, using the growing number of full text articles in PubMed Central, it should be possible in the future to measure the frequency of duplication of portions of manuscripts. The ethical implications of duplicating different sections of a manuscript may vary, for novelty may not be as important in introductory material or methods details, but results, discussion and conclusion sections should be considered sacrosanct.

Among the 8.7 million Medline citations with abstracts, and taking into account our sensitivity of 50.3%, a thorough search using our method would add approximately $58\,700 \pm 20\,000$ new duplicate citations. This would add substantially to the 421 new duplicates we discovered in this study and the 171 annotated duplicates that are a subset of all those labeled by Medline as a 'Duplicate Publication'. We also project we would

discover approximately 1700 ± 800 duplicate citations with different authors, each of which represents a potential case of plagiarism. Given the importance of properly gauging these computer-identified highly-similar citations and assigning them to proper classifications, we feel it is critical to inspect each manually. This manual verification is very labor intensive, but the high specificity of our method enables efficient verification. Even with our low sensitivity the number of duplicates we expect to find greatly exceeds the number of duplicates currently known.

There are limitations to our approach in addition to those imposed by the sensitivity and by the presence of abstracts in Medline citations. First, eTBLAST performs best with citations containing a large number of words, and it has limited capacity to detect duplication for citations without abstracts. Second, this methodology is presently incapable of detecting 'smart duplication', such as rewording another work while reproducing the substance. Indeed, there are technologies designed to aid authors to evade other similarity detection systems by rearranging sentences and using synonym substitution (http://www.radio-active.net.au/web/articles/articles/turnitin.html). Third, although eTBLAST searches other non-Medline databases, this approach has not yet been extended to cross database interrogations, so duplicate publications in journals not indexed in Medline will not be found. Fourth eTBLAST cannot distinguish between acceptable duplication (i.e. after conference proceedings...) and questionable duplication. Lastly eTBLAST identified 13 duplicates with differing authors, and some of our observations are based on this limited size sample. It is critical to increase the number of duplicates identified in this category to attain better statistical significance. Last, this is a study of the similarity among citations, which may or may not be representative of its corresponding full text manuscript. In some cases, a pair of unique full text manuscripts may have highly similar citations, so it is critical that users of the database inspect the full text manuscripts. Conversely, the identification of those full text manuscripts that contain inappropriate (or appropriate) duplicate text internally, but have unique citations, will not be found by this approach.

5 CONCLUSION

This study concludes that duplicate publications are a persistent problem, yet perhaps not as pervasive as some have estimated. It underscores the need for journals to enforce their submission and copyright protection policies by demonstrating that reliance upon affirmations provided by authors has not been sufficient. Duplicate detection via analysis of citations at submission time can be performed by journals now using the eTBLAST tool on the web. When combined with the Déjà vu database, this real time function can be a deterrent to this questionable scientific behavior and is a step forward in making detection easier for authors, editors and reviewers (Giles, 2005; Marris, 2006). These resources can be customized to perform on any text-based database, and although this study focused on analysis of Medline, several publicly available bibliographic databases are available including the Institute of Physics (http://journals.iop.org/), NASA (http://ntrs.nasa.gov/) and the NIH CRISP-funded grant abstract databases (http://crisp.cit.nih.gov/). Each offers opportunities to study the prevalence of duplicate publication and potentially even duplicate grant summaries.

The success of efforts to curb unethical practices will be reflected in reductions in the rates of duplicate citations and the articles they represent as monitored by processes like those discussed here. Follow-up studies may be warranted to quantitatively monitor the impact that other advances, such as the availability of electronic text databases and methods to analyze them, will have on the publication process.

ACKNOWLEDGEMENTS

We thank the NIH NCBI for providing us a free license for the Medline database. We thank Chris Renard for assisting with the computation, Karen Adams J.D. for legal advice and John W. Fondon III PhD for comments and review of this manuscript. This work was supported by the Hudson Foundation (HG), the NIH/NLM grant 1 R01 LM009758-01 (HG) and NSF-EPSCoR Grant no. EPS-0447262 (JW).

Conflict of Interest: none declared.

REFERENCES

Bailey,B.J. (2002) Duplicate publication in the field of otolaryngology-head and neck surgery. *Otolaryngol. Head Neck Surg.*, **126**, 211–216.

Barnard,H. and Overbeke,A.J. (1993) [duplicate publication of original manuscripts in and from the nederlands tijdschrift voor geneeskunde]. *Ned. Tijdschr. Geneeskd.*, **137**, 593–597.

Blancett,S.S. *et al.* (1995) Duplicate publication in the nursing literature. *Image J. Nurs. Sch.*, **27**, 51–56.

Bloemenkamp,D.G. *et al.* (1999) [duplicate publication of articles in the dutch journal of medicine in 1996]. *Ned. Tijdschr. Geneeskd.*, **143**, 2150–2153.

Budinger,T.F. and Budinger,M.D. (2006) *Ethics of emerging technologies, scientific facts and moral challenges*, Wiley & Sons, New York.

Chennagiri,R.J.R. *et al.* (2004) Duplicate publication in the journal of hand surgery. *J. Hand Surg. [Br].*, **29**, 625–628.

Durani,P. (2006) Duplicate publications: redundancy in plastic surgery literature. *J. Plast. Reconstr. Aesthet. Surg.*, **59**, 975–977.

Giles,J. (2005) Special report: taking on the cheats. *Nature*, **435**, 258–259.

Gotzsche,P.C. (1989) Multiple publication of reports of drug trials. *Eur. J. Clin. Pharmacol.*, **36**, 429–432.

Grinnell,F. (2005) Misconduct: acceptable practices differ by field. *Nature*, **436**, 776.

Kostoff,R.N. *et al.* (2006) Duplicate publication and 'paper inflation' in the fractals literature. *Sci. Eng. Ethics.*, **12**, 543–554.

Lehmann,S. *et al.* (2006) Measures for measures. *Nature*, **444**, 1003–1004.

Lewis,J. *et al.* (2006) Text similarity: an alternative way to search medline. *Bioinformatics*, **22**, 2298–2304.

Marris,E. (2006) Should journals police scientific fraud? *Nature*, **439**, 520–521.

Martinson,B.C. *et al.* (2005) Scientists behaving badly. *Nature*, **435**, 737–738.

Mojon-Azzi,S.M. *et al.* (2004) Redundant publications in scientific ophthalmologic journals: the tip of the iceberg? *Ophthalmology*, **111**, 863–866.

Roig,M. (2005) Re-using text from one's own previously published papers: an exploratory study of potential self-plagiarism. *Psychol. Rep.*, **97**, 43–49.

Rosenthal,E.L. *et al.* (2003) Duplicate publications in the otolaryngology literature. *Laryngoscope*, **113**, 772–774.

Schein,M. and Paladugu,R. (2001) Redundant surgical publications: tip of the iceberg? *Surgery*, **129**, 655–661.

von Elm,E. *et al.* (2004) Different patterns of duplicate publication: an analysis of articles used in systematic reviews. *JAMA*, **291**, 974–980.

Yank,V. and Barnes,D. (2003) Consensus and contention regarding redundant publications in clinical research: cross-sectional survey of editors and authors. *J. Med. Ethics*, **29**, 109–114.