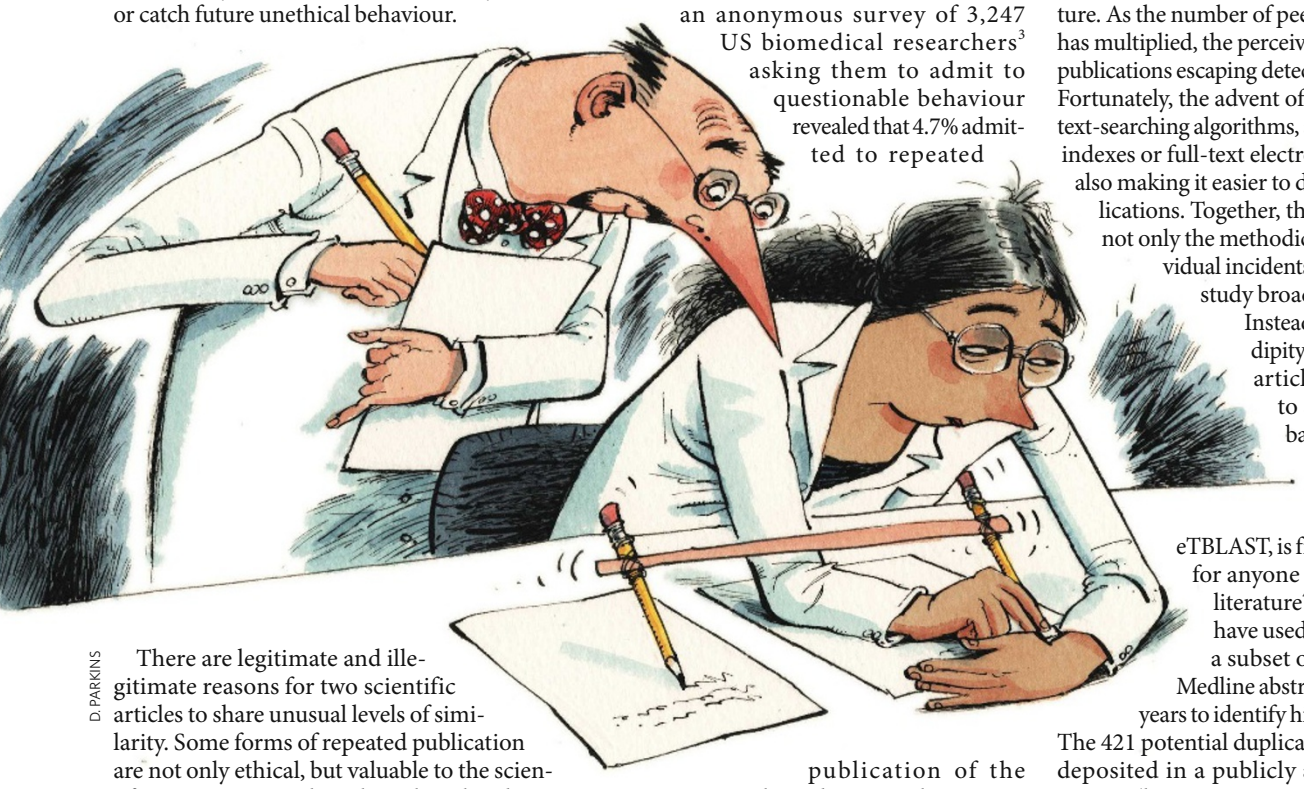
COMMENTARY

A tale of two citations

Are scientists publishing more duplicate papers? An automated search of seven million biomedical abstracts suggests that they are, report **Mounir Errami** and **Harold Garner**.

With apologies to Charles Dickens, in the world of biomedical publications, "It is the best of times, it is the worst of times". Scientific productivity, as measured by scholarly publication rates, is at an all-time high[1]. However, high-profile cases of scientific misconduct remind us that not all those publications are to be trusted — but how many and which papers? Given the pressure to publish, it is important to be aware of the ways in which community standards can be subverted. Our concern here is with the three major sins of modern publishing: duplication, co-submission and plagiarism. It is our belief that without knowing whether these sins are becoming more widespread, the scientific community cannot hope to effectively deter or catch future unethical behaviour.



D. PARKINS

There are legitimate and illegitimate reasons for two scientific articles to share unusual levels of similarity. Some forms of repeated publication are not only ethical, but valuable to the scientific community, such as clinical-trial updates, conference proceedings and errata. The most unethical practices involve substantial reproduction of another study (bringing no novelty to the scientific community) without proper acknowledgement. If such duplicates have different authors, then they may be guilty of plagiarism, whereas papers with overlapping authors may represent self-plagiarism. Simultaneous submission of duplicate articles by the same authors to different journals also violates journal policies.

Previous studies that have tried to gauge the level of unethical publishing have mostly relied on small surveys of specific communities. One of the largest to date used text-matching software to trawl more than 280,000 entries in arXiv, an open-access archive of mathematics, physics, computer science, biology and statistics papers. The study suggested a low number of suspected acts of plagiarism (0.2% of arXiv papers), but a much higher number of suspected duplicates with the same authors[2] (10.5%). In 2002, an anonymous survey of 3,247 US biomedical researchers[3] asking them to admit to questionable behaviour revealed that 4.7% admitted to repeated publication of the same results and 1.4% to plagiarism.

"The duplication of scientific articles has been largely ignored by the gatekeepers of scientific information."

In general, the duplication of scientific articles has largely been ignored by the gatekeepers of scientific information — the publishers and database curators. Very few journal editors attempt to systematically detect duplicates at the time of submission. The US National Library of Medicine, based in Bethseda, Maryland, curates the primary biomedical citation index, Medline, and currently reports fewer than a thousand cases of duplication since the 1950s, discovered mainly by serendipity. Yet if the results of the anonymous survey[3] are extrapolated to the Medline database (more than 17 million citations and growing steadily), then you would expect to find closer to 800,000 cases. Where between these two vastly different figures does the true number lie?

The academic arms race

Establishing a baseline is a crucial first step, but in our view, monitoring trends is even more important to the health of the scientific literature. As the number of peer-reviewed journals has multiplied, the perceived odds of unethical publications escaping detection have improved. Fortunately, the advent of new computational text-searching algorithms, along with electronic indexes or full-text electronic manuscripts, is also making it easier to detect unethical publications. Together, these advances enable not only the methodical discovery of individual incidents, but also a means to study broad trends.

Instead of relying on serendipity to identify duplicate articles, we have chosen to search online databases, such as Medline, using text-similarity software. The search engine, eTBLAST, is freely available online for anyone to use to search the literature[4]. In recent work, we have used eTBLAST to search a subset of more than 62,000 Medline abstracts from the past 12 years to identify highly similar entries[5]. The 421 potential duplicates found have been deposited in a publicly available database, Déjà vu (http://spore.swmed.edu/dejavu), and after manual inspection were confirmed as duplicates with different authors (0.04%; based on inspection of full-text articles), or duplicates with the same authors (1.35%; based on inspection of the abstracts). The rate of false positives in this study was only 1%. But without full text it may be difficult to determine if suspected duplicates properly attributed the earlier work.

Whether or not the duplications are legitimate papers has yet to be established.

Extrapolating to the subset of Medline records that have abstracts (8.7 million), this would correspond to roughly 117,500 duplicates with the same authors[4]. Although this number is far higher than the 739 records currently annotated as duplicates in Medline, these duplication rates are substantially lower than those found in arXiv, perhaps reflecting differences in the database formats (preprints versus journal papers), or disparities between these fields in what is considered acceptable practice. There is also variation in how these estimates were reached, including the subjective nature of manual inspection (we used two manual checkers in each case). The Medline database, unlike arXiv, is limited to titles and abstracts, and so automated comparison of full-text articles is not possible, perhaps making it harder to detect more sophisticated duplications.

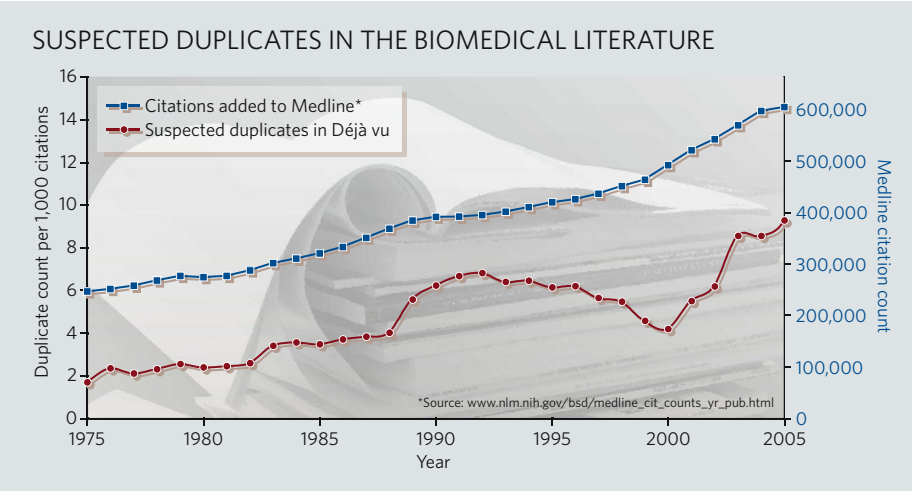


Figure 1. Increasing opportunity? The number of biomedical papers indexed in the citation database, Medline, has grown steadily over the past 30 years. A search of 7 million abstracts, using the text-matching software eTBLAST, reveals tens of thousands of highly similar articles (unpublished data), which are also growing in number. Are these legitimate or illegitimate publications?

Closer than close

Because of the sheer size of the Medline database, scaling up the eTBLAST search to all 17 million records would be extremely time consuming even though each search takes only about 40 seconds. Fortunately, we observed that 73% of the Medline duplicates identified in our initial study and curated in Déjà vu also feature as the 'most related article' in Medline (calculated by a Medline algorithm). So, we downloaded the related abstracts for 7,064,721 Medline records, and compared the original and related abstracts against one another using eTBLAST. This approach allowed us to complete our analysis in 10 days rather than 10 years. In this way we have identified a further 70,458 highly similar records, all of which have been deposited in Déjà vu.

"Automated text-matching systems are used by high schools and universities. We hold our children up to a higher standard than we do our scientists."

Given the limitations of our process, we expect around 50,000 of these to be true duplicates. This is partly because we used a less stringent duplication threshold for the latest data set and so after manual checking 27% of the records turn out to be false positives (see http://spore.swmed.edu/dejavu/statistics). To date, 2,600 of the Déjà vu records have been manually inspected alongside the original, but until that is done the status of each entry remains unverified. However, extrapolating to the entire database, we estimate there are potentially more than 200,000 duplicates in Medline, after various correction factors have been applied.

Although manual verification of the Déjà vu database is very much a work in progress, and so analysis of the full data set should be interpreted with caution, we have started looking for trends in the approximately 70,000 candidate duplicates. With the articles so far captured within the Déjà vu database, merged with analysis of other data extracted from full-text versions of Medline articles available in PubMed Central (such as publication date, language of article and country of origin), it is possible to begin to identify broad trends in publication behaviour. Perhaps the most obvious is a steady rise in the rate of such publications in the biomedical literature since 1975 (Figure 1).

Medline indexes over 5,000 journals published in the United States and more than 80 other countries worldwide. Rising duplicate publication rates documented in Figure 1 are therefore a global phenomenon. Potential factors contributing to this trend are the explosion in the number of journals with online content (increasing opportunities for unethical copying), and a body of literature growing so fast that the risk of being detected seems to diminish. This last factor may be the most important, and we believe that automated detection processes that can provide an effective deterrent may be our best weapon in fighting duplicate publications.

One argument for duplicate publication is to make significant works available to a wider audience, especially in other languages. However, only 20% of manually verified duplicates in Déjà vu are translations into another language. What of the examples of text directly translated with no reference or credit to the original article? Is this justified or acceptable? And is such behaviour more widespread for review-type articles for which greater dissemination may be justified? We do not yet have answers to these questions.

In general, we find that the duplication rate extracted from the total Déjà vu database for each country is roughly proportional to the number of manuscripts that country contributes to Medline (Figure 2). The top eight contributors to Medline are the United States, Japan, Germany, China, the United Kingdom, Italy, France and Canada, representing close to 75% of all Medline records. However, two of these countries, China and Japan, have estimated duplication rates that are roughly twice that expected for the number of publications they contribute to Medline. Perhaps the complexity of translation between different scripts, differences in ethics training and cultural norms contribute to elevated duplication rates in these two countries.

Simultaneous submission

With few exceptions, the repeated publication of the same results by those who conducted the research is ethically questionable. It not only artificially inflates an author's publication record but places an undue burden on journal editors and reviewers, and is expressly forbidden by most journal copyright rules.

Examination of typical submission and publication dates from 10,000 articles randomly selected from PubMed Central, shows that on average the review process takes 4.3 months and that 97% of articles complete this process within 10 months (see Supplementary information). Curiously, as many as one-third of the manually verified duplicate abstracts in Déjà vu sharing at least one author are also published less than

five months after the original. Examination of the submission and publication dates of these pairs confirms that many of these duplicates must have been submitted simultaneously to different journals in violation of journal policies and accepted norms. For instance, the Déjà vu database contains many pairs of highly similar abstracts with overlapping authors that appear in the same month, all apparently acts of simultaneous submission to multiple journals.

Duplication by different authors

Articles sharing excessive similarity with other papers with different authors do not necessarily represent plagiarism, as there are sometimes valid or trivial reasons (such as a simple author name change). However, considering only those duplicates in Déjà vu where the full text of both articles has been manually inspected, we have found 73 plagiarism candidates, most of which were previously undetected. Discerning the difference between legitimate and illegitimate duplication is beyond the capacity of automated algorithms (and apparently many scientists), and so it is critical to withhold judgement of any candidate duplicates until evaluated by a suitable body such as an editorial board or a university ethics committee. As part of our study, we have started to send out requests for additional information for such cases, one of which has initiated an investigation by a journal. It is our intent to send such requests for information to all individuals and journals involved in, or affected by, duplicate records with different authors.

Many duplicate articles without authors in common go undiscovered. Are the perpetrators then likely to repeat the offence? Searching the Déjà vu database reveals several repeat practitioners, and manual inspection of full-text articles confirms some of these as suspected serial offenders. As with any potential illegitimate duplication, caution and careful human judgement must be exercised, and detailed comments and manual assessments for these and other duplicate pairs can be found within the Déjà vu database.

Unlike repeated publication by the same authors, simultaneous publication is rarely observed for duplicates that do not share authors (see Supplementary information), undoubtedly due to the fact that it is usually difficult to re-use someone else's work before it appears in print — unless the duplicating author also happens to have been a referee of the original. Although anecdotes abound of referees stalling a publication in order to give themselves time to duplicate and publish the same result first, the general lack of duplicates with different authors appearing in rapid succession suggests that this is either rarer than feared, or that the perpetrators do a good job of concealing it.

In general, duplicates are often published in journals with lower impact factors (undoubtedly at least in part to minimize the odds of detection) but this does not prevent negative consequences — especially in clinical research. Duplication, particularly of the results of patient trials, can negatively affect the practice of medicine, as it can instill a false sense of confidence regarding the efficacy and safety of new drugs and procedures. There are very good reasons why multiple independent studies are required before a new medical practice makes it into the clinic, and duplicate publication subverts that crucial quality control (not to mention defrauding the original authors and journals).

What can be done?

Although duplicate publication and plagiarism are often discussed, it seems that discussion is not enough. Two important contributing factors are the level of confusion over acceptable publishing behaviour and the perception that there is a high likelihood of escaping detection. The lack of clear standards for what level of text and figure re-use is appropriate (for example in the introduction and methods) is a well known problem; but the belief that one can get away with re-use is probably the single most important factor.

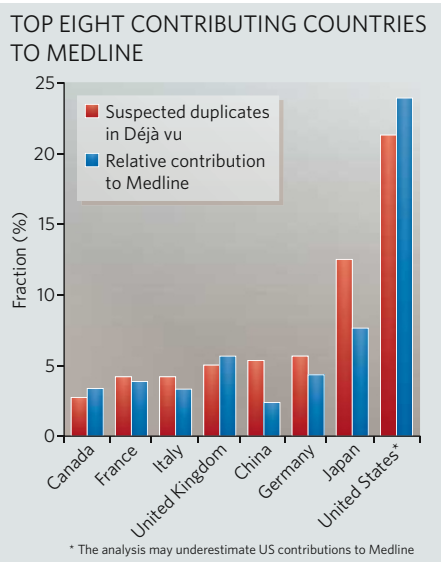


Figure 2. Duplication is a global activity. The proportion of suspected duplicates in the Déjà vu database for each country was estimated (unpublished data) by assigning articles to countries based on the corresponding author's address. Also presented is each country's relative contribution to Medline estimated from 180,000 randomly selected Medline articles.

Addressing these two aspects could be relatively quick and easy. If journal editors were to use more frequently the new computational tools to detect incidents of duplicate publication — and advertise that they will do so — much of the problem is likely to take care of itself. We find it odd that automated text-matching systems are used regularly by high schools and universities, thereby enabling us to hold our children up to a higher standard than we do our scientists. In our view, it would be fairly simple to fold these tools into electronic-manuscript submission systems, making it a ubiquitous aspect of the publication process.

Although text-comparison algorithms have come a long way in the last decade, they are still in their infancy, and experience with student software shows that as tools to detect duplicate publication improve, determined and skilled cheats will find ways to defeat them. But as in any arms race, the winners are usually determined by the cost–benefit balance, and the costs entailed in unethical duplication practices will quickly rise to a level that makes them prohibitively expensive to all but the most desperate (or most skilled) practitioners.

"The fear of having some transgression exposed in a public and embarrassing manner could be a very effective deterrent."

There are additional practical avenues for improving Medline and other databases, such as more aggressive enforcement of copyrights by journals, and the creation of an 'update' publication category under which clinical updates and longitudinal surveys in sociology or psychology could be categorized, and these should be explored.

But above all, the fear of having some transgression exposed in a public and embarrassing manner could be a very effective deterrent. Like Dickens's Ebenezer Scrooge, the spectre of being haunted by publications past may be enough to get unscrupulous scientists to change their ways. ■

Mounir Errami is in the Division of Translational Research Department and Harold Garner is in the McDermott Center for Human Growth and Development, The University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-9185, USA.

1. http://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html
2. Sorokina, D., Gehrke, J., Warner, S. & Ginsparg, P. *Sixth International Conference on Data Mining* 1070–1075 (2006).
3. Martinson, B. C., Anderson, M. S. & de Vries, R. *Nature* **435,** 737–738 (2005).
4. Errami, M., Wren, J. D., Hicks, J. M. & Garner, H. R. *Nucleic Acids Res.* **35,** W12-5 (2007).
5. Errami, M. *et al. Bioinformatics* advance online publication, doi:10.1093/bioinformatics/btm574 (2007).

Supplementary information is linked to the online version of this article at www.nature.com/nature