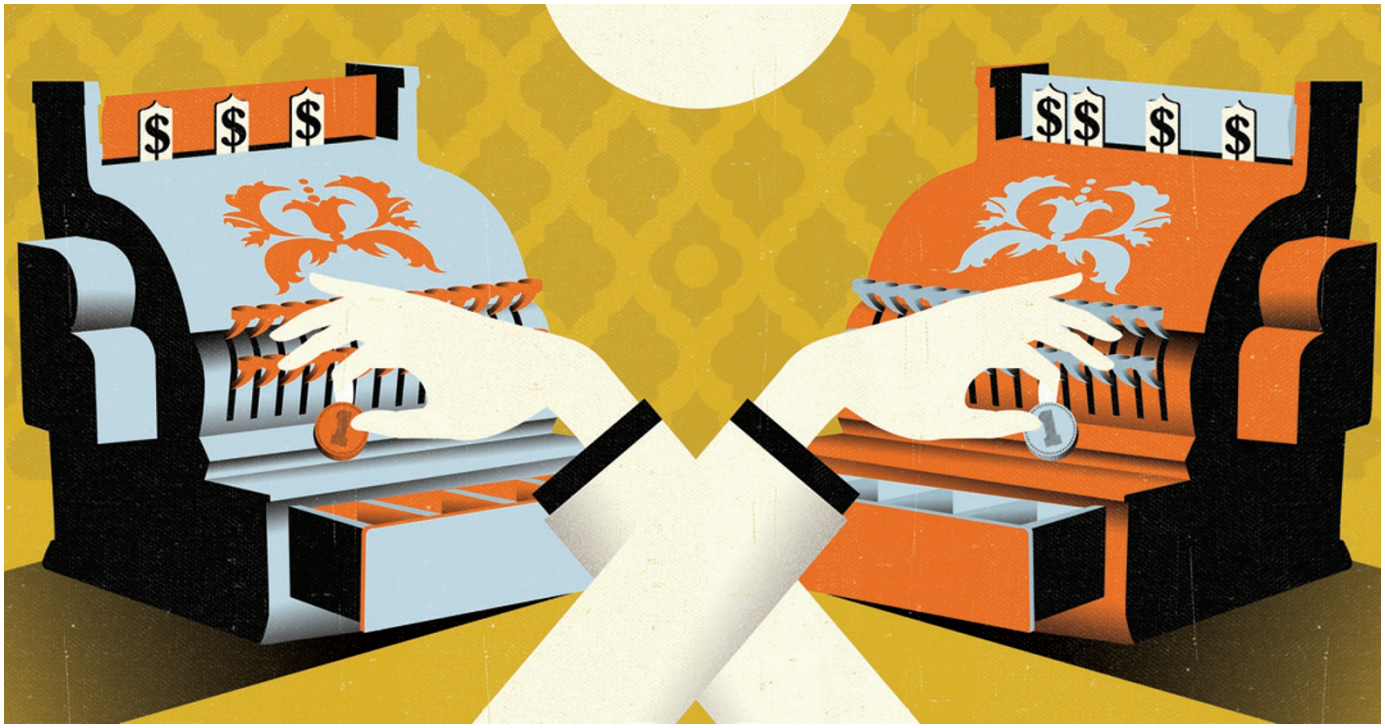
COMMENT

NEIL WEBB



Same work, twice the money?

Funding agencies may be paying out duplicate grants, according to an analysis by **Harold R. Garner**, **Lauren J. McIver** and **Michael B. Waitzkin**.

With grant success at an all-time low[1], scientists are working harder than ever to fund their research. They respond to the competitive economic times by submitting more applications. They may also simultaneously or serially submit applications to multiple funding agencies to increase their odds of getting funding. Some grant agencies allow the submission of applications with identical or highly similar specific aims, goals, objectives and hypotheses. But we believe that researchers should not accept duplicate funding for the same work — either the whole study or any part of it.

In February last year, the US Government Accountability Office audited the three federal agencies that provide about 94% of all federal funding for medical-sciences research in the United States — the National Institutes of Health (NIH), Department of Defense (DOD) and the Veterans Administration — and found "a potential for unnecessary duplication"[2]. It suggested that the agencies "improve the ability of agency officials to identify possible duplication". The NIH responded to the audit by requiring a detailed evaluation of all proposals from researchers who receive more than US$1.5 million a year in funding to detect any possible "dual/overlapping support"[3]. It has not yet reported any results.

To estimate the extent of double-funding, we systematically compared more than 850,000 funded grant and contract summaries submitted to five of the largest US funders of biomedical research using text similarity software that one author (H.R.G.) invented; we then reviewed a subset of the summaries manually.

We could not determine definitively whether the similar grants we identified were true duplicates — this would require access to the full grant files, which are not available to us without Freedom of Information Act (FOIA) requests. But we did find ▶

evidence that, since 1985 — the earliest year for which grant summaries are available — tens of millions of dollars may have been spent on projects in which at least a portion of the research was already being funded. The problem probably continues today — in the most recent 5 years (2007–11), there were 39 concerningly similar grant pairs, involving more than $20 million. Some of the potential duplicate grants we discovered with our software may have already been identified by the relevant agencies, which may have adjusted the award amount accordingly without updating summaries. But we suspect that there may be many more cases of duplication than our analysis implies.

These findings suggest to us that the research community must launch a more thorough investigation into the true extent of duplication. There should be better, clearer and more consistent coordination and guidance about duplication of funding across agencies, both public and private. A central database for all grant proposals would be an excellent first step.

FINDING PAIRS

Government agencies require the disclosure of all current or pending research support, as well as whether the same or a similar proposal is being submitted to another agency. Although explicit rules for every possible scenario do not exist, a good and safe practice is to report any new resources for a study to all funders, and allow them, with this full disclosure, to make a determination. Any violations may open up the grant applicant to criminal prosecution for fraud, civil liability for filing false claims or administrative sanctions, such as debarment from government contracting or suspension as an investigator. Despite these rules, there have been very public cases, often found by serendipity, in which principal investigators have accepted multiple sources of funding for the same project without declaring the existence of other sources[4].

Early last year, we downloaded funded grant summaries (corresponding to 858,717 grants or contracts) from public websites in the NIH, the National Science Foundation (NSF), the DOD, the Department of Energy (DOE) and the Susan G. Komen for the Cure, the largest charitable funder of breast cancer research in the United States. We eliminated more than one-quarter (227,380) of the summaries because they contained fewer than 50 words, so could not be processed accurately by our computational methods. As a result, 2×10^{11} $((858{,}717 - 227{,}380)^2 / 2)$ text-similarity comparisons were possible, so our analysis would only be able to find 54% of all possible duplicates. The number and dates of applications obtained varied greatly across agencies (see 'Double-dip analysis'), with awards totalling more than $200 billion. We recently revisited these databases and found that the DOE had removed the entire database of funded grant summaries.

Our text similarity engine[5], called eTBLAST, calculates a similarity score between each pair of grant summaries using the same approach we established to identify potentially plagiarized scientific literature[6,7]. A two-pass, full-text algorithm accurately and efficiently calculates similarity scores on the basis of the number of shared words placed in the same order in sentences[5]. For this collection of documents, the similarity score for all pairs ranged from 0 to 1.8 (with 1 indicating identical text in two same-length documents, and more than 1 representing identical text in one piece that is longer than the other). We manually reviewed all documents from the federal agencies that received a score of 0.8 or more, and all those from the Susan G. Komen for the Cure that received a score of at least 0.65 — arbitrary cut-offs that in our view provided a sufficiently large sample (1,300 summaries) to capture most of the duplicates. We excluded grants that were obviously associated with an inter-agency combined effort, such as, for example, support for workshops or conferences, large equipment purchases and research involving national laboratory partners.

SAME DIFFERENCE

We focused our manual inspection on comparing specific aims, objectives, goals and hypotheses — because high similarity can result from reuse of introductory or background material.

We found that 11% of the pairs with a similarity of more than 0.8 (or 0.65) had overlapping aims, hypotheses or goals. For these 167 pairs the total money involved was around $200 million (including both grants of the pair) over the entire time records are available. The average size of the first award was 1.9 times that of the potentially overlapping one, so an estimated $69 million of possible overlap funds were found.

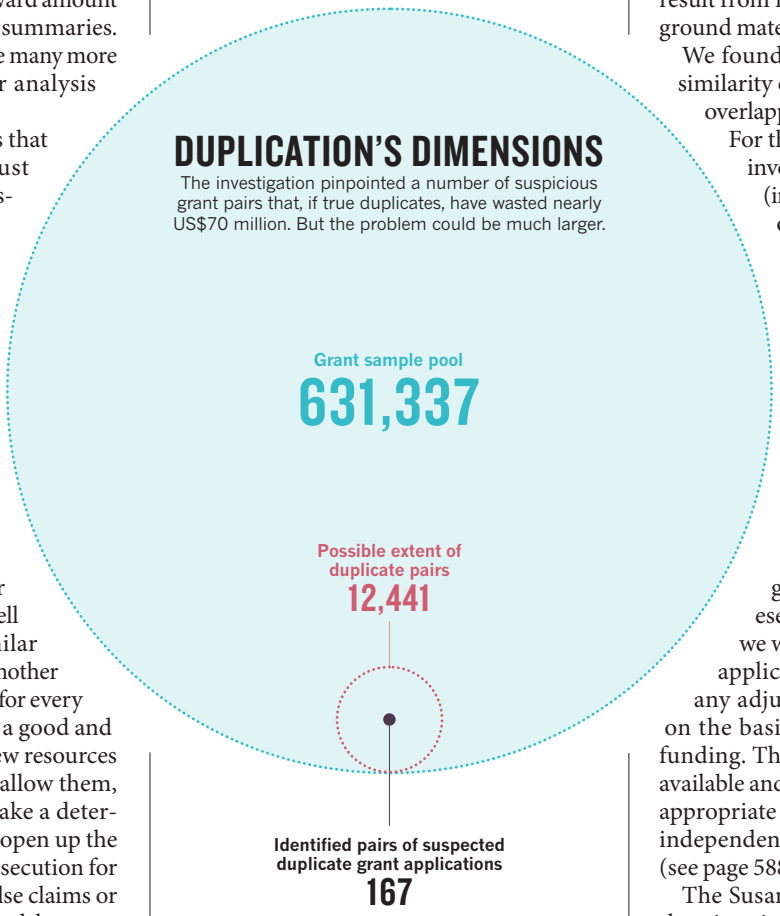
Our analysis does not determine whether any likenesses in funded grant pairs are inappropriate, only that the short summaries contained highly similar aims, goals, objectives and hypotheses. To identify true duplicates, we would need to compare the full applications, the awards made and any adjustments made to the awards on the basis of disclosures of duplicate funding. This information is not publicly available and would need to be analysed by appropriate governmental agencies, or by independent groups using FOIA requests (see page 588).

The Susan G. Komen for the Cure did share its private analysis of the highly similar grant pairs we identified (but not the entire applications). At our request, it evaluated 30 pairs, four that it had funded in advance of other agencies, eight that it had funded concurrently and 18 that it had funded subsequent to other agencies. Only four of these pairs had a similarity score of less than 0.8, suggesting that this threshold captured a large fraction (87%) of similar grants.

For two of the pairs, the Susan G. Komen for the Cure reported that its ongoing internal administrative review of grant funding had already identified and made adjustments to funding, which was not reflected in its online summaries. It immediately began to review two active projects with potential overlap that it had not identified. The remaining suspected duplicates would have required the foundation to obtain complete grant applications either from its

DUPLICATION'S DIMENSIONS

The investigation pinpointed a number of suspicious grant pairs that, if true duplicates, have wasted nearly US$70 million. But the problem could be much larger.



archives or from other agencies (using FOIA requests) to make a thorough review, which it did not do.

A better estimate of duplicate funding would account for the grant summaries that were too short to analyse and for those grant pairs with similarity of less than 0.8 that may nonetheless overlap. Using the amount of possible overlap funding ($69 million) inferred from the grants we reviewed, and accounting for the two sensitivities (54% and 87%, the portion of grants captured by the 0.8 threshold), our view is that an exhaustive analysis of all grants could reveal twice as much overlapping funding (69 million/ $(0.87 \times 0.54) = \$147$ million).

Among the summaries with identical or highly similar specific aims, objectives, goals and hypotheses, we found that about 31% ran concurrently. The rest may have been 'recycled', whereby a principal investigator had sent a previously successful grant to another agency. Strangely, the later grant sometimes proposed studies that were cited as preliminary data in the earlier grant, suggesting that the hypothesis had already been resolved and that the proposed research had already long been completed.

With some pairs, a principal investigator seemed to have received a grant that included support for lab members, then sent the grant to another agency for support of graduate students or postdoctoral fellows. There seem to be no clear standards on whether this is an acceptable practice. A full review would be necessary to determine whether the additional funding for the fellow or student on an already funded project was fully disclosed and whether the science project grant budget was adjusted appropriately.

"We believe that our analysis may actually have missed duplications."

We also found similar grants that had different principal investigators (at the same or different institutions) and were funded by different agencies, suggesting that principal investigators may be sharing successful grants with others, thereby enabling them to amplify the amount of funds for a given project, and technically bypassing the formal definition of 'double dipping'. Some of these pairs included current or former co-authors on journal publications, which we discovered from searches of published literature.

In a sampling of around 20 similar grant pairs, we looked in PubMed for publications resulting from the funding, and found that some acknowledged only one agency. Some publications acknowledged additional grant numbers, which, after review, revealed further highly similar grant summaries. These did not meet our 0.8 threshold, but indicate another technique for discovering potential overlap.

DOUBLE-DIP ANALYSIS

After comparing grant and contract summaries with software, all those with a high similarity score were reviewed manually to identify possible duplicates.

Funder	Dates available	Number of applications	Reviewed digitally	Reviewed manually	Suspicious overlaps
Susan J. Komen for the Cure	2003–2011	1,209	1,208	98	30
Department of Defense	1993–2011	10,201	10,086	157	68
Department of Energy*	1995–2009	38,408	9,731	20	3
National Science Foundation	1985–2012	299,332	221,513	446	92
National Institutes of Health	1985–2012	509,567	388,799	579	141
Total		858,717	631,337	1,300	334

*Stopped reporting these data in 2009.

Justifiably, critics will counter that our limited analysis overestimates the problem by not factoring in whether funding agencies have adjusted awards for previous support, and that it suffers from a lack of access to the full grant application and data from all years from all agencies.

However, from our experience in detecting plagiarism using text similarity, we believe that our analysis may actually have missed duplications. For instance, the same comparison techniques have detected plagiarism in 0.04% of biomedical manuscripts[7]. Yet 1.4% of scientists in one survey[8] admitted to plagiarism — that's 35 times the estimated number of duplications in that analysis. If — and it is a very big 'if' — a similar level of duplication did apply in grant applications, the problem could have involved 12,441 pairs of applications (see 'Duplication's dimensions') and up to $5.1 billion since 1985 (or 2.5% of the total funds).

Even if $200 million in duplicated grants represents the full extent of the problem, then some may argue that less than 0.1% of funding since 1985 is too small an amount to warrant concern. But that it is research money that cannot be used to fund the next scientific breakthrough.

GRANT DATABASE

Our findings indicate a need for clearer and more consistent guidance and coordination of grant and contract funding across agencies, both public and private. They may also indicate a need to clarify the standards on what constitutes duplicate funding and to strengthen the surveillance of proposals and funded projects for overlap to ensure adherence to regulations and intent.

We feel that funding agencies and recipient institute administrations could curb duplicate funding more than they are currently doing by using text-similarity comparisons to identify applications and funded grant summaries that warrant closer human scrutiny. That said, similarity software must be continuously updated to respond to changes in grant formats and attempts to evade detection.

Most importantly, creating a central database of grant information from all agencies would enable thorough direct comparisons of all awarded funding, and the prospective identification of similar grant proposals. Such a database, including detailed information within grant applications, and its analysis should remain confidential to ensure that only appropriate facts are released beyond government agencies and their staff. Although this will add some administrative burden, it would help agencies prioritize their awards and target the available dollars more efficiently. ■

Harold R. Garner *and* **Lauren J. McIver** *are at the Virginia Bioinformatics Institute, Virginia Tech, Washington Street, Blacksburg, Virginia.* **Michael B. Waitzkin** *is at Genomeon, Floyd, Virginia.*
e-mails: garner@vbi.vt.edu; mbwaitzkin@gmail.com

1. Kaiser, J. *Science Insider* (20 January 2012).
2. US Government Accountability Office *Report to Congressional Addressees, 2012 Annual Report: Opportunities to Reduce Duplication, Overlap and Fragmentation, Achieve Savings, and Enhance Revenue.* (GAO, 2012); available at go.nature.com/bufrae.
3. Rockey, S. Piloting the $1.5M Special Review. National Institutes of Health Office of Extramural Research Extramural Nexus (8 May 2012). available at go.nature.com/ymffdj
4. Samuel Reich, E. *Nature* **482,** 146 (2012).
5. Lewis, J., Ossowski, S., Hicks, J., Errami, M. & Garner, H. R. *Bioinformatics* **22,** 2298–2304 (2006).
6. Errami, M. *et al. Bioinformatics* **24,** 243–249 (2008).
7. Errami, M. & Garner, H. R. *Nature* **451,** 397–399 (2008).
8. Martinson, B. C. Anderson, M. S. & de Vries, R. *Nature* **435,** 737–738 (2005).

H.R.G. declares competing financial interests. For full details see go.nature.com/q6fkrt.

Editor's note *Nature* is not publishing the grant summaries analysed in this Comment, nor the names of their authors (*Nature*, like most journals, requires Comment authors to make these data available on request). This is because a definitive demonstration of duplication would require access to documents that are not available to the Comment authors. Moreover, the grant authors were not approached for their responses to the analysis. This Comment, although less formal than a *Nature* Letter or Article, underwent peer review.