

## Metrics Associated With NIH Funding: A High-Level View

Kevin W. Boyack<sup>1</sup>, Paul Jordan<sup>2</sup>

<sup>1</sup>*kboyack@mapofscience.com*

SciTech Strategies, Inc., 8421 Manuel Cia Pl. NE, Albuquerque, NM 87122 (USA)

<sup>2</sup>*jordan1@od.nih.gov*

NIH OD OER ORIS, P.O. Box 12233, MD K3-05, Room 3108,  
Research Triangle Park, North Carolina 27709 (USA)

### Abstract

**Objective:** To introduce the availability of grant-to-article linkage data associated with NIH grants and to perform a high-level analysis of the publication outputs and impacts associated with those grants.

**Design:** Articles were linked to the grants they acknowledge using the grant acknowledgement strings in PubMed using a parsing and matching process as embodied in the NIH SPIRES system. Additional data from PubMed and citation counts from Scopus were added to the linkage data. The data comprise 2,572,576 records from 1980-2009.

**Results:** The data show that synergies between NIH institutes are increasing over time; 29% of current articles acknowledge grants from multiple institutes. The median time lag to publication for a new grant is 3 years. Each grant contributes to approximately 1.7 articles per year, averaged over all grant types. Articles acknowledging U.S. Public Health Service (PHS, which includes NIH) funding are cited twice as much as U.S. authored articles acknowledging no funding source. Articles acknowledging both PHS funding and a non-U.S. government funding source receive on average 40% more citations than those acknowledging PHS funding sources alone.

**Conclusion:** The U.S. PHS is effective at funding research with higher than average impact. The data are amenable to further and much more detailed analysis.

### Introduction

Although science policy studies have been conducted for decades, interest in such studies is currently on the rise in the United States, as well as other countries. This is evidenced by the number of recent workshops highlighting “science of science policy” as well as the establishment and funding of a Science of Science and Innovation Policy (SciSIP) program at the National Science Foundation (NSF). Despite the long historical interest in science policy, quantitative input-output studies establishing the impact of programs at different agencies and institutes have been very difficult owing to the fact that data explicitly linking articles with the grants from which they were funded have been lacking. One place where these data do exist is the PubMed database, which has been indexing grant numbers for U.S. Public Health Service (PHS) grants since at least 1981.

The fact that PubMed records contain grant acknowledgements does not, however, mean that they have been systematically used for research evaluation. In fact, the opposite is true. Although these data exist, they were not systematically mined, standardized, and placed in a publicly available database for bulk download until just recently. The National Institutes of Health (NIH) Research Portfolio Online Reporting Tool Expenditures and Results (RePORTER) website was made available in mid-2009. This site allows one to query a wide variety of NIH funding and publication data. The ExPORTER extension to make selected data available for bulk download ([http://projectreporter.nih.gov/exporter/ExPORTER\\_Catalog.aspx](http://projectreporter.nih.gov/exporter/ExPORTER_Catalog.aspx)) was introduced several

months later. Data have been added to the ExPORTER site at intervals over the past year – data back to 1985 are currently available, including tables for projects, links (articles to projects), articles, and patents.

Prior to the introduction of RePORTER and ExPORTER, it was possible to obtain funding and publication data for input-output studies. Grant data were available from CRISP and RaDiUS ('Computer Retrieval of Information on Scientific Projects' and 'Research and Development in the United States' databases, now both defunct), and publication and grant acknowledgement data were available from PubMed. However, significant efforts were required to obtain, clean and link such data, and were thus a limiting factor in the number and scope of input-output studies (1).

This paper reports work on two separate tasks. First, we report on work done at NIH to generate and provide a clean, standardized source of article-to-grant linkage data from grant acknowledgements in PubMed. Second, we have linked those data to records from Scopus to add citation counts, and have used the combined data for a first set of high-level analyses to show distributions and metrics related to NIH grants. The balance of this paper proceeds as follows. First, relevant literature on linking of grant and article data will be briefly reviewed. The NIH process to link grant and article data will then be described. A high-level characterization of the linked data and various metrics associated with those data is then presented. The paper concludes with a discussion of benefits, limitations, and suggestions for future work.

## **Background**

### *Input-Output Studies*

Perhaps the most comprehensive input-output studies were done in the 1980's by Computer Horizons, Inc. For example, McAllister and colleagues studied the relationship between R&D expenditures and publication outputs for U.S. colleges and universities (2) and U.S. medical schools (3) on a large scale using aggregated funding amounts, and publication and citation counts. Bourke and Butler (4) reported on the efficacy of different modes of funding research in biological sciences in Australia. Their work aggregated funding to the sector level, and concluded impact was correlated with researcher status. Butler (5) followed this work up with a study of funding acknowledgement, finding that, although acknowledgement data on the whole accurately reflected the total research output of a funding body, there was no ability to track research back to the grant level. This inability to track research back to an individual grant precludes analyses of research vitality at the finest levels. Additional studies using aggregated results are also available in the literature (cf., 6, 7, 8).

Far fewer studies are available in which actual linking of grant data to individual articles has been reported. Computer Horizons, Inc. mined and maintained funding data from the acknowledgements in journal articles, and used them for a variety of studies for the U.S. NIH in the 1980's (9). However, neither their grant-article linkage data nor their reports to NIH are readily available. Lewison and colleagues (10-12) used citation data from the Science Citation Indexes and acknowledgement data from the UK Research Outputs Database to study national level impacts in various biomedical fields. Although they mention looking up articles and extracting funding information, no grant-article level analysis is reported. Boyack and colleagues linked grants to individual articles through common author/PI and institution using data supplied

by the National Institute on Aging (13), and showed that citation impact increased with grant size (14). They also showed funding profiles for NIH and NSF on a map of science (15), using grant data from 1999 and article data from 2002 linked through author/PI and institution. Recently, Zhao (16) examined a set of 266 articles in library and information science journals, and found that articles acknowledging grant funding were cited over 40% more on average than those without a grant acknowledgement. Lyubarova et al. (17) investigated the impact of NIH funding on cardiovascular research and found that the mean journal impact factor of NIH-funded research was significantly higher than that of non-NIH-funded research for most article types.

### *Data History*

NIH has recently replaced its CRISP database with the Research Portfolio Online Reporting Tools Expenditures and Results (RePORTER) tool. Much of the data in RePORTER comes from the Scientific Publication Information Retrieval & Evaluation System (SPIRES), which was begun by one of us (Jordan) while at the National Institute of Environmental Health Sciences (NIEHS) in 2001. The initial version of the SPIRES database, completed in Spring 2002, was limited to data for three NIH institutes – NIEHS, NICHD, and NIA – and linked articles acknowledging grants from these three institutes back to 1995. The initial system was based on parsing grant numbers from PubMed records, and then matching those grant numbers to a table of known grant numbers. The initial version of SPIRES was updated monthly, with grant numbers being parsed from PubMed records and matched to tables of grant numbers using SQL.

Prior to the creation of SPIRES, multiple attempts had been made to create accurate linkages between publications and NIH grant numbers. These had all been based on matching of author names. These efforts failed primarily due to the large number of similar names and the fact that in earlier years PubMed stored author names in the format of last name plus first initial.

In 2005, SPIRES became part of the NIH Electronic Research Administration project, with initial funding provided by NIEHS to expand the database to include all NIH institutes. Over time, the process was upgraded to contain all NIH publications from PubMed from 1980 to the present, update the database daily by adding new publications and updating revised publication records, automate the database maintenance process and include a new, improved grant number parser, and provide a scoring system for rating the quality of the grant number to publication matches. In 2008, publication data from the NIH Manuscript System and the NIH Intramural database were added. Throughout its history, the SPIRES system has only been available within NIH.

## **Methods**

### *SPIRES Parsing and Matching Process*

SPIRES uses the following steps to explicitly link articles to grants through matching of grant acknowledgement strings to known grant numbers.

- All publication records in PubMed since 1980 that contain either the full NIH institute abbreviation (e.g., NCI for the National Cancer Institute, NHLBI for the National Heart, Lung and Blood Institute, etc.) or a two letter PHS organization code (e.g., CA for NCI, HL for NHLBI, etc.) are downloaded in bulk from PubMed. In PubMed XML, these values are found in the <Grant><Agency> and <Grant><Acronym> fields, respectively.

- Various data elements, including PubMed IDs (PMID), grant agency, acronym, and number (field <Grant><GrantID>) are extracted and loaded into a relational database.
- The values extracted from <GrantID> are processed through a parsing engine that attempts to decode the string by finding three valid NIH project number components – a two letter organization code(e.g., CA, HL, etc.), the research activity code (e.g., R01, P30) and the 5 or 6 digit serial number. This is not trivial because the <GrantID> values appear in a variety of formats (see Table 1). Grant type prefixes (e.g. 1, 3, 5) and suffixes (support years) are ignored since they are seldom included in grant acknowledgement strings.

**Table 1. Instances of grant number P30 ES 006694 in PubMed. Only 138 of 457 occurrences contain the activity code P30; only 10 contain a suffix (-xx); only 11 contain a prefix (1 or I).**

<GrantID>	count	<GrantID>	count
ES06694	212	1P30 ES 06694	1
ES 06694	54	1P30ES06694	1
ES-06694	42	ES 6694	1
P30-ES-06694	34	ES06694-01	1
P30 ES06694	33	ES06694-02	1
P30-ES06694	18	ES06694-06	1
P30 ES 06694	14	ES-6694	1
P30ES06694	11	P30 ES 006694	1
ES006694	6	P30 ES006694	1
IP30 ES06694	4	P30 ES006694-12	1
P30 ES-06694	3	P30 ES06694-01	1
P30-ES 06694	3	P30ES006694	1
1P30ES06694-01	2	P30-ES006694	1
IP30ES06694	2	P30-ES-006694	1
P-30-ES-06694	2	P30ES06694-02	1
1 P30ES06694-01	1	P30-ES06694-9L	1

- SPIRES then matches what is referred to as the *core* project number against a list of known project numbers dating from 1972 to the present. The core project number consists of the three components mentioned above: the activity code, two letter PHS organization code, and grant serial number. Although this list of core project numbers is maintained in the SPIRES system, a similar list can be constructed from grant data in the RePORTER system.
- As matches are made, the match quality is assessed using a match case scoring system. This scoring system is based on two conditions: the particular components of the NIH project number that could be unambiguously matched, and the number of core project numbers that are either fully or partially matched. Full unambiguous matches can be obtained for grant strings that contain all three components; these are given a score of 5 as shown in Table 2. For cases where the activity code is not available (which is a large fraction of cases, as shown in Table 1) only the organization code and serial number can be matched. In these cases the activity code is inferred from the list of known project numbers; a score of 4 is assigned if only one possible project number can be found.
- Matches (PMID and core project numbers) are placed in a table along with the match case scores. Matches with scores of 4 and 5 have been made publicly available in

RePORTER linkage tables; matches with scores of 3 or less have not been included in the public files.

**Table 2. SPIRES match case scoring criteria.**

<b>Score</b>	<b>Activity code</b>	<b>Org code</b>	<b>Serial number</b>	<b># Core matched</b>
5	match	match	match	One – unambiguous
4	N/A	match	match	One – AC inferred
3	N/A	match	match	Multiple – AC inferred
2	match	match	N/A	No match
1	match	N/A	match	No match

Match case scores of 4 or 5 represent extremely accurate matches that almost invariably are matched to the correct core grant number. Match scores of 1 or 2 indicate cases where not enough data was provided to result in any match. Combined, records with scores of 1 or 2 represent less than 1% of the grant acknowledgement strings, while records with scores of 4 or 5 represent approximately 74% of the data. Records with a match score of 3 are more problematic. A match score of 3 indicates that based on the decoded NIH organization code and serial number, more than one grant number match was found. This is possible because for many years NIH would recycle a series of serial numbers across multiple activity codes. Therefore, cases occur where there is a match against a T32 training grant and an R01 research project grant. In most of these cases, only one of these can be correct, and it is usually not possible from these limited data (without PI names, project titles, etc.) to know which of the two is correct. However, there are other situations where match case 3 records are subsequently upgraded to a score of 4 based on known pairings of activity codes. For example, a pair of case 3 matches involving the activity codes R01 and R37 invariably represents a single research project because the activity code was changed at some point in the history of the project. The SPIRES matching engine now contains a number of these “score upgrade rules” based on known pairings of activity codes, which has resulted in an overall reduction in the number of case 3 matches. Currently, case 3 matches represent 26% of the total SPIRES data.

### *Linking Citation Data*

The final data source used was Scopus, a citation database produced by Elsevier. Scopus is available at the many institutions worldwide that license use of these data. Scopus purports to index all Medline records (and presumably most PubMed records). We matched Scopus records to PubMed records at the article level using a variety of information (e.g., journal, volume, page, publication year, author, title), thus creating a map of PMID to Scopus article ids. These were used to link citation counts to PubMed articles where available.

## **Results and Analysis**

### *Matching Results*

As mentioned above, the ultimate output from the SPIRES linking process is a table of PMID-to-project number matches (with ancillary data such as grant start and end dates) that can be the seed for a variety of additional analyses. Full Medline records for each linked article were also available, and data from those records were used for some of the analysis.

We did not limit our analysis to the match case 4 and 5 records that are publicly available, but included the match case 3 records as well for completeness. We did, however, limit the data in other ways. First, records were limited to those matching NIH grants (as opposed to grants from other PHS agencies such as CDC). Second, records with activity codes starting with N or Z were removed to focus on extramural research grants. Third, although the SPIRES matching process uses a list of project numbers that were assigned both to funded projects and applications that were never funded, we limited this analysis to those matching project numbers where funding was actually awarded. Fourth, records were limited to projects and articles between the years of 1980 and 2009.

One additional processing step was made to further merge records with a match case score of 3. Although, as mentioned above, NIH at times recycled serial numbers across multiple activity codes, this recycling did not occur within Major Activity Code (indicated by the first letter of the activity code). For example, serial numbers might have been recycled between R type and T type grants, but not within R type grants. We thus de-duplicated the set of match records such that each record had a unique combination of “PMID + activity code first letter + organization code + serial number”.

Overall results are shown in Table 3, along with numbers of unique matched grants and unique matched articles associated with each NIH institute. Note, however, that many articles acknowledge multiple grants, both within and across institutes, and also many grants are acknowledged by multiple articles. Thus, we report the number of matches per grant, meaning that on average each grant ‘contributes to’ a certain number of articles rather than that there are a certain number of ‘articles per grant’ produced. Note also that the sum of the number of unique articles for the individual institutes (1,779,893) is higher than the overall number of unique articles (1,386,410) because some articles are associated with multiple institutes. This is also true for the numbers of multi-institute articles.

As shown in Table 3, each grant acknowledged in this dataset has contributed on average to 12.7 articles. Values vary widely by NIH institute, from a low of 4.2 for NINR (National Institute of Nursing Research) to a high of 33.4 for NCRR (National Center for Research Resources). Values should not be strictly compared at the institute level because the different institutes each have very different grant portfolios consisting of different distributions of grant types (i.e., research, equipment, training, etc.), different topic mixes, different dollar amounts, and different goals (i.e., research, clinical, public health, etc.).

The final column of Table 3 shows the percentage of articles that reference grants from multiple institutes. For articles citing NCRR grants, 67% of those articles also cite grants from other NIH institutes. This is not surprising in that NCRR grants include equipment and infrastructure grants along with regional research facilities, and thus are logically overlapped by research grants that use those facilities. By way of comparison, of the institutes with at least 4000 grants, 25% of the articles citing NEI grants also cite grants from other institutes. In general there is a high degree of interlinkage between the NIH institutes – the average fraction of multi-institute articles across the NIH institutes is 41%. However, if one considers the NIH as a whole, and de-duplicates across institutes, the fraction of total articles that reference grants from multiple institutes is only 24% (see row labelled ‘UNIQUE’ in Table 3).

**Table 3. Statistics on grant-article linkages by NIH Institute.**

Institute	matched grant strings	# unique grants	matches/grant	# unique articles	# multi-inst articles	% multi-inst art
NCI	398,691	27,363	14.6	253,398	86,976	34.3%
NHLBI	332,084	23,430	14.2	214,751	74,775	34.8%
NIGMS	273,407	21,142	12.9	209,425	77,120	36.8%
NIAID	203,617	17,976	11.3	140,866	57,839	41.1%
NINDS	180,675	15,911	11.4	130,122	53,834	41.4%
NIDDK	173,001	12,798	13.5	123,433	55,781	45.2%
NIMH	136,871	14,185	9.6	90,475	35,015	38.7%
NCRR	135,394	4,053	33.4	108,163	72,839	67.3%
NICHD	126,680	11,011	11.5	92,452	41,967	45.4%
NIA	96,499	7,148	13.5	66,485	33,461	50.3%
NEI	93,526	6,814	13.7	55,219	13,734	24.9%
NIDA	80,612	6,925	11.6	48,630	19,770	40.7%
NIADDK	68,972	6,704	10.3	50,441	20,620	40.9%
NIEHS	60,824	3,562	17.1	41,559	19,455	46.8%
NIAMS	51,078	4,173	12.2	38,878	18,599	47.8%
NIAAA	39,953	3,881	10.3	25,268	10,421	41.2%
NIDCR	39,127	4,098	9.5	27,891	9,356	33.5%
NIDCD	32,641	3,137	10.4	20,657	5,888	28.5%
NIBIB	11,365	1,416	8.0	9,916	5,403	54.5%
NHGRI	9,707	1,075	9.0	7,760	3,913	50.4%
FIC	9,477	1,995	4.8	8,085	4,508	55.8%
NINR	7,330	1,751	4.2	6,615	1,488	22.5%
NLM	5,964	685	8.7	4,962	1,622	32.7%
Other	5,081	799	6.4	4,442	2,430	54.7%
TOTAL	2,572,576	202,032	12.7	1,779,893	726,814	40.8%
UNIQUE	2,572,576	202,032	12.7	1,386,410	333,331	24.0%

Table 4 shows statistics from the same data broken down by article publication year. The fraction of multi-institute articles has risen over time; the number was relatively steady at around 22% from 1980 through 2000, and has increased steadily since then to a rate of 28.5% in 2009. This increase may simply reflect the dramatic increase in NIH funding over that time period spilling over into increased overlap or synergy between grants. The number of matches per article (or unique grants acknowledged per article) has also risen slightly since 2000, from around 1.7 to 2.1. The changes in matches per article and fraction of multi-institute articles have mirrored each other closely over the past 30 years.

Table 5 gives statistics by initial grant year, and gives rise to a number of interesting observations. Articles referencing multiple grants are counted for each grant they reference, and are counted for the first year the grant was active. For example, an article published in 2000 that references a grant that was active from 1991-1998 appears in the 1991 numbers. Table 5 shows that the average grant durations in years, and the average numbers of papers published per grant have been decreasing slightly over time. These decreases are undoubtedly due in part to the fact that many grants that started years ago are still active; a future look at these same quantities that includes the next several years' data will show larger values. Nevertheless, the decreases cannot

be explained solely by active grants. There appears to be a slight trend toward shorter grant durations. In addition, the data from 1980 to 1989 showed that the average number of articles contributed to by each grant (for those grants that produced articles) was between 14 and 15. These values are larger than the 12.7 that was shown in Table 3. However, the 12.7 value was based on a combination of older and newer grants and is thus a low estimate.

**Table 4. Statistics on matches to grant strings by article publication year.**

<b>Art pub year</b>	<b>matched grant strings</b>	<b># unique grants</b>	<b># unique articles</b>	<b># multi-inst articles</b>	<b>% multi-inst art</b>	<b>match/article</b>
1980	25,686	12,849	14,141	3,090	21.9%	1.82
1981	57,266	20,237	31,164	7,281	23.4%	1.84
1982	58,785	20,929	32,405	7,253	22.4%	1.81
1983	59,545	21,197	32,827	7,388	22.5%	1.81
1984	61,885	22,002	34,405	7,823	22.7%	1.80
1985	63,100	22,742	35,207	7,938	22.6%	1.79
1986	65,620	23,782	36,462	8,481	23.3%	1.80
1987	67,592	24,762	37,749	8,670	23.0%	1.79
1988	68,246	25,473	38,593	9,035	23.4%	1.77
1989	70,840	26,170	40,246	9,317	23.2%	1.76
1990	72,960	26,918	41,569	9,433	22.7%	1.76
1991	74,960	27,467	42,795	9,667	22.6%	1.75
1992	75,709	27,615	43,275	9,847	22.8%	1.75
1993	76,871	28,430	43,931	9,821	22.4%	1.75
1994	78,438	28,723	45,382	9,738	21.5%	1.73
1995	78,935	29,369	45,687	9,629	21.1%	1.73
1996	78,265	29,674	45,413	9,644	21.2%	1.72
1997	78,167	30,182	45,472	9,689	21.3%	1.72
1998	80,630	31,281	46,273	10,141	21.9%	1.74
1999	82,233	32,227	47,314	10,693	22.6%	1.74
2000	84,226	33,541	48,420	10,853	22.4%	1.74
2001	94,575	35,961	49,765	12,239	24.6%	1.90
2002	100,226	38,042	51,513	12,721	24.7%	1.95
2003	105,231	40,037	54,251	13,531	24.9%	1.94
2004	114,715	42,657	58,354	15,084	25.9%	1.97
2005	123,247	45,309	62,368	16,455	26.4%	1.98
2006	132,193	47,253	66,136	17,798	26.9%	2.00
2007	140,738	50,102	70,089	18,791	26.8%	2.01
2008	152,391	53,006	73,787	20,903	28.3%	2.07
2009	149,301	52,521	71,417	20,378	28.5%	2.09
UNIQUE	2,572,576	202,032	1,386,410	333,331	24.0%	1.86

**Table 5. Statistics on matches to grant strings by initial grant year.**

Grant year	# grants	avg duration (yrs)	%grants w/articles	# articles	# art/grant <sup>a</sup>	# Scopus articles	avg cites
1980	6,504	6.64	71.26	67,584	14.58	14,301	34.82
1981	5,453	7.26	73.81	61,559	15.29	13,829	34.69
1982	5,374	7.09	75.81	57,930	14.22	14,352	37.20
1983	7,126	6.61	69.25	71,574	14.50	20,105	35.53
1984	7,367	6.63	69.72	73,688	14.35	21,596	34.93
1985	8,252	6.87	71.28	91,440	15.55	31,484	35.18
1986	7,358	6.61	69.49	77,272	15.11	27,880	35.93
1987	7,458	6.51	66.10	77,671	15.75	29,672	35.89
1988	7,402	6.34	65.44	69,798	14.41	28,691	33.79
1989	7,630	6.12	61.26	66,632	14.26	30,696	38.45
1990	7,443	6.14	57.77	59,894	13.93	31,625	35.94
1991	8,257	6.07	59.59	66,192	13.45	39,113	37.58
1992	7,984	6.21	65.13	68,231	13.12	45,584	38.17
1993	7,108	6.09	61.04	55,383	12.76	41,482	39.03
1994	8,085	6.45	65.41	66,355	12.55	54,631	38.06
1995	7,768	6.30	66.79	64,255	12.39	56,326	37.83
1996	7,399	6.19	70.56	61,111	11.70	54,737	36.84
1997	8,489	6.05	70.64	63,180	10.54	56,317	34.73
1998	8,616	6.17	73.43	65,868	10.41	58,160	31.06
1999	10,133	6.28	73.85	79,067	10.57	68,735	27.51
2000	11,056	6.28	73.66	88,664	10.89	74,037	23.33
2001	11,070	5.92	72.14	73,224	9.17	62,703	19.56
2002	11,639	5.90	69.69	64,944	8.01	54,646	15.73
2003	13,690	5.73	62.59	62,566	7.30	50,197	13.08
2004	14,322	5.35	57.11	50,402	6.16	38,007	10.31
2005	13,668	5.09	54.23	39,314	5.30	26,777	7.26
2006	12,188	4.70	53.63	27,894	4.27	16,185	5.00
2007	13,015	4.69	48.67	20,166	3.18	8,829	2.81
2008	13,782	4.29	31.19	9,227	2.15	2,074	1.44
2009	17,956	3.89	7.85	2,109	1.50	159	3.31

<sup>a</sup> #articles/grant is calculated only for grants with articles.

Several quantities from Table 5 have been plotted in Figure 1. Grant durations and articles per grant have already been mentioned above. The fraction of grants with articles (Figure 1 upper right) acknowledging them was at or above 70% from 1980-1985, dipped to around 60% for the years 1986-1995, with a low of 58% in 1990, and was then once again relatively constant at over 70% from 1996-2003. This suggests that roughly 30% of awarded grants never produce an article. We suspect that there are many valid reasons for this (possible factors may include small exploratory grants, training grants, lack of coverage in PubMed, etc.), but have not looked into this question any further. The rapid decrease in the fraction of grants with publications from 2003 to the present is an artefact of the time lag between the time a grant starts and the first date at which results from that grant are published. The recent tail of this curve suggests that roughly 2/3 of those grants that will end up with publications will have published something within the first three years, and that around 95% of such grants will have produced a publication within 5

years. This is consistent with a recent analysis by Börner et al. (8) showing that around 65% of grants produce an article within 3 years.

Börner et al. (8) also state that “it is generally accepted that publications from earlier years were less likely to cite their grant support.” This could explain the dip in the grants with articles curve at 1990. However, if this is true, then Figure 1 also suggests that articles written prior to 1990 were more likely to cite grant support and that the dip at 1990 could be due to laxness of acknowledgement at that time period that was not historical.

The number of articles published per grant per year (Figure 1 lower left) has been decreasing slowly but steadily since the late 1980's. The slope of the curve decreased monotonically from 1988-1997, and was followed by a flat period of about 4 years where each grant produced roughly 1.7 articles per year. The value decreases again with each initial grant year after 2001, suggesting that long duration grants reach a publication steady state at around 8 years. Although we have not broken these data down by publication year and grant type, the data would support such an analysis which would tell us much more about publication time lags. We leave this work to a future study.

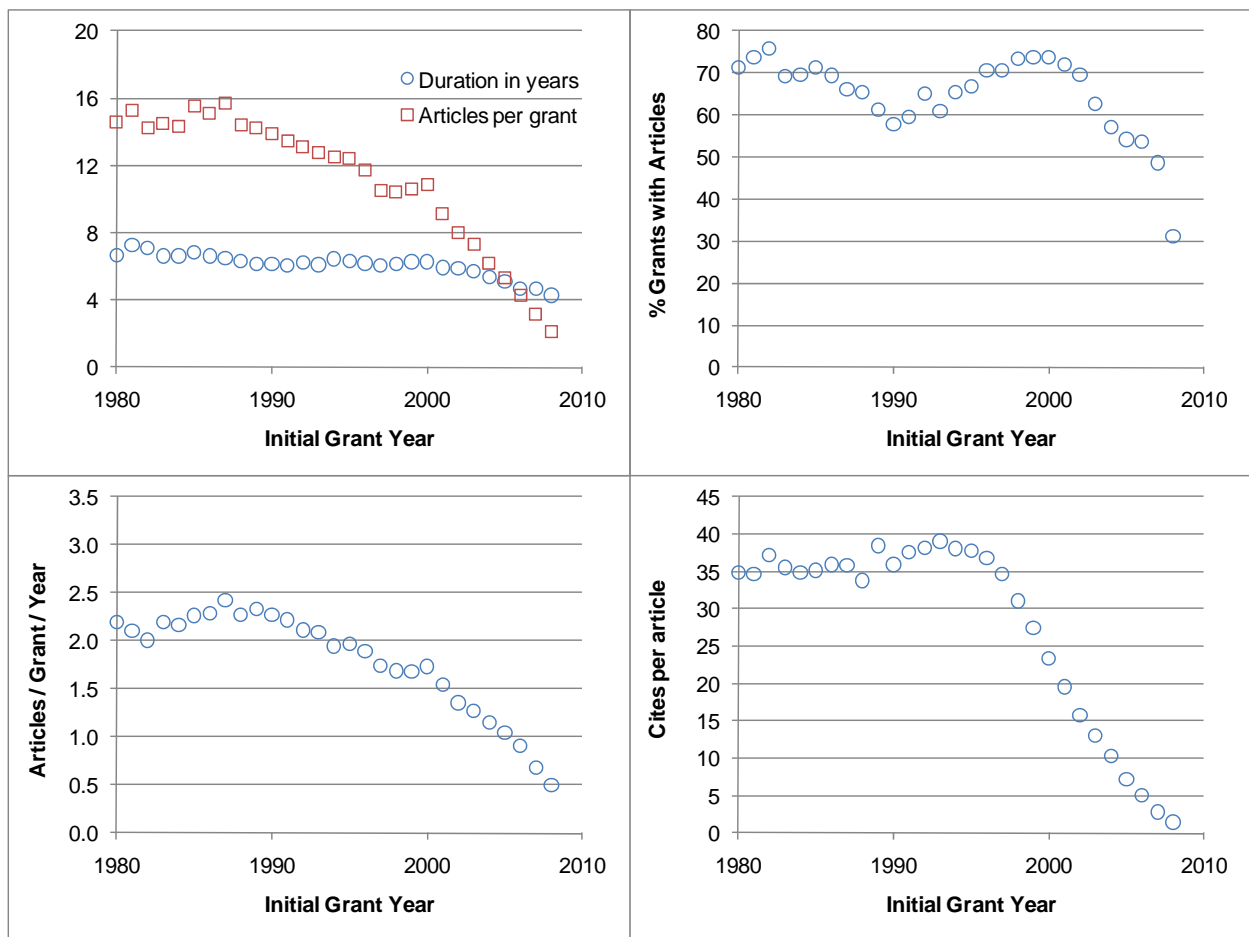


Figure 1. Time series of grant-related quantities by initial grant year.

The final quantity shown in Figure 1 is the average number of citations per article as a function of initial grant year. Citation counts were added to each record by finding the article in Scopus, as mentioned earlier. Citation counts were based on the 1996-2008 Scopus annual files, and are thus current as of the end of 2008. This quantity is relatively constant through the 1980s at around 35 citations per article, rising to a peak of 39 citations per article in 1993, and then dropping off almost linearly for grant years from 1997 to 2008. This drop does not indicate that the articles from more recent grants will be less cited than those citing earlier grants, but rather reflects the fact that newer grants are accompanied by newer articles that have had less time to be cited.

### PI-author Order

We are also interested in principal investigators (PI) and where they appear as authors of articles. The author orders for the PI's for each of the grant-article matches for papers published since 1996 were found in the PubMed data and percentages are shown in Figure 2. Over the entire time period, it can be clearly seen that if the PI was listed as an author on the article, it was more often as last author (38.6%) than as the first (10.7%) or a middle (18.0%) author. This correlates well with the historical convention in biomedical publication for the leading author to be listed last. The largest shift in authorship over the time period was a decrease in first authorship (from 12% to 9%) and corresponding increase in middle authorship (from 15% to 20%).

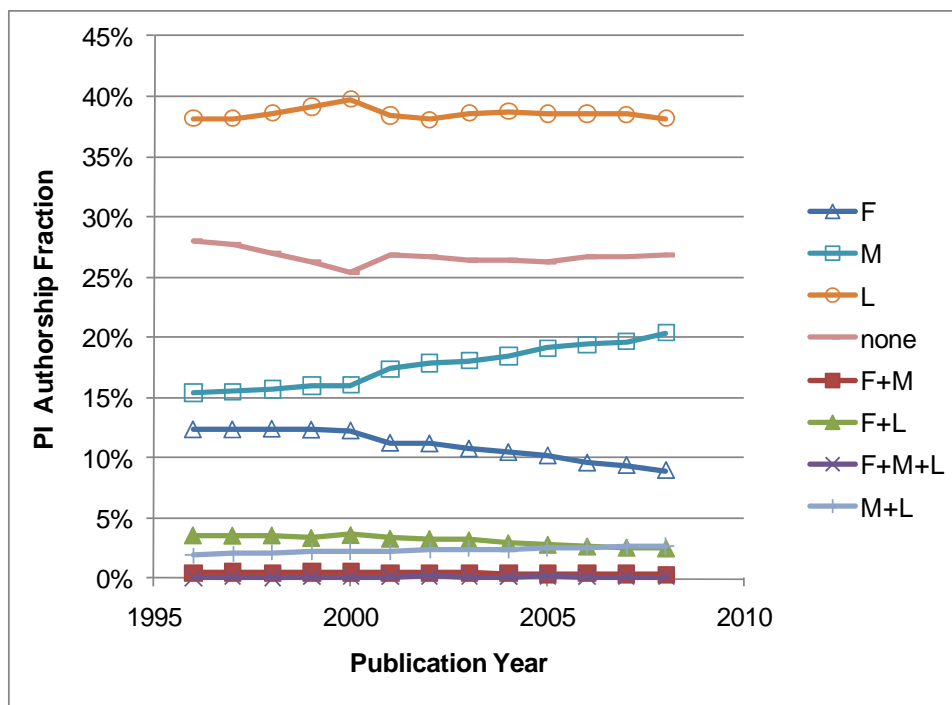


Figure 2. Author order for principal investigators from the grant-to-article matches by publication year. F = first author, M = middle author, L = last author.

Figure 2 also shows that the PI was not listed as an author on 27% of the matched articles. This is not surprising in that many grants are certainly large enough that not all work is directly overseen or published by the PI. In this analysis only overall PI's were considered. If component PI's had been considered, the fraction of articles not authored by a PI would undoubtedly have

been lower. This fact also shows why any attempted matching of grants to articles based on matching of authors to PI's cannot be fully successful. Note also that combinations of author orders (e.g. first+last) are also shown in Figure 2. These reflect cases where multiple people were PI's on the same grant at different times, and more than one of those PI's co-authored the article. Since we do not know the exact time lag between publication and when the particular work reported on in article was funded, we have not limited each grant to a single PI, but report all possible combinations.

### *Research Support Type and Impact*

In addition to the grant string information, PubMed contains more general information about research support. The article type field, in addition to listing article types, lists the following funding-related tags:

*Research Support, N.I.H., Extramural*

*Research Support, N.I.H., Intramural*

*Research Support, U.S. Gov't, P.H.S.*

*Research Support, U.S. Gov't, Non-P.H.S.*

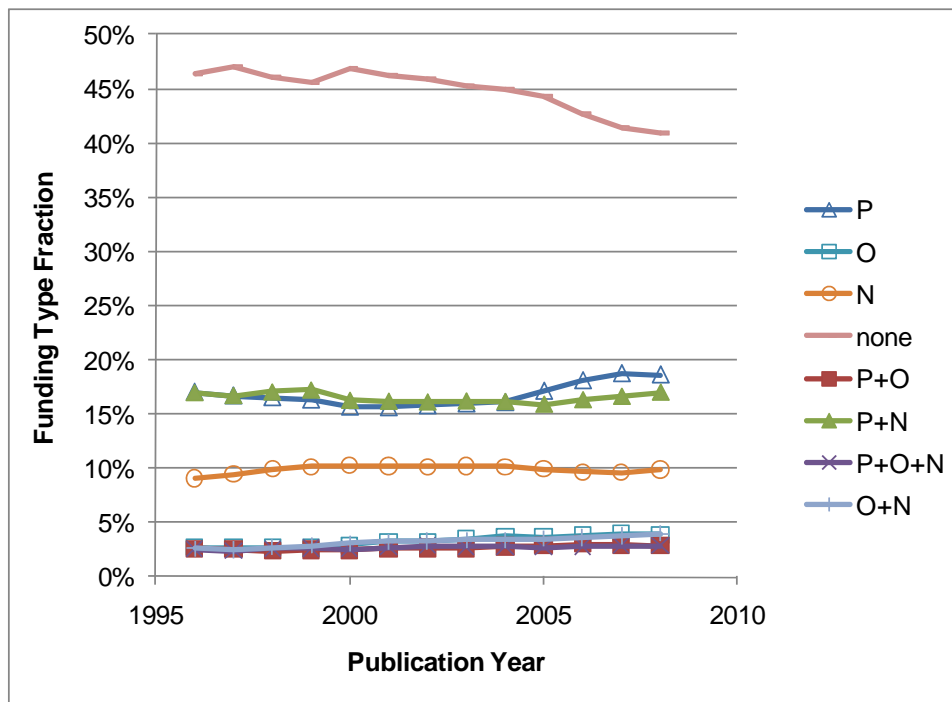
*Research Support, Non-U.S. Gov't*

Most other *Research Support* tag variations are typographical errors that can be mapped back to these five tags. Since NIH is the largest grant awarding arm within the U.S. Health and Human Services agency (and thus the U.S. Public Health Service), we have combined the first three of these tags (*N.I.H., Extramural, N.I.H., Intramural,* and *U.S. Gov't, P.H.S.*) into a single group denoting funding from the U.S. Public Health Service (PHS), listed as (P) in the following discussion. *U.S. Gov't, Non-P.H.S.* (O) includes other U.S. funding agencies such as the NSF, DOE, NASA, etc. *Non-U.S. Gov't* (N) includes government funding from other countries as well as industry, non-profits, foundations, and university endowments from around the world. Some articles acknowledge multiple funding types and thus have multiple tags.

Figure 3 shows the fractional distribution of 2,167,393 PubMed articles by funding type and publication year. This analysis has been limited to those articles that can be justified as having a US address. This includes articles with U.S. first author addresses in PubMed and those additional articles to which NIH grant strings have been matched. The latter criterion assumes that NIH funding goes primarily to US institutions and addresses. In addition to the numbers listed in Table 5, there are 5,229,721 PubMed articles over the same time period that do not reference a U.S. first author address and that do not reference a matched NIH grant string. We assume that these articles represent contributions from non-U.S. countries. Of these, 1,740,170 articles (33%) have a *Research Support, Non-U.S. Gov't* funding tag, while the remaining 3,489,551 articles have no acknowledgement of a funding source.

Figure 3 shows that the fraction of U.S. articles with no funding acknowledgement has decreased from 47% to about 41% over the past decade. The fraction of articles funded exclusively by the US PHS is currently around 18%, while another 17% are joint funded by the US PHS and non-US government sources (P+N). The total contribution of the US PHS to the U.S. biomedical publication output, summed over the four types (P, P+O, P+N, P+O+N), is currently just over 41%. The fraction of articles funded exclusively by non-US government sources is 10% of the

total. Other funding sources and combinations of funding sources each account for less than 4% of the total.



**Figure 3. Funding type distribution by year. P = PHS funding, O = other US government funding, N = non-US government funding.**

It is difficult to know if the grants indexed in PubMed comprise the majority of the actual grant-to-article relationships or not. Figure 3 suggests that around 45% of the U.S. articles indexed in PubMed since 1996 have no acknowledgement of funding. Lewison (10) reported that 46% of nearly 13,000 UK gastroenterology papers had no acknowledged funding source, but that 85% of those were from UK National Health Service hospitals, and thus had an implied funding source by association. Further, Lewison, Dawson, and Anderson (18) found that while 39% of papers in the UK Research Outputs Database did not contain acknowledgements of funding, 7/8 of those could not be expected to have them. Lewison's studies thus suggest that a fraction of 40-45% of articles without acknowledgements is reasonable and does not suggest that authors are ignoring the proper acknowledgement of grant funding from public sources.

By contrast, Cronin and Franks (19) examined over 1000 articles from the journal *Cell* and found that over 90% of them had financial acknowledgements. We note that of the 840,942 articles associated with NIH or PHS funding types in Figure 3, over 85% of them are in the grant-article match list. This leaves around 15% of the articles noted to have received NIH or PHS funding, but for which the actual grant information was not indexed; these could be considered as false negatives. Taken in total, these studies suggest that biomedical researchers do, for the most part, acknowledge government funding in a consistent and representative, but not totally complete, manner.

Combining the funding information tags and grant information strings from PubMed records with citation counts from Scopus allows us to examine the impact of different funding types in the biomedical research area. Matching of Scopus to PubMed records produced positive one-to-one matches for nearly 96% of the records acknowledging funding from Figure 3, and for 88% of the records in the “none” category from Figure 3. The no acknowledgement category likely has a lower matching rate to Scopus records because this category will include a higher fraction of article types that do not represent technical advances (e.g., editorials, book reviews, etc.) and thus are far less likely to acknowledge funding sources.

Average citation counts per article were calculated using Scopus data for the articles in each of the funding categories and publication years from Figure 3. Figure 4 shows that articles acknowledging PHS funding (P) receive twice as many citations on average as those that have no funding acknowledgement (none). This can be taken as evidence that the PHS system is working in terms of funding work that is of higher than average impact. However, the data also show that impact increases with the number of different types of funding sources acknowledged. Citation counts rise by about 10% if funding from another U.S. government agency is acknowledged in addition to that from PHS (P+O). However, they rise even further, by 40% or more, if funding from a non-U.S. government source (e.g., a foundation, non-profit, or society) is acknowledged in addition to PHS funding (P+N). This is consistent with previous results by Boyack (14) showing that articles jointly funded by the U.S. PHS and non-U.S. government sources have higher impact than those funded by the U.S. PHS alone. The highest impact articles on average are those that acknowledge funding from three different types of funding sources (P+O+N). These findings are robust in that each data point shown in Figures 3 and 4 is based on a minimum of 3,000 articles per category-year.

The analysis of Figure 4 includes all PubMed articles to which we could attach Scopus citation counts. Figure 5 shows the corresponding results if the analysis is limited to the 690,325 unique PubMed articles in the grant-article match list. Although the citation numbers for the more limited set are about 5% lower on average than those in Figure 4, the ordering of the curves is maintained, further suggesting that the findings mentioned above regarding impact and funding types are robust. The 5% decrease in average citations from the larger set suggests that the 10% of articles noted to have received PHS funding in the article type tags but that are missing grant information strings may be preferentially from higher impact journals.

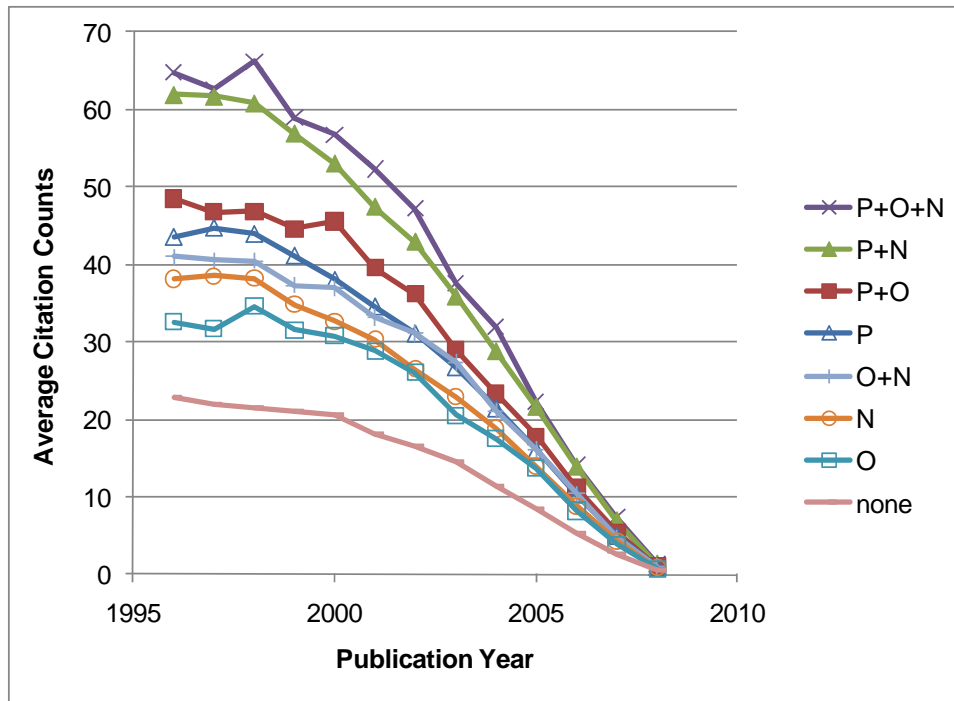


Figure 4. Average citation counts per paper as a function of funding type using the publication groupings of Figure 3. P = PHS funding, O = other US government funding, N = non-US government funding.

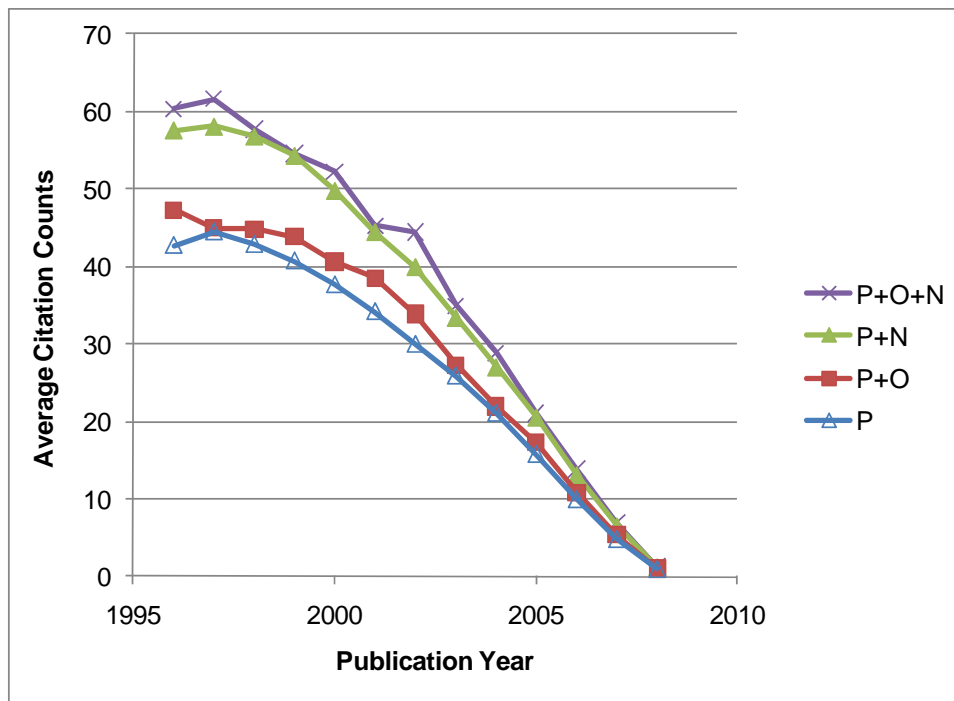


Figure 5. Average citation counts per paper as a function of funding type using the only those publications that are matched by NIH grant strings in PubMed records. P = PHS funding, O = other US government funding, N = non-US government funding.

## Limitations and Suggestions

There are, of course, limitations to the data described here that limit the depth of the analyses that can be undertaken. Although it has not been mentioned until now, there are certainly false positive matches between grants and articles in these data. The fraction of false positives is unknown, but is expected to be small given that we have excluded all matches to unawarded grants. One other type of potentially false positive match is one where the article appears to have been published before the grant was awarded. These amount to roughly 1% of the data, and can occur due to a variety of errors in either the grant data or parsing steps. In some cases they may also reflect a researcher incorrectly acknowledging a grant that is due to start, but has not yet. We have not excluded these cases since we expect that the majority of them can be explained with sufficient investigation.

Another key limitation is that despite PubMed's broad and increasing coverage of the literature surrounding the life sciences and biomedicine, it does not index all articles that acknowledge NIH grants. Thus, article counts and publication rates derived from this analysis should be considered lower bounds, and may be more underspecified in boundary areas with many journals and conference proceedings that are not indexed by PubMed (e.g., informatics) than in core areas (e.g., oncology). Another limitation is that the grants acknowledgements in PubMed do not, for the most part, include suffix information, and thus cannot be linked to individual grant years. Thus, time lags must be either assumed or ignored.

In the analyses here we have not made use of any funding data. With the addition of such data a variety of detailed input-output studies could be done. For example, time histories showing funding, publication counts, and citation counts could be constructed for individual grants, or for groups of grants by agency, program, funded institution, PI, etc.

The data that have been recently made available through NIH RePORTER and ExPORTER, based on the matching of grants-to-articles using the SPIRES system, are a great resource from which to pursue input-output studies of biomedical fields in the United States. Similar data exist for the UK in the Research Outputs Database. However, we note that no similar widely accessible data exist outside the biomedical area. For example, such data linking grants and articles are lacking for the U.S. NSF and other agencies. We hope that such data will be made more widely available for other agencies in the future.

## Summary

This work reports on NIH activities that have provided grant-to-article linkage data that are now available at the NIH RePORTER and ExPORTER websites. In addition, we have performed high level analyses of NIH data input, output, and impact data at a large scale, and have shown high level features of those data that have not been reported before. Although key findings have been given in each section of the paper in context, we summarize them here for easy access.

- The fraction of articles reflecting synergies between multiple NIH institutes is increasing over time; 29% of current articles acknowledge grants from multiple institutes. Synergy between multiple grants from the same institute is also increasing over time.
- Grant durations are decreasing slightly over time.
- The median time lag to publication from a new grant is 3 years. Most grants that will produce an article will have done so within 5 years.

- Each grant contributes to roughly 1.7 articles per year, averaged over all grant types.
- Principal investigators are not listed as authors on roughly 27% of articles that acknowledge their grants.
- Articles acknowledging PHS funding receive on average twice as many citations as articles that acknowledge no funding source whatsoever.
- Articles acknowledging both PHS funding and a non-U.S. government funding source (e.g., foundation, non-profit, society) receive on average 40% more citations than those acknowledging PHS funding sources alone.

Although we have stayed with a high level view in the analyses reported here, the data support analysis at much finer-grained levels – by institute, grant type, etc. – and even to the level of individual grants. We plan to pursue additional studies using these data, and encourage others to do the same.

### **Acknowledgments**

Associated contributions: We acknowledge Dr. Bennett van Houten, former chief of the Program Analysis Branch at NIEHS for his work and support at the genesis of the SPIRES project, and Ying Gao and F.O. Finch for their work on the SPIRES parsing and matching engines.

Competing interests: Although author Boyack is employed by SciTech Strategies, Inc., no patent or product based on this work is under development.

Funding: This work was partially supported by NSF award SBE-0738111. The sponsors had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

### **References**

1. Boyack KW. Linking grants to articles: Characterization of NIH grant information indexed in Medline. 12th International Conference of the International Society for Scientometrics and Informetrics. 2009 July 15-17, 2009:730-41.
2. McAllister PR, Wagner DA. Relationship between R&D expenditures and publication output for U.S. colleges and universities. *Research in Higher Education*. 1981;15(1):3-30.
3. McAllister PR, Narin F. Characterization of the research papers of U.S. medical schools. *Journal of the American Society for Information Science*. 1983;34(2):123-31.
4. Bourke P, Butler L. The efficacy of different modes of funding research: Perspectives from Australian data on the biological sciences. *Research Policy*. 1999;28:489-99.
5. Butler L. Revisiting bibliometric issues using new empirical data. *Research Evaluation*. 2001;10(1):59-65.
6. Jimenez-Contreras E, Moya-Anegón F, Lopez ED. The evolution of research activity in Spain: The impact of the National Commission for the Evaluation of Research Activity (CNEAI). *Research Policy*. 2003;32:123-42.

7. Kwan P, Johnston J, Fung AYK, Chong DSY, Collins RA, Lo SV. A systematic evaluation of payback of publicly funded health and health services research in Hong Kong. *BMC Health Services Research*. 2007;7:121.
8. Börner K, Ma N, Biberstine JR, Wagner RM, Berhane R, Jiang H, et al. Introducing the Science of Science (Sci2) Tool to the Reporting Branch, Office of Extramural Research/Office of the Director, National Institutes of Health. Science of Science Measurement Workshop; December 2-3, 2010; Washington, DC2010.
9. Hicks D, Kroll P, Narin F, Thomas P, Ruegg R, Tomizawa H, et al. Quantitative methods of research evaluation used by the U.S. federal government: Second Theory-Oriented Research Group, National Institute of Science and Technology Policy (NISTEP), Japan2002.
10. Lewison G. Gastroenterology research in the United Kingdom: Funding sources and impact. *Gut*. 1998;43:288-93.
11. Lewison G, Devey ME. Bibliometrics methods for the evaluation of arthritis research. *Rheumatology*. 1999;38:13-20.
12. Lewison G, Grant J, Jansen P. International gastroenterology research: Subject areas, impact, and funding. *Gut*. 2001;49:295-302.
13. Boyack KW, Börner K. Indicator-assisted evaluation and funding of research: Visualizing the influence of grants on the number and citation counts of research papers. *Journal of the American Society for Information Science and Technology*. 2003;54(5):447-61.
14. Boyack KW. Mapping knowledge domains: Characterizing PNAS. *Proceedings of the National Academy of Sciences*. 2004;101(Suppl. 1):5192-9.
15. Boyack KW, Börner K, Klavans R. Mapping the structure and evolution of chemistry research. *Scientometrics*. 2009;79(1):45-60.
16. Zhao D. Characteristics and impact of grant-funded research: A case study of the library and information science field. *Scientometrics*. 2010;84:293-306.
17. Lyubarova R, Itagaki BK, Itagaki MW. The impact of National Institutes of Health funding on U.S. cardiovascular disease research. *PLoS ONE*. 2009;4(7):e6425.
18. Lewison G, Dawson G, Anderson J. The behaviour of biomedical scientific authors in acknowledging their funding sources. 5th International Conference of the International Society for Scientometrics and Informetrics; June 7-10, 1995; River Forest, IL1995. p. 255-63.
19. Cronin B, Franks S. Trading cultures: Resource mobilization and service rendering in the life sciences as revealed in the journal article's paratext. *Journal of the American Society for Information Science and Technology*. 2006;57(14):1909-18.