Linking Grants to Articles: Characterization of NIH Grant Information Indexed in Medline

Kevin W. Boyack

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

Abstract
Existing input-output studies designed to inform science policy have been largely based on aggregated and unlinked funding and literature data rather than actual grant-article linkages. Linked grant-article data are lacking in most cases, and have not been cleaned and systematically studied on a large scale in cases where they do exist. This study presents the first large scale cleaning and linking of grant string information from the Medline database to the grant numbers in grant databases. Distributions and analyses derived from 577,941 individual grant-article linkages are reported. Limitations and suggestions for future research are also described.

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 are lacking. One place where these data do exist is the Medline database, which has been indexing selected public health system related grant information since 1981.[1]

The fact that these Medline references to grant numbers exist does not, however, mean that they have been systematically used for research evaluation. In fact, the opposite is true. Although these data exist, they are not standardized, and have been difficult to link to grant databases. In addition, many grant databases (including NIH's CRISP database) do not contain dollar amounts for the grants. The significant efforts required to clean and link such data along with the lack of funding information have contrived to limit the number and scope of input-output studies.

The US National Institutes of Health (NIH) has recently made a significant investment in the SPIRES and IMPAC-II systems, which promise to link grants, funding amounts, and article information. However, these databases are currently for internal NIH use only. Since it is unclear if the public will gain access to these data in the near future, we have decided to embark upon our own linking of grant and article data, and report on that effort here.

The balance of this paper will proceed as follows. First, relevant literature on linking of grant and article data will be briefly reviewed. Next, our data sources will be described, along with the logic that was used to clean and link the data sources. A short study of the characteristics of the linked data will then be presented. The paper will conclude with a discussion of benefits, limitations, and suggestions for future work.

Background
Perhaps the most comprehensive input-output studies were done in the 1980's by CHI Research. For example, McAllister and colleagues studied the relationship between R&D

[1] http://www.nlm.nih.gov/bsd/funding_support.html

Proceedings of ISSI 2009, pages 730–741.
Edited by B. Larsen and J. Leta.

expenditures and publication outputs for U.S. colleges and universities (1) and U.S. medical schools (2) on a large scale using aggregated funding amounts, and publication and citation counts. Bourke and Butler (3) 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 (4) 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., 5, 6).

Far fewer studies are available in which actual linking of grant data to individual articles has been reported. CHI Research 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 (7). However, neither their grant-article linkage data nor their reports to NIH are readily available. Lewison and colleagues (8-10) 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 (11), and showed that citation impact increased with grant size (12). They also showed funding profiles for NIH and NSF on a map of science (13), using grant data from 1999 and article data from 2002 linked through author/PI and institution.

Data and Methodology

Data

Three separate data sources were used for this study: NIH grant data from CRISP, NIH grant data from the (now defunct) RaDiUS database, and article data from Medline, as shown in Figure 1.

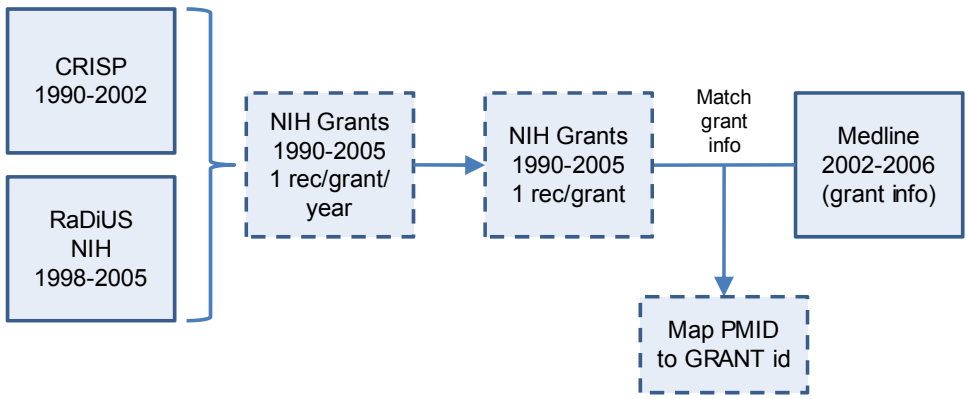

Figure 1. Schematic showing data sources and the general order of processing. The CRISP, RaDiUS, and Medline data were linked together on grant numbers.

The CRISP data were comprised of 525,850 records covering fiscal years 1990-2002, where 2002 contained only partial year data.[2] These data contain most essential grant information,

[2] CRISP data were obtained from Indiana University, and are the same NIH data which populated the Scholarly Database at IU through 2008. CRISP data are available at http://crisp.cit.nih.gov/.

including grant number, PI, institution, dates, title, abstract, and thesaurus terms. However, they do not contain funding amounts.

RaDiUS (R&D in the United States) was a database maintained by the RAND Corporation for many years. Data from many US agencies were collected and placed in a standard format and made available through a web interface to users. Although RaDiUS is no longer available, it appears to have been replaced by USAspending.gov, where similar data are now available for download.[3] NIH data from RaDiUS were downloaded by the author in 2006, and are used in this study. These data are comprised of 396,577 records covering fiscal years 1998-2005. The RaDiUS data contain nearly all of the information available in CRISP data, and contain funding information as well. Data in both databases are listed by fiscal year; thus grants spanning multiple fiscal years have multiple entries.

The Medline data were comprised of 2,826,380 distinct records covering publication years 2002-2006, and were downloaded as tagged record files in early 2007. These Medline data are not a complete set for all years listed; comparison with numbers available online suggests that these data contain virtually all records for 2002-04, but only 87% of what is now available for years 2005-06.[4] As mentioned above, Medline does index public health system related grant information. A total of 647,888 separate pieces of grant information, containing grant number information listed in indexed articles, were listed in these Medline data. Of these, 616,562 records appear to relate to NIH and its institutes.

Cleaning and Linking Methodology

One would think that with grant numbers available in all three of these databases, linking of articles to grants would be an easy task. However, the grant numbers appear in the three databases with a large number of format variations; thus, cleaning of each database and identification of the component parts of the grant number in each is required to match and merge the data. An example of the various grant string formats is shown in Table 1.

Table 1. Grant string variations in the three databases for grant P30ES006694.

CRISP	RaDiUS	Medline
		1P30 ES 06694/ES/NIEHS
1P30ES006694-01		ES 06694/ES/NIEHS
5P30ES006694-02		ES006694/ES/NIEHS
5P30ES006694-03		ES06694/ES/NIEHS
5P30ES006694-04		ES-06694/ES/NIEHS
2P30ES006694-06		ES06694-06/ES/NIEHS
3P30ES006694-06S1		P30 ES 06694/ES/NIEHS
5P30ES006694-07		P30 ES06694/ES/NIEHS
3P30ES006694-07S1		P30 ES-06694/ES/NIEHS
3P30ES006694-07S2		P30ES006694/ES/NIEHS
3P30ES006694-07S3		P30ES06694/ES/NIEHS
5P30ES006694-08		P30-ES06694/ES/NIEHS
3P30ES006694-08S1	P30ES006694	P30-ES-06694/ES/NIEHS
5P30ES006694-09	P30ES06694	P30-ES06694-9L/ES/NIEH

NIH grant numbers consist of two main parts, a two character abbreviation indicating the institute (e.g., ES, abbreviation for NIEHS)[5], and a six character integer (e.g., 006694). NIH grant numbers also have a three character prefix designating the grant type (e.g., P30 or R01). The grant strings in CRISP are the most standardized of the set, each with 12 characters followed by a hyphen and a suffix that is typically either two or four characters long. This 12

[3] http://www.usaspending.gov/

[4] Compare with publication counts at http://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html.

[5] Abbreviations and associated institutes are listed at http://www.nlm.nih.gov/bsd/grant_acronym.html.

character string can be decomposed into the three character grant type (characters 2-4), the two character abbreviation indicating the institute (characters 5-6), and the six character integer (characters 7-12). Different suffixes are used for different years, and four character suffixes with an "S" in the third position indicate subawards.

The grant strings in RaDiUS are either 10 or 11 characters long, and contain the grant type, two-character institute abbreviation, and integer. The grant information strings in the Medline article records are by far the most diverse and least standardized.[6] In general they contain three parts[7], separated by "/" characters, where the first section contains the grant type, two-character abbreviation, and integer, the second section is the two-character institute abbreviation, and the third section is the full institute abbreviation. However, as shown in Table 1, the grant type (e.g., P30) is missing in about half the entries, the integer can have differing numbers of leading zeroes, and there may or may not be suffix information. These variations are typical for the full set of grant information strings; there are many examples that are far more obtuse than those shown here. In addition, there may be additional separators (spaces, commas, dots, hyphens), additional two-character abbreviations, and additional forward slashes, all of which make a complex parsing logic a necessity. To summarize, the number and type of variants in grant strings across three databases is what has precluded people from any large scale study of NIH grants and their impacts, and is what makes the current work to merge these data sources a timely one.

Figure 1 shows the order in which the data were processed. First, the three grant number components (type, abbreviation, and integer) were extracted from both the CRISP and RaDiUS data. These three components, along with the fiscal year (present in both databases) were used to merge the two data sources into a common schema, with some fields coming from each data source. The schema is far too large (over 50 fields) to present here. Subawards from the CRISP data were not included in the merging algorithm to avoid duplication of funding amounts (from RaDiUS) within a single fiscal year. Results of the CRISP:RaDiUS merge are shown in Table 2, where the combined data contained one record per grant per year (with the exception of subawards). Of the 185,301 combined records shown in Table 2, CRISP and RaDiUS returned the same principal investigator name for 96% of the matches. The remaining 4% can be accounted for by null PI entries in the sources, or by changes in the PI name occurring in different years in the two sources.

The merged dataset was then reduced to a *grant summary* table containing a single record per grant number using an 11 character representation of the grant number (e.g. P30ES006694). This table contained 197,716 separate grant numbers, comprising a relatively complete set of NIH grants that were active during the years 1990-2005.

Table 2. Statistics on the merging of CRISP and RaDiUS data (numbers of records).

FY	TOTAL	CRISP	JOINED	RaDiUS
1990-1997	320,021	320,021		
1998-2001	204,373	18,987 [a]	161,369	24,017 [b]
2002 [c]	51,353	1,541	23,932	25,880
2003-2005	161,379			161,379
ALL	737,126	340,549	185,301	211,276

[a] unmatched CRISP records are typically subawards
[b] unmatched RaDiUS records are typically zero cost extensions
[c] 2002 is listed separately because our CRISP data were only partial year

[6] "*Authors and publishers present the grant numbers in a variety of formats. NLM does not attempt to standardize the format of the published grant numbers. The data are only as accurate as the original source.*" from http://www.nlm.nih.gov/pubs/techbull/mj06/mj06_grant_numbers.html

[7] Starting in 2009, country names have been added to the end of the grant strings in Medline.

The final steps of this process were to extract grant number components from the grant information strings in the Medline data and then match those components to those in the table of grant numbers. Parsing of grant information strings contained the following general steps:

1) All characters following the final "/" were stripped and placed in a field designating institute. The "/" was dropped from the string.
2) The two characters following the final remaining "/" were placed in a field designating the two character abbreviation. The "/" was dropped from the string.
3) The location of the two character code extracted in step (2) was found in the remaining portion of the grant string. All characters before the code were placed in a field designating the grant type. All characters after the code were placed in a text field that was assumed to contain the integer (and perhaps additional information).
4) Leading integers and leading spaces, hyphens, or slashes were removed from the grant type field. Any "O" characters where replaced by zeroes (e.g. R01 in place of RO1).
5) If any "/" characters remained in the integer field, they typically were followed by an additional two-character abbreviation. These additional two-character codes were placed in a separate field as an alternate code that could be used for later matching.
6) If the first character of the integer field was one of the letters (R,P,U,K,L), it typically denoted that the grant type had ended up in the integer field. These grant types were located, stripped, and placed in the grant type field.
7) Leading hyphens and slashes were removed from the integer field. All periods, commas, and extra spaces were also removed.
8) If the first two characters of the integer field were "A1", they were replaced by "AI".
9) Step (6) was repeated on the integer field. Leading hyphens and spaces were also removed.
10) The integer field was searched for hyphens. All characters after a hyphen were assumed to be a grant suffix and were removed. Any leading zeroes were also removed.
11) If the number of characters in the integer field was greater than six, alternate five character integers were constructed (e.g. characters 1-5, characters 2-6, etc.) to be used as possible matches.

At this point each Medline grant information record had been separated into a grant type, one or two abbreviation codes, and one or two integers. These records were then matched against the grant summary table. Matches were assigned a score using the following logic:

- Score = 1.0 if all three fields (grant type, two-character abbreviation, and integer) were unambiguously matched, where unambiguous means that one and only one grant number could be matched with the information parsed from the Medline grant string.
- Score = 0.9 if the grant type was missing in the Medline grant string, but the two-character abbreviation and integer were unambiguously matched, where unambiguous means that only one grant type was found.
- Score = 0.8 if the grant type was missing in the Medline grant string, and the two-character code and integer were matched to two entries in the grant summary table, with two different grant types.
- Score = 0.7 if the grant type was missing in the Medline grant string, and the two-character code and integer were matched to three or more entries in the grant summary table, each with different grant types.

Results and Analysis

The ultimate outputs from the entire cleaning and linking process are represented by the three dashed boxes shown in Figure 1: the merged CRISP/RaDiUS information described in Table

2, a table of unique grant numbers and durations, and a list of PubMed id to grant number matches that can be the seed for a variety of additional analyses.

Overall results from matching of the grant information strings to actual grant numbers listed in the grant summary table are shown in Table 3. Almost 94% of all of the grant strings were matched, with the large majority of those matched unambiguously. Note that Table 3 only contains statistics for matches to grants assigned to NIH institutes. The Medline grant strings also contained references to grants from other agencies and institutions (e.g. PHS, CDC, Wellcome Trust, etc.) that could not be matched because the grant information for those institutions was not available.

Table 3. Statistics on matches to grant strings in Medline (2002-2006) by NIH Institute. Numbers of unambiguous (only one grant type) and ambiguous (multiple grant type) matches are shown.

Institute	possible matches	% matched	unambig	ambig	no match	# unique grants	# unique articles	% multi-inst arts
NCI	93,897	92.0%	82,539	3,883	7,475	11,314	51,521	36.1%
NHLBI	82,525	93.5%	72,172	4,952	5,401	9,600	41,901	41.6%
NIGMS	58,749	95.3%	49,886	6,103	2,760	8,421	43,640	35.3%
NIDDK	52,390	95.4%	45,857	4,125	2,408	6,987	31,405	49.5%
NIAID	51,953	92.5%	43,087	4,976	3,890	8,348	30,149	42.8%
NINDS	37,054	94.9%	32,774	2,377	1,903	5,954	24,467	46.7%
NIMH	36,859	93.8%	31,392	3,186	2,281	6,092	21,401	40.0%
NCRR	31,373	95.1%	27,601	2,233	1,539	1,470	24,271	72.7%
NIA	27,424	93.9%	24,104	1,659	1,661	3,369	16,489	50.4%
NICHD	26,691	93.1%	22,596	2,248	1,847	3,975	17,041	49.3%
NIDA	21,145	95.3%	18,234	1,924	987	3,394	11,812	43.1%
NEI	18,835	95.6%	16,183	1,824	828	2,604	10,610	27.8%
NIEHS	16,220	94.3%	14,280	1,008	932	1,540	10,064	52.1%
NIAMS	15,401	93.4%	13,522	856	1,023	2,236	9,931	50.3%
NIAAA	10,643	94.3%	8,885	1,154	604	1,700	5,973	43.3%
NIDCD	9,200	95.0%	7,706	1,033	461	1,916	5,830	29.9%
NIDCR	9,094	94.3%	8,025	554	515	1,536	5,922	38.6%
NIBIB	4,381	95.5%	4,124	60	197	727	3,415	56.5%
FIC	2,813	87.7%	2,404	64	345	547	2,178	54.1%
NINR	2,661	88.2%	2,314	32	315	784	1,996	23.2%
NHGRI	2,559	93.2%	2,098	286	175	492	2,023	50.3%
NCCAM	1,724	93.0%	1,580	23	121	331	1,335	48.5%
NLM	1,609	85.6%	1,362	15	232	232	1,109	35.1%
NCMHHD	559	74.2%	413	2	144	65	373	62.5%
WHI	205	97.1%	199	0	6	41	35	40.0%
Others	598	4.5%	27	0	571	15	26	46.2%
Totals	616,562	93.7%	533,364	44,577	38,621	83,690	374,917	44.0%

Numbers of unique grants and articles associated with each NIH institute are also shown in Table 1. Note, however, that the numbers of articles per grant cannot be obtained from the ratio of the two numbers because the multiple grant-to-article relationships (in both directions) are not accounted for. Note also that the total number of articles listed (374,917) is higher than the number of unique articles (283,413) because some articles are associated with multiple institutes.

The final column of Table 1 shows the percentage of articles that reference grants from multiple institutes. From this standpoint, NCRR is the most highly interlinked of all the institutes, with nearly 73% of its articles also referencing grants from other institutes. Of the larger institutes NEI is the most insular with only 28% multi-institute articles. In general there is a high degree of interlinkage between the NIH institutes – the overall fraction of multi-institute articles across all of NIH is 44%. These grant co-occurrences have been used to

generate a map of the NIH institutes from the perspective of potential topical overlaps. Raw co-occurrence values were used as edge weights, resulting in the map shown in Figure 2.

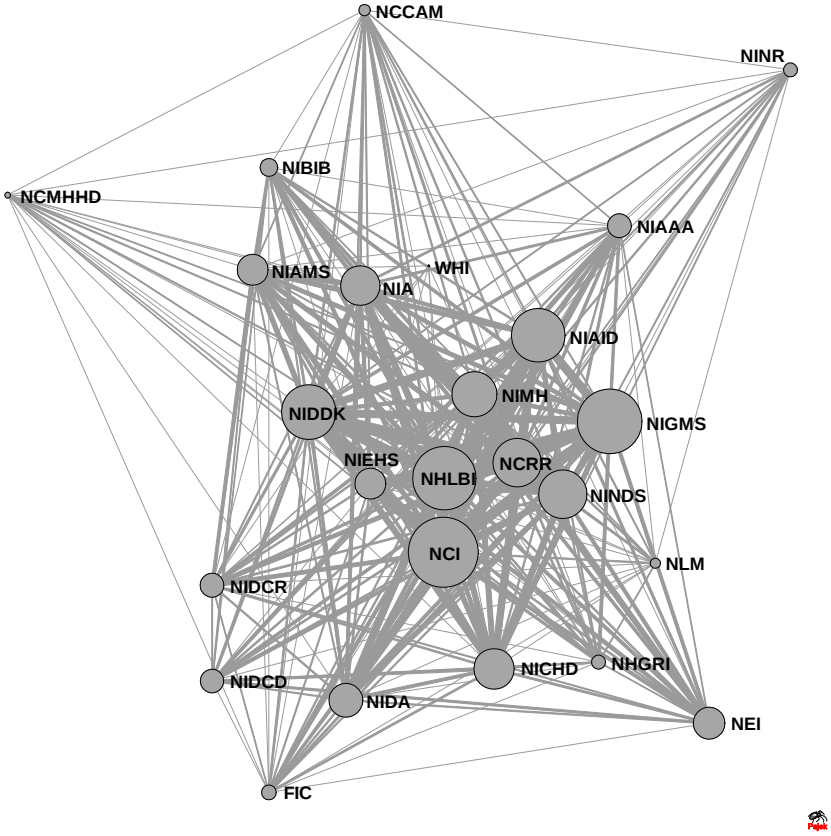

Figure 2. Map of NIH institutes based on grant co-occurrence (grants from multiple institutes being referenced in the same articles). Node size represents number of articles, while edge widths represent relative (square root) co-occurrence values.

Although Figure 2 positions the various institutes in relation to their overlaps, the layout obscures the edge widths, which are scaled to the actual co-occurrence values. The data have been re-plotted in a circular graph in Figure 3, with a randomized ordering of nodes, so that the edge widths can be viewed and compared. The two largest absolute overlaps in the set are NCI/NIGMS (with 4,044 co-funded articles) and NHLBI/NIDDK (with 4,020).

It was noted above that the number of articles per grant could not be calculated from the aggregated numbers in Table 1. Table 4 breaks down the numbers of grants by the initial grant year, along with the total numbers of articles referencing those grants. For example, an article published in 2000 that references a grant that was active from 1991-1998 will appear in the 1991 numbers. Articles referencing multiple grants are counted for each grant they reference.

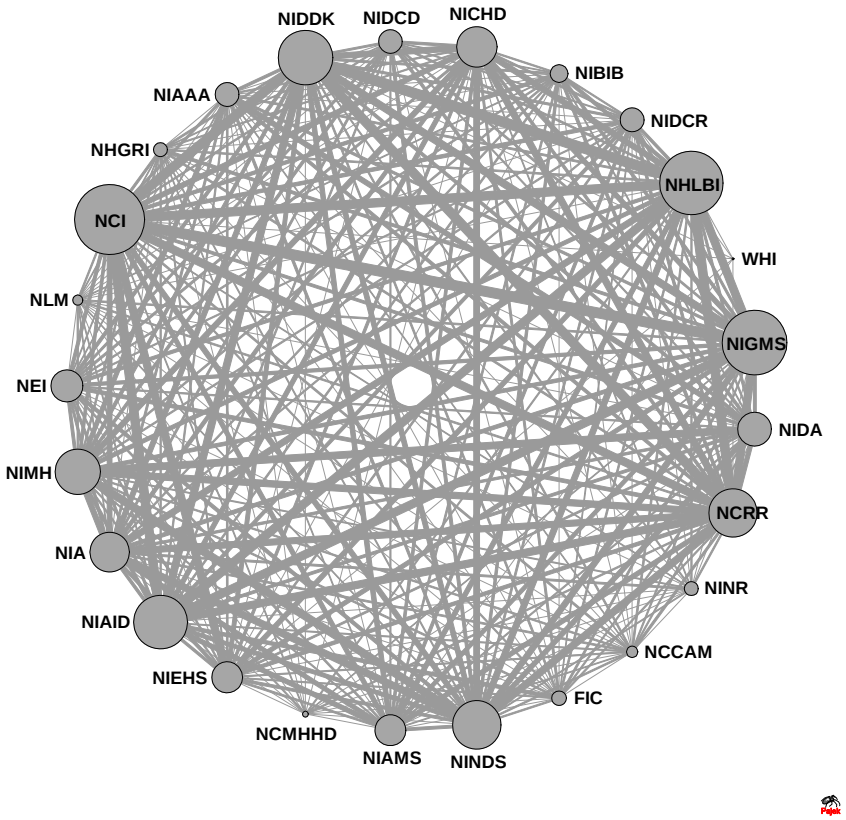

Figure 3. Circular map of NIH institutes based on grant co-occurrence with node positions randomized. Edge widths represent relative (square root) co-occurrence values.

Table 4. Numbers of NIH grants, and numbers of grants referenced by articles by initial grant year for unambiguous matches. Numbers of articles per grant are also shown.

Initial year	# grants	# gr/w/art	%gr/w/art	# art	# art/grant
≤ 1990	37,347	8,402	22.5%	104,336	12.42
1991	11,078	1,835	16.6%	17,719	9.66
1992	9,693	1,702	17.6%	13,222	7.77
1993	8,668	1,696	19.6%	12,684	7.48
1994	9,656	2,277	23.6%	16,800	7.38
1995	7,979	2,371	29.7%	17,494	7.38
1996	9,786	3,591	36.7%	22,735	6.33
1997	9,159	3,991	43.6%	23,790	5.96
1998	10,935	5,684	52.0%	40,810	7.18
1999	11,192	6,647	59.4%	47,186	7.10
2000	11,759	7,143	60.7%	51,285	7.18
2001	12,047	7,442	61.8%	50,765	6.82
2002	11,594	7,101	61.2%	39,974	5.63
2003	12,868	7,139	55.5%	32,776	4.59
2004	11,772	5,580	47.4%	18,457	3.31
2005	12,183	3,936	32.3%	9,460	2.40
ALL	197,716	76,537	38.7%	519,493	6.79

Note that the articles covered in Table 4 are limited to publication years 2002-2006; thus, the numbers of articles and articles per grant will be artificially low for the earlier years in the table (i.e., through 1999). However, given the commonly accepted lag time between grant and article of 3-4 years, the numbers for 2000 and later should be representative. The change in

the percentage of grants referenced by articles shows the effects of using a small publication window; the increasing percentages through the 1990s reflect the amount of time from the initial grant year to 2002. However, the fact that the percentage peaks from 2000-2002 suggest that this peak could reflect an actual ceiling in the fraction of grants that produce articles. It is also interesting to note that the fraction of grants producing articles and the number of articles per grant are greater for years up to 1990 than for the years 1991-1993. We expect that this is an artefact of our grant database, which starts in 1990, and also speculate that this grouping contains a larger than average fraction of very large multi-year grants than do the other yearly bins in Table 4. Further studies should be done to show the effect of grant duration on the values in Table 4.

The distribution of the number of articles produced per grant per year is shown in Figure 4 for different grant durations and varies widely. The average numbers of articles/grant/year for the six different grant durations are 2.35, 1.76, 1.77, 1.46, 1.50, and 2.85, from shortest to longest durations respectively. Thus, shorter grants actually seem to produce larger numbers of articles per year than most long grants. This is a counter-intuitive finding; one would assume that greater continuity would lead to larger numbers of publications. The grouping for grants of duration 16 years and longer contains large numbers of multi-institution grants with multiple subawards. This may be a contributing factor to the very large number of articles/year associated with this grouping. Median numbers are, of course, much smaller, and are 1.0 for most of the grant duration groupings. Note that these statistics consider all grant sizes equally. If grant sizes were corrected for, the distributions would most likely be much narrower. We leave this work to a future study.

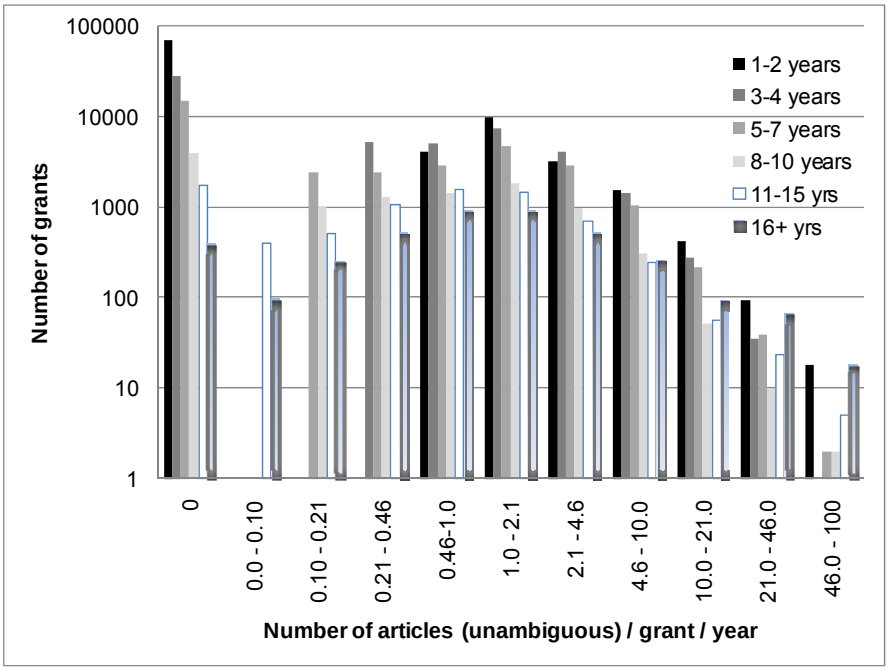

Figure 4. Number of grants by number of articles per grant/year for 6 different grant durations. The average grant duration over all grants considered here was 3.74 years.

We are also interested in the principal investigators and where they appear as authors of articles. The author orders for the PI's for each of the 533,364 unambiguous grant-article matches in Table 1 were found in the Medline data and counts are shown in Figure 5. It can be clearly seen that if the PI was listed as an author on the article, it was more often as last author (35%) than as the first (10%) or a middle (17%) author. This correlates well with the convention in biomedical publications for the leading author to be listed last. The PI was not listed as an author in 28.5% of the matches. This is not surprising in that many grants are

certainly large enough that not all work is directly overseen or published by the PI. Note also that combinations of author orders (e.g. first/last) are also shown in Figure 5. 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.

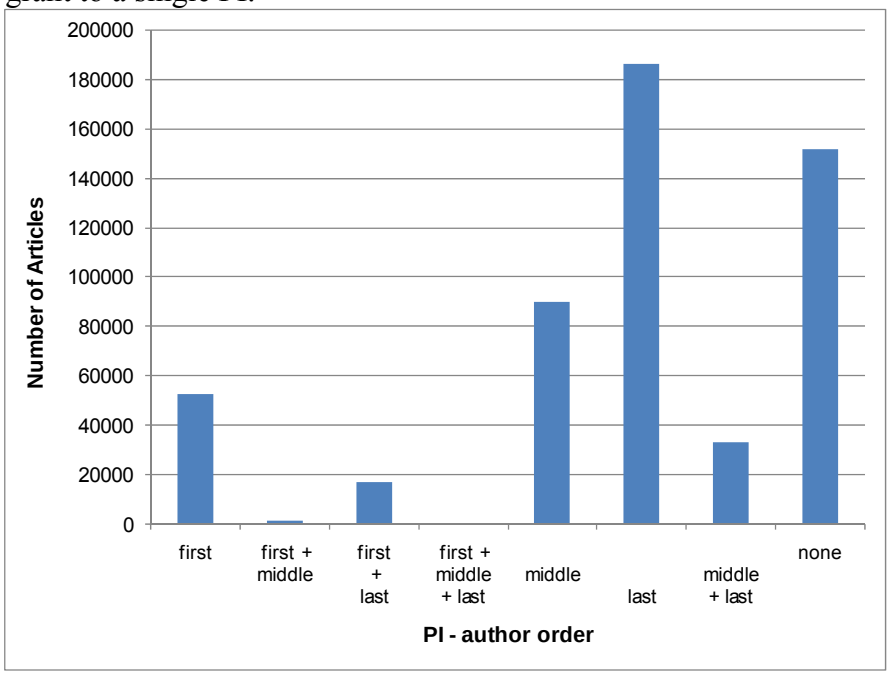

Figure 5. Author order counts for principal investigators from the matched grant-article data.

Limitations and Suggestions

There are, of course, limitations to the data described here that limit the depth of the analyses that can be undertaken. The grants information strings in Medline 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. We have chosen to ignore them in this study.

In the analyses above we have not made use of the funding data from RaDiUS that were merged into our grant database. Our funding data only exist for eight years (1998-2005) and thus effectively limit any comprehensive analysis to grants whose full durations fall within those eight years.[8] If assumptions are made about time lags between funding and publication, the window can be widened somewhat. Within these limitations 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. Adding citation counts to these data would require matching of the Medline articles to those in either the Thomson Reuters or Scopus databases.

In addition to the grant string information, Medline contains information about funding type. This information appeared in the MeSH terms through 2003, and has appeared in the article type field since 2004. The major funding types listed are *N.I.H., Extramural*, *N.I.H, Intramural*, *U.S. Gov't, P.H.S.*, *U.S. Gov't, Non-P.H.S.*, and *Non-U.S. Gov't*. Most other variations are typographical errors stemming from these five types. Figure 6 shows the fractions of the 797,369 Medline articles with confirmed U.S. first author addresses by funding type. The *N.I.H., Extramural*, *N.I.H, Intramural*, and *U.S. Gov't, P.H.S.* types have

[8] Funding data for additional years to overcome this limitation are available from USAspending.gov.

been combined into the PHS type shown in the figure since the NIH is the dominant part of the U.S. PHS funding system. *U.S. Gov't, Non-P.H.S.* (othGOV) includes other U.S. agencies such as the NSF, DOE, NASA, etc. *Non-U.S. Gov't* (non-USG) includes foreign funding sources as well as U.S. industry, non-profits, foundations, university endowments, etc. Some papers have multiple funding types, as shown by the multi-type funding types in Figure 6.

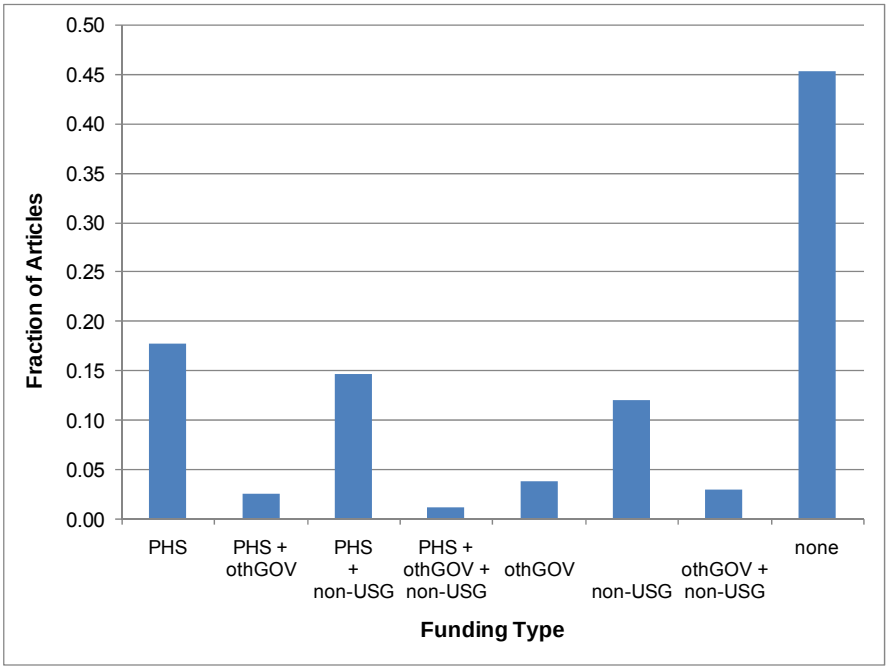

Figure 6. Funding types for Medline articles 2002-2006 where the first author address is in the United States. Articles comprise those publication types in Medline that are associated with scientific advances (e.g. journal articles, case reports, clinical trials, comparative studies, etc.)

It is difficult to know if the grant information strings indexed in Medline comprise the majority of the actual grant-to-article relationships or not. Figure 6 suggests that over 45% of the U.S. articles indexed in Medline have no acknowledgement of funding. Lewison (8) reported that 46% of nearly 13,000 UK gastroenterology papers had no acknowledged funding source, but that 85% of those were from National Health Service hospitals, and thus had an implied funding source by association. Further, Lewison, Dawson, and Anderson (14) found that while 39% of papers in the Research Outputs Database did not contain acknowledgements of funding, 7/8 of those could not be expected to have them. By contrast, Cronin and Franks (15) examined over 1000 articles from the journal *Cell* and found that over 90% of them had financial acknowledgements. We note that of the 286,911 articles associated with NIH or PHS funding types in Figure 6, 91.5% of them had grant information strings. This leaves only 8.5% (a relatively small number) of the articles noted to have received NIH or PHS funding, but for which the actual grant information was not indexed. Taken in total, these studies suggest that biomedical researchers do, for the most part, acknowledge government funding in a consistent and representative (if not totally complete) manner.

The indexing of grant information strings in Medline provides 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. Certainly, such data linking grants and articles are lacking for the US NSF and other agencies. Studies using the grant-to-article linkages identified in this work might be able to generate heuristics that could be used to infer grant-to-article linkages in other fields where direct linkage data do not exist. It is possible that some

combinations of text analysis with author matching might be sufficient for this purpose, and we suggest that this line of research be pursued.

Acknowledgments

This work was supported by NSF award SBE-0738111.

References

1. MCALLISTER, P. R. & WAGNER, D. A., Relationship between R&D expenditures and publication output for U.S. colleges and universities, *Research in Higher Education*, 15 (1981) 3-30.
2. MCALLISTER, P. R. & NARIN, F., Characterization of the research papers of U.S. medical schools, *Journal of the American Society for Information Science*, 34 (1983) 123-131.
3. BOURKE, P. & BUTLER, L., The efficacy of different modes of funding research: Perspectives from Australian data on the biological sciences, *Research Policy*, 28 (1999) 489-499.
4. BUTLER, L., Revisiting bibliometric issues using new empirical data, *Research Evaluation*, 10 (2001) 59-65.
5. JIMENEZ-CONTRERAS, E., MOYA-ANEGÓN, F. & LOPEZ, E. D., The evolution of research activity in Spain: The impact of the National Commission for the Evaluation of Research Activity (CNEAI), *Research Policy*, 32 (2003) 123-142.
6. KWAN, P., JOHNSTON, J., FUNG, A. Y. K. et al., A systematic evaluation of payback of publicly funded health and health services research in Hong Kong, *BMC Health Services Research*, 7 (2007) 121.
7. HICKS, D., KROLL, P., NARIN, F. et al. (2002) Quantitative methods of research evaluation used by the U.S. federal government *NISTEP Study Material* (Second Theory-Oriented Research Group, National Institute of Science and Technology Policy (NISTEP), Japan).
8. LEWISON, G., Gastroenterology research in the United Kingdom: Funding sources and impact, *Gut*, 43 (1998) 288-293.
9. LEWISON, G. & DEVEY, M. E., Bibliometrics methods for the evaluation of arthritis research, *Rheumatology*, 38 (1999) 13-20.
10. LEWISON, G., GRANT, J. & JANSEN, P., International gastroenterology research: Subject areas, impact, and funding, *Gut*, 49 (2001) 295-302.
11. BOYACK, K. W. & 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*, 54 (2003) 447-461.
12. BOYACK, K. W., Mapping knowledge domains: Characterizing PNAS, *Proceedings of the National Academy of Sciences*, 101 (2004) 5192-5199.
13. BOYACK, K. W., BÖRNER, K. & KLAVANS, R., Mapping the structure and evolution of chemistry research, *Scientometrics*, 79 (2009) 45-60.
14. LEWISON, G., DAWSON, G. & ANDERSON, J. (1995) The behaviour of biomedical scientific authors in acknowledging their funding sources *5th International Conference of the International Society for Scientometrics and Informetrics*, pp. 255-263 (River Forest, IL).
15. 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*, 57 (2006) 1909-1918.