Journal of Informetrics 13 (2019) 226–237

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regular article

Are NIH-funded publications fulfilling the proposed research? An examination of concept-matchedness between NIH research grants and their supported publications

Kai Li*, Erjia Yan

Drexel University, Philadelphia, PA, 19104, USA

ARTICLE INFO

Article history:
Received 13 September 2018
Received in revised form 3 January 2019
Accepted 3 January 2019

Keywords:
Research topics
Natural language processing
National Institutes of Health
Research grants

ABSTRACT

The conceptual connections between scientific grants and publications are important, yet often overlooked in quantitative studies of science. An analysis of such connections could offer important insights into how science is conducted by individual researchers and research teams under the social and economic conditions of science. This study aims to offer the first piece of evidence towards this endeavor by analyzing the ratio of keyword matchedness between accepted NIH research grants from 2008 to 2015 and their funded publications. By applying linear regression method, we identified and examined three identified predictors of the outcome: 1) the funding rate of an NIH research program in a specific year, 2) the year difference between grant and publication, and 3) the funding size of a grant. Our findings suggest that these three factors contribute to the outcome in different capacities. Moreover, all of them may have different performances in individual funding programs, which highlights the importance of understanding the differences among individual funding mechanisms.

1. Introduction

Scientific vocabularies are an essential representation of science. Topics used in scientific publications, such as keywords, have been frequently studied to understand the distribution of specialties in the overall scientific community and how research interests shift over time. However, much of the relationship between research content and the socio-economic conditions in which science is produced is yet to be studied, which motivates the present paper.

Researchers have given various explanations for the nature of science. The two most important ideas during the 20th century are scientific realism (i.e., scientific developments approaching to the ultimate truth) and social constructivism (i.e., science being a social construct), with the former being gradually challenged by the latter (Boyd, 1983). The new constructivist school shares the understanding that there is a symmetrical relation between scientific knowledge and the non-scientific infrastructure in which such knowledge is produced. As such, one should attend to the ecosystems and processes in which knowledge products or claims are created, in order to better understand scientific knowledge.

An important component of the socio-economic conditions of science is the funding to support scientific activities. Scientific funding has received increasing attention since the late 20th century, when external scientific grant became a scarcer resource. This shift, famously named as *academic capitalism*, was thoroughly examined by Sheila Slaughter and her colleagues

* Corresponding author.
E-mail addresses: kl696@drexel.edu, ey86@drexel.edu (K. Li).

https://doi.org/10.1016/j.joi.2019.01.001

(Slaughter & Leslie, 1997; Slaughter & Rhoades, 2004). The term of academic capitalism denotes activities undertaken by universities and individual researchers to “secure external moneys” (Slaughter & Leslie, 1997, p. 8). Such activities include not only direct market activities (e.g., patents, licensing agreements, and university-industry partnerships) but also indirect, market-like activities, most notably the competition for external grants and contracts. Driven by a significant decrease in research funding from state governments, higher-education institutions have no choice but to make stronger efforts to seek resources from national funding agencies and industry funders (Slaughter & Leslie, 1997). Under this “triple helix” model, named after the relationship among governments, universities, and the industry, scientific norms have shifted dramatically from those established under the “pure science” paradigm of the 1940s (Etzkowitz & Leydesdorff, 2000; Etzkowitz, 2008; Leydesdorff & Etzkowitz, 1996).

This recent structural shift of scientific funding makes it further urgent to investigate how the changing availability of grant money has affected the evolution of research topics. A few researchers have offered some efforts to answer this question by investigating how scholars adapt to their topics in accordance with grant requirements (Laudel, 2006, Gläser & Laudel, 2016) and the evolution of research topics between grants and publications within an entire research community (Shi, Nallapati, Lescovec, Mcfarl, & Jurafsky, 2010). However, the influence of scientific grants on research contents still remains a blind spot in this overall research agenda (Gläser & Laudel, 2016; Hottenrott & Lawson, 2014). Such a gap is reflected by a lack of quantitative and large-scale analyses addressing the relationship between concepts discussed in grants and pursued in their funded publications. This is an especially meaningful question under the new paradigm of academic capitalism, because when scientific grants become more difficult to acquire and a more valuable reward, it could potentially affect how researchers pursue their research content. One important strategy identified in the literature is *bootlegging*, which is that researchers use grant money under a specific project to try new research ideas or even changing the research trial (Gläser & Laudel, 2016; Hackett, 1987; Laudel, 2006). Given the lack of evidence, it is difficult to quantify, not to say interpret, behaviors of how researchers use their funding, which is a major barrier faced by both universities and scientific funders.

In this study, we present a method to compare research topics between connected research grants and publications in large corpora and investigated factors that affect the level of affect the level of topic closeness within the same funding cycle. To conduct our analysis, we analyzed research grants from the National Institutes of Health (NIH). This choice is due to the immense roles played by NIH’s investments in scientific research. NIH invests more than 37 billion dollars annually in medical research, 80% of which is awarded through competitive grant.[1] At present, NIH offers these grants under 241 *funding programs* organized into seven groups.[2]

Combining the metadata about NIH research grants and their funded publications, we measured the keyword matchedness between abstracts of connected grants and publications. We then applied statistical methods to test the following three hypotheses concerning keyword closeness between the two sources.

Hypothesis 1 (H1). The keyword matchedness rate is affected by the funding rate of a specific program in a given year.

One central question we are interested is how funding rate by program and year, taken as a representation of overall pressure or competition for funding, affects the level of keyword-matchedness between grants and their funded publications. Studies on the socioeconomic conditions of science have proven that academicians now suffers from stronger stresses to pursue external grants (Slaughter & Leslie, 1997), and a popular strategy towards this goal is to pursue multiple lines of research under the same grant (Gläser & Laudel, 2016; Hackett, 1987; Laudel, 2006). Based on these observations, we hypothesized that with higher pressure (or lower success rate in securing scientific grants), researchers would be more likely to pursue research topics less related to the original grant descriptions in subsequent publications. Funding pressure may motivate researchers to maintain their other line(s) of research for the sake of the next grant proposal, or it may drive them to develop new and potentially more productive research interests more quickly.

Hypothesis 2 (H2). The keyword matchedness rate is affected by the year gap between the project and publication.

The second hypothesis is inspired by the general observation that individual researchers and research groups naturally shift their research interests over time, and this shift is reflected in the topics of emerging publications (Hamilton, Leskovec, & Jurafsky, 2016; Yan & Zhu, 2018). To apply this observation to the research cycle, this hypothesis holds that as time goes by, there are weaker conceptual connections between publications and the original research ideas stated in the grant proposal. This may happen independently of the deliberate choices implied in the first hypothesis.

Hypothesis 3 (H3). The keyword matchedness rate is affected by the funding size of the accepted grant.

Despite the evidence that there is a positive correlation between funding size and researchers’ productivity and impact (Bloch & Sørensen, 2014; Boyack, 2004), its relationship with the level of keyword matchedness within the same research cycle is nevertheless unknown. Aiming to provide preliminary evidence to fill this gap, we hypothesized that funding size per year of a grant could have impact on the keyword-matching ratio of researchers in subsequent publications.

To test these three hypotheses, we analyzed all R-series research grants awarded by NIH between 2008 and 2015 and all publications published between 2008 and 2017 linked to these grants. All data were acquired from the NIH Research

[1] https://www.nih.gov/about-nih/what-we-do/budget.
[2] https://grants.nih.gov/grants/funding/funding_program.htm.

Portfolio Online Reporting Tools (ExPORTER) platform.[3] We used a popular keywork-extraction program, RAKE (Rose, Engel, Cramer, & Cowley, 2010), to extract keywords from grant and paper abstracts. Keywords within each project-publication pair were then matched to represent the conceptual similarity between grant and publication. Finally, linear regression method was applied to evaluate the relationship between the level of keyword matchedness in a project-publication pair and other independent variables that might affect the outcome, especially the three variables connected to our hypotheses.

The following section presents a review of literature related to our research question. We then provide more details about our methods and data, followed by major results from the descriptive and statistical analyses. At the end of the paper, we discuss our hypotheses and their implications in light of our findings.

2. Literature review

2.1. How grants affect the productivity and impact of researchers

The relationship between scientific grants and outputs has been a popular topic for researchers in quantitative studies of science. Many studies have identified ways in which funding affects researchers' productivity and scientific impact. A series of studies conducted by Grant Lewison and his colleagues (Lewison & Dawson, 1998; Lewison, 1998; Lewison, Grant, & Jansen, 2001), have observed a positive correlation between the number of scientific funds acknowledged in a publication and its impact. Focusing on nanotechnology studies in Canada, Beaudry and Allaoui (2012) have demonstrated that only public funding, rather than private funding or contracts, has a positive correlation with higher scientific productivity. Moreover, several recent empirical studies have proved that NIH funding is connected to higher scientific productivity and/or impacts for individual researchers in various medical science fields (Colaco, Svider, Mauro, Eloy, & Jackson-Rosario, 2013; Rezek, McDonald, & Kallmes, 2011; Svider et al., 2013; Venable et al., 2014).

Alongside this positive evidence, it is important to note that some studies have also identified weak or negligible connections between grants and the number and impact of publications. The work of Carter and colleagues (Carter, Winkler, & Biddle-Zehnder, 1987) provides an early example: examining the NIH Research Career Development Awards (NCDA) program, they reported that most of the variance of publication-based measures of research productivity can be explained by the quality of researchers considered for the award, rather than the award per se. Similarly, Fortin and Currie (2013) reported a weak connection between funding and scientific impact in their examination of Natural Sciences and Engineering Research Council of Canada (NSERC) grants.

The discrepancies between these sets of findings can be partly explained by social variables. For example, in their series of papers, Jacob and Lefgren (Jacob & Lefgren, 2011b, 2011a) found that being funded by NIH has different effects on faculty researchers and post-doctoral researchers: it increases productivity to a much higher degree (20% vs. 7%) for the latter group than for the former over the next five years. Moreover, Arora and Gambardella (2005) have shown that the positive effect of NSF funding on the productivity of basic economic studies is much stronger for junior researchers than for their senior counterparts. Last but not least, Yan, Wu, and Song (2018) have demonstrated that the number of authors and institutions in the same publication is an important mediator between funding and paper impact.

2.2. Evolution of scientific vocabularies

Compared to studies examining the quantitative relationship between grants and their outputs, the influence of grants on research content has been much less thoroughly investigated, especially from the quantitative perspective.

A number of qualitative studies have been conducted to investigate how researchers adapt their research contents based on the requirements of research grants. For example, Laudel (2006) observed that scientists have to decide the fundability of their research topics; in order to increase the likelihood of funding, scientists can take strategies like diversifying research, avoiding risky research, and selecting externally predetermined research contents. An important factor underlying such strategies is that, given the conservatism of grant application review, researchers have to accommodate both their research contents and rhetoric to the requirement of *do-ability* (Fujimura, 1987) which has been confirmed by a variety of qualitative analyses (Glaser et al., 2010; Hottenrott & Lawson, 2014).

From a macro-level perspective, a number of researchers have also attempted to understand the relationship between the amount of research grant and the shift of research interests. Such studies include the work of Ahlin and Gibbs (2012), which reported a reciprocal relationship between funding for research on community-oriented policing (COP) and the overall popularity of this topic. Another line of inquiry is to use the lead-lag method to understand the cascade of research topics between grant proposals and scientific publications. An important example of this method was implemented by Shi et al. (2010), who extracted research topics using latent Dirichlet allocation (LDA) from a sample of grant proposals and scientific publications in computer science, then analyzed the lead-lag relationship between concepts in these two corpora.

Despite these analyses, as stated in Gläser & Laudel review's (2016), a lack of evidence, both qualitative and quantitative, prevents us from gaining even a brief understanding of how research policy (of which research grant is the most important

[3] https://exporter.nih.gov/default.aspx.

vehicle) is affecting research content. Most of the evidences presented above, however, focus on either how research content is shifted before grant application is submitted, or the cascade of research within the whole research community. In the present study, we used a different approach, i.e., tracking the change of research topics within the same research cycle after research funding are granted. We are hoping this new approach could offer different yet informative evidence towards quantitative studies of science.

From a different angle, researchers have also aimed to understand whether the topical connections between grants and publications is a factor determining the success rate of one's grant proposal. Boyack et al.s' recent article is similar with our study in terms of the method. By conceptualizing the similarity of topics between one's proposal and previous publications as an applicant's strengths related to a research topic, Boyack, Smith, and Klavans (2018) found that the past topic profile is an advantage for individuals, rather than universities, in grant applications. This study reached similar results with some earlier efforts (Hörlesberger et al., 2013) and offers an interesting context of the present study.

3. Materials and methods

3.1. Methodological pipeline

In order to pursue our research questions, we developed a text-analysis pipeline based on a number of NLP techniques using the Python programming language. This section describes each step of our methodological pipeline in further detail and presents basic descriptive information about our sample.

Step 1: Data preparation

In order to compare keywords from NIH grant and publication abstracts, we combined data from the NIH ExPORTER system and the PubMed database.

We first downloaded all metadata for grants awarded by NIH from 2008 to 2015 (under the "PROJECTS" and "ABSTRACTS" tabs) from the ExPORTER system. This includes information for each grant, such as its starting and ending years, activity code, applicant information, funding size, and abstract. Because some NIH grants are not research-driven, we selected only research projects (activity codes beginning with "R") as the sample of this study.

We also downloaded the publication metadata from 2008 to 2017 ("PUBLICATIONS" tab) from ExPORTER. Our publication window extends two years beyond our grant window because we wanted 2015 grants to have at least three years of publication records, so that publication trends for these grants could be compared with those in other years. The NIH data include the PubMed ID (PMID) for each publication, but not the abstract. Using this identifier, we downloaded the abstract for each paper from the PubMed API as implemented in the *metapub* Python package.[4] Publications lacking a PMID or PubMed abstract were removed from our sample.

All sampled grants and publications were linked via the link tables ("LINK TABLES" tab) offered by NIH. Such project-publication pairs were the basis of the keyword comparison described in Step 4. Any item, project, or publication unable to be linked was removed from our sample.

Step 2: Keyword extraction

We extracted keywords from each grant and paper abstract using a Python implementation of Rapid Automatic Keyword Extraction (RAKE),[5] an unsupervised, domain-independent keyword extraction algorithm (Rose et al., 2010). By calculating ratio of degree to frequency (deg(w)/freq(w)) of terms not in the stop word list, RAKE selects keywords that appear more frequently and dominantly within longer keyword candidates. Moreover, this algorithm favors n-grams over unigrams: longer phrases without a stop word are assigned higher scores (Rose et al., 2010).

However, RAKE's tendency to favor long phrases is not always consistent with the grammatical traits of keywords. Based on the findings on the linguistic characteristics of keywords (Justeson & Katz, 1995), we selected only bi- and trigrams with the following part-of-speech combinations from the results of RAKE as the final keywords, where *NN* stands for a noun, *JJ* stands for an adjective, and *IN* stands for a preposition (such as *of* and *with*):

- NN + NN
- JJ + NN
- JJ + JJ + NN
- JJ + NN + NN
- NN + JJ + NN
- NN + NN + NN
- NN + IN + NN

Step 3: Keyword validation

In order to evaluate the meaningfulness of the extracted keywords from the RAKE algorithm, two coders individually rated all 475 extracted keywords from 20 randomly selected grant abstracts. Based on their conceptual relationship with the

[4] https://pypi.org/project/metapub/.
[5] https://github.com/aneesha/RAKE.

Table 1
Examples of direct and indirect matched terms.

Project Number	Paper PMID	Project keywords	Paper keywords
R01AI080884	22797772	dendritic cell	dendritic cell
R01AA016798	22381123	substance use	substance use disorder

abstract where they were extracted, we defined three levels of meaningfulness of keywords as the guideline of our coding work:

- **Level 1 (score 1):** the extracted phrase is meaningful and is able to reflect the topic of the original text. Such examples include phrases representing the concepts discussed in the abstract.
- **Level 2 (score 0.5):** the extracted phrase is meaningful but cannot reflect the topic of the original text. Most of these examples are keyword-like, but too general compared to the concepts in the abstract, such as "negative consequence" and "robust response."
- **Level 3 (score 0):** the extracted phrase is meaningless. Most of these phrases can be easily distinguished from keywords. They were extracted because the linguistic rules specified in the previous subsection were too broadly applied or the POS attribute of some words were mistakenly identified by the algorithm. Examples of this type of phrases include "good understanding" and "low level."

Given the ordinal levels in our rating system, we tested inter-coder reliability using Cohen's weighted kappa (Ben-David, 2008), instead of the regular Cohen's kappa method. Applying this method, the resulting Cohen's kappa (κ) is 0.774. Based on the guidelines from Landis and Koch (1977), a kappa (κ) of.774 represents a substantial strength of agreement. Furthermore, since $p = .000$, our kappa (κ) coefficient is statistically significantly different from zero.

Based on our manual coding, the mean value of keyword meaningfulness across the 20 grant abstracts is 0.786, with the mean scores on the abstract-level ranging from 0.542 to 0.941. Among all 457 extracted phrases, 337 (73.7%) of them were rated into Level 1. These results suggest that most of the keywords extracted by RAKE are meaningful terms that can be used to summarize the concepts discussed in the abstracts.

Step 4: Keyword comparison and statistical analysis

Two further actions were taken before keywords from both sources were matched. First, we added author-supplied project keywords (those that were either bi- or trigrams) included in the project datasets to the automatically-extracted keywords from the project abstract. The combined list of keywords is hereafter referred to as *project terms*. Second, all project terms and automatically-extracted terms from paper abstracts (*paper terms*) were lemmatized using the *Spacy* Python package (Honnibal & Johnson, 2015), so that keywords with name variations could be correctly matched.

To compare project terms and paper terms within each project-publication pair, we adopted a combination of indirect and direct matching methods. That is, apart from identical bigram and trigram keywords from both keyword lists, we counted matches in which a bigram from one list matched a part of a trigram from another, or vice versa. It should be noted that trigrams are split and only split into two bigrams. For example, the term "intravenous alcohol administration" would be matched with either "intravenous alcohol" and "alcohol administration," rather than "intravenous administration."

Any matched keyword within a project-publication pair makes this pair a matching pair, regardless of the number of matched terms within the pair. Matched and unmatched pairs were assigned *Matchedness* values of "1" and "0" respectively. Examples of two matched pairs are offered in Table 1. The first example is a direct match, the second an indirect match.

For all project and publication abstracts we examined, we found a number of items with extremely short lengths. Many of these abstracts seem to be incomplete ones caused by mistakes of either the ExPORTER system or the PubMed API. We manually removed all abstracts shorter than 50 words to ensure the quality of abstract texts.

In order to pursue RQ1, we acquired the acceptance of each funding program from 2008 to 2018 from the NIH website.[6]

In order to answer RQ3, we classified all grants into three groups based on its funding size per year. Funding size per year is a size independent indicator that effectively eliminates multicollinearity caused when both funding size and project duration are used in the regression model. The three groups are: lower than 60,000 dollars (Group 1), between 60,000 and 90,000 dollars (Group 2), and over 90,000 dollars (Group 3). The two cutoffs were selected in order to categorize all grants into three groups of similar sizes (28%, 42%, and 30% for Groups 1–3, respectively).

Based on this information, we calculated the ratio of matched papers by project year, and further divided these groups by project-publication year gap, funding program, and funding size. We did not consider any group of projects with fewer than 50 projects, in order to avoid radical keyword-matching ratios. We further applied linear regression models to these final groups of data to test the hypotheses specified in the Introduction. Linear regression analysis was conducted using R (R Core Team, 2016).

[6] https://report.nih.gov/success_rates/.

Table 2
Grant abstract length (number of words) and keyword counts for the top three funding programs.

Activity code	Mean project abstract length	Mean project keywords	Keyword / Abstract length correlation
R01	430.84	23.03	0.616
R03	416.93	21.81	0.63
R21	422.12	22.55	0.633

Table 3
Paper abstract length (number of words) and keyword counts for the top three funding programs.

Activity code	Mean paper abstract length	Mean paper keywords	Keyword / Abstract length correlation
R01	201.69	10.82	0.567
R03	212.41	11.14	0.562
R21	204.03	10.99	0.575

3.2. Keywords profiles

On the population level, 925,610 unique keywords were extracted from all publications for a total of 3,255,276 times, while 391,798 keywords were extracted from all grants for 1,185,656 times. It is not surprising that the distributions of both keyword lists are extremely long-tailed. For example, in the paper keyword list, only 108 keywords were extracted from more than 1000 publications among more than 400,000 publications. Similarly, only 84 keywords were extracted from more than 500 grant abstracts among the 52,000 grants in our sample. On the other side of the distribution, both lists have over 70% of keywords that only appear in one abstract (71.5% for paper term list and 72.9% for grant term list).

We listed the top 50 keywords from both paper keyword list and grant keyword list in the Appendix. As is expected from the results of the baseline analysis stated above, not all of the extracted keywords are fully meaningful ones: top terms have even higher likelihood not to be very meaningful. Despite these facts, the majority of the top keywords on both lists are meaningful keywords. Moreover, all keywords have a relatively low presence among all items, whether publications or grants. The top keywords from grants, such as “mouse model” and “molecular mechanism,” are found in less than 6% of all grants. Similarly, the top keywords from papers, such as “mouse model” and “gene expression,” are only extracted in about 1% of all publications.

For both grant and publication abstracts, there is a strong positive correlation between the length of abstract and the number of extracted keywords. For both grant and publication abstracts, there is a strong positive correlation between the length of abstract and the number of extracted keywords. The correlations between these two parameters for project abstracts and paper abstracts are 0.635 and 0.567, respectively.

Tables 2 and 3 show the relationship between abstract length and the number of keywords extracted from them, by focusing on the three largest funding programs. The abstract length and the number of extracted keywords per abstract are largely identical among the three funding programs, so is the correlation between these variables within each group.

3.3. Description of the sample

52,457 projects and 455,572 papers are included in our final sample. Table 4 shows a summary of descriptive information grouped by project year. All year groups have a comparable number of projects, number of paper terms, and keyword match ratio. The number of papers for each project year is largely a function of the total potential number of years between the project year and the last year for publication: projects in earlier years have larger numbers of publications on average. There is a noticeable difference in number of project terms between projects in 2008 and all other years, but this difference is solely attributable to the large number of PI-supplied keywords in 2008. Beyond this difference, year groups have a highly consistent number of RAKE-extracted terms from the project abstracts.

Moreover, as shown in Fig. 1, all project years have similar distributions of papers in terms of the project-publication year gap: projects normally reach the peak number of publications by the 3rd to 4th year after the project begins. This pattern seems to be consistent even for the most recent project years, despite the incomplete data.

Table 4
Sample summary by project year.

Project year	Papers	Projects	Funding size per year (in dollars)	Papers per project	Mean project terms	Mean paper terms	Match ratio
2008	74,003	6,542	71,874	11.31	111.57	10.83	0.604
2009	95,294	9,385	108,528	10.15	56.05	10.91	0.563
2010	74,560	7,076	90,805	10.54	55.18	10.91	0.563
2011	59,877	6,051	79,949	9.9	55.48	10.91	0.553
2012	54,490	6,452	79,434	8.45	55.27	10.84	0.548
2013	42,405	5,753	79,181	7.37	53.17	10.81	0.541
2014	33,748	6,087	84,651	5.54	54.33	10.68	0.551
2015	21,195	5,111	87,417	4.15	55.01	10.6	0.551

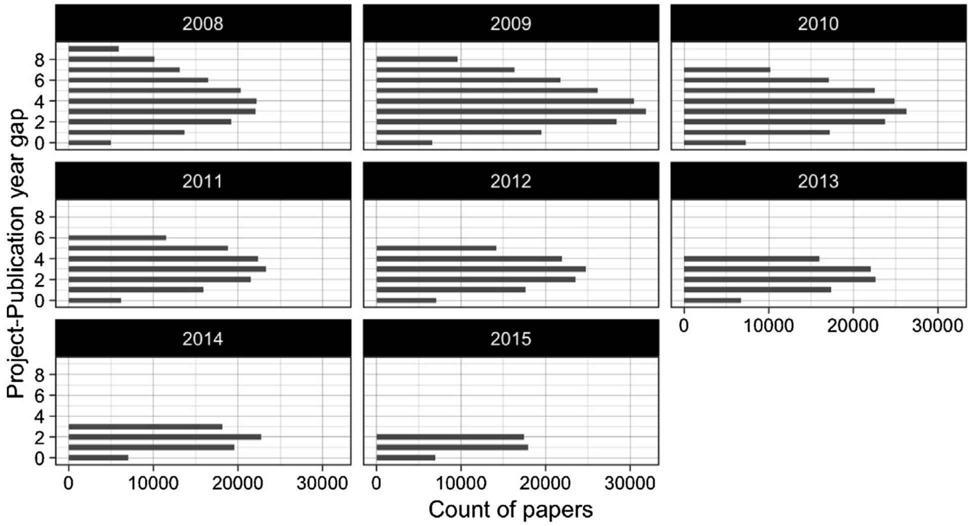

Fig. 1. Year distribution of papers by project year.

Table 5
Sample summary by activity code.

Activity code	Papers	Projects	Funding size per year	Papers per project	Mean project terms	Mean paper terms	Match ratio
R01	340,556	28,637	86,201	11.89	63.76	10.82	0.566
R03	13,861	3,582	31,068	3.87	59.01	11.14	0.643
R21	56,875	12,534	74,561	4.54	61.09	10.99	0.593

It is worth noting that there is a highly uneven distributions of projects and papers by funding program: the top program, R01, contains 54.5% of all projects and 74.8% of all publications, while the second largest program, R21, accounts for 24% of projects and 12.5% of publications in our sample. Given this uneven distribution, we decided to include only the top three programs (R01, R21, and R03) in our statistical analysis; these are also, not coincidentally, the only three programs with more than 10,000 publications in our dataset. In total, these three programs account for 77.8% of all unique projects and 90.3% of publications in our sample. Descriptive information for each program is summarized in Table 5.

4. Results

We first plotted the ratio of matched papers over the project-publication year gap, as shown in Fig. 2. Two observations seem to emerge from the graph. First, there seems to be a consistently decreasing trend in keyword matching ratio as the time between project and publication increases. Second, more recent projects tend to have lower keyword matching ratio than older ones—even though, apart from 2008, this difference is not very large between years.

To address the hypotheses stated in Introduction, we constructed a model with *mean ratio of matched publications* (m_ratio) as the dependent variable and the following as independent variables:

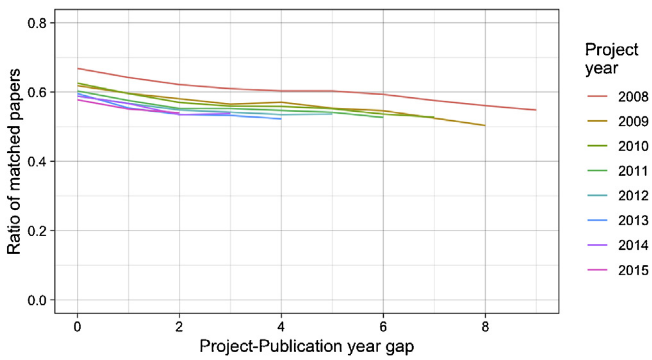

Fig. 2. Ratio of matched papers by project-publication year gap.

Table 6
Results of the linear regression model (p < .001 ***; p < .01 **; p < .05 *).

Variable	Est. Coefficient	Std. Error	*t*-value	*p*-value
(Intercept)	0.413	0.087	4.750	0.000***
acc	0.151	0.315	0.479	0.632
y_gap	−0.007	0.001	6.529	0.000***
p_year2009	−0.016	0.009	1.751	0.081
p_year2010	−0.032	0.009	3.666	0.000***
p_year2011	−0.043	0.014	3.183	0.002***
p_year2012	−0.049	0.013	3.864	0.000***
p_year2013	−0.059	0.016	3.782	0.000***
p_year2014	−0.059	0.014	4.186	0.000***
p_year2015	−0.067	0.014	4.961	0.000***
length_2	0.017	0.006	3.078	0.002***
programR03	0.065	0.011	6.164	0.000***
programR21	0.021	0.008	2.509	0.013*
fundingG2	0.007	0.005	1.295	0.196
fundingG3	−0.001	0.006	0.101	0.920

- Project year (p_year)
- Project-publication year gap (y_gap)
- Program (prog)
- Program acceptance rate (acc)
- Project term length (length_1)
- Paper term length (length_2)
- Funding size per year (funding)

To ensure the suitability of the linear regression model, we assessed the following criteria:

- Linearity of residuals
- Independence of residuals
- Normal distribution of residuals
- Equal variance of residuals

Among all independent variables, *Project term length* alone failed to pass the requirement of independence of residuals (with $GVIF^{1/(2*Df)} = 7.84$). It was then removed from the model. The revised model, whose results are reported below, passed all the requirements.

The adjusted R-squared value of this model is 0.451, which suggests that this model is able to explain 45.1% of the variance in the data. Table 6 shows the results for each independent variable.

Both *Project year* and *Funding size per year* were treated as a categorical variable. In the case of *Project year*, different years are displayed as “p_yearxxxx” in our results, with 2008 used as the baseline for all other years. For Funding size per year, Group 1 (the low-funded group) was used as baseline, and other groups are displayed as “fundingGX.” Our results show that more recent project years tend to have lower keyword matching ratios, consistent with Fig. 2. Projects in all years but 2009 have significantly higher levels of keyword matchedness than those in 2008. All other results that are relevant to our RQs are discussed in more detail in the next section.

5. Discussion

This section discusses how the foregoing results shed light on the three hypotheses of this study and their implications.

H1. The keyword matchedness rate is affected by the funding rate of a specific program in a given year.

At face value, the results shown in Table 6 seem to reject the claim that there is a significant, positive relationship between the funding rate of a project and grant/publication keyword matchedness. However, one limitation of this statement should be kept in mind: the relationship between funding rate and keyword matchedness varies by program as shown in Fig. 3. Fig. 3B shows the relationship between ratio of matched papers and project year grouped by program. This graph, as compared to Fig. 3A, shows that R01, the dominant program in our sample, displays a pattern distinct from the other two programs we examined: while its funding rate largely remained parallel to that of other programs, its keyword-matching ratio dropped markedly over the years.

To understand how this relationship performs differently in different funding programs, we conducted grouped linear regression within each funding program. We used the linear regression model discussed above (with *Program* removed). The results concerning the overall model and the acceptance rate are shown in Table 7.

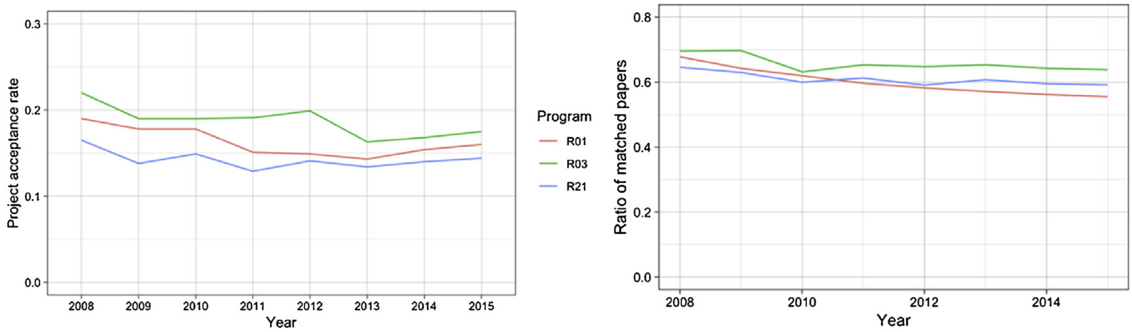

Fig. 3. Project acceptance rate (left) and ratio of matched papers (right) for each grant program by year.

Table 7
Results of grouped linear regression (p < .001 ***; p < .01 **; p < .05 *).

Group	Model adjusted r-squared	Est. Coefficient	Std. Error	*t*-value	*p*-value
R01	0.689	2.72	0.29	9.32	0.00***
R03	0.078	1.57	0.7	2.24	0.03*
R21	0.295	3.28	0.86	3.83	0.00***

The results demonstrate that, despite the different goodness-of-fit values of the three grouped models,[7] the acceptance rate of a specific funding program in a given year is shown to be a strong predictor of the outcome. For example, a unit (100%) increase in the acceptance of R01 program could result a 272% increase in the keyword matchedness (95% CI [243%, 301%]). Pragmatically, the acceptance rates of the three programs during 2008 and 2018 range from 12% to 22%. This 10% of difference could result a variation between 24% and 30%. This effect is even stronger for the R21 program, whose 95% CI spans from 242% to 414%.

The strong positive connection between these two variables in both R01 and R21 points to the idea that increased funding pressure could drive a quicker shift in research concepts from grants to publications within the same research setting. In other words, different research topics are more likely to emerge in publications if the grant rate drops. Our future work will focus on the interpretation of this phenomenon.

H2. The keyword matchedness rate is affected by the year gap between project and publication.

The second hypothesis is supported both by the consistent pattern across *Project year* groups (Fig. 2) and by our linear regression model. In the overall regression results (Table 6), *Project-publication year gap* is shown to be a significant predictor of the outcome: a unit (one year) increase corresponds to a 0.7% decrease in keyword-matching ratio (95% CI [0.6%, 0.8%]). Similar trends were also shown in the R01 program, but not in R21. In R01, one unit of change in the year gap could bring a 1% of change in the keyword-matching ratio (95% CI [0.9%, 1.1%]), while in R21, this effect fails to show any statistical significance.

Our results suggest that keywords use changes over time within the same research cycle, regardless of other factors. This finding is definitely consistent with our daily experience: researchers shift their research interests, and thus keyword, regardless of all other factors. However, our finding demonstrates that this phenomenon exists even in the same research cycle. However, this seemingly natural phenomenon is nevertheless complicated by our finding that this trend is only present in certain funding programs, which is an important question to be answered by future studies.

Indeed, many factors could contribute to the natural shift of scientific vocabularies within the same research cycle, such as researchers' shifted interests, new scientific theories and instruments, and even the natural shift of scientific vocabulary outside individual research cycles. In this study, we are assuming the results are attributable to the combination of all these potential reasons, rather than identifying every possible factor and their relationship.

Hypothesis 3 (H3). The keyword matchedness rate is affected by the funding size of the accepted grant.

Our general model does not show any significant effects on the outcome exerted by the funding size. Again, this could be because of the different performances of the funding size in the three programs we examined. For example, even though our three-category scheme works fine in the R01 and R21 programs, it is much less so in the R03 "Small Grant Program" program, which features a much smaller mean funding size per year than the other two programs.

Moreover, in R01 and R21 programs, funding size seems to affect the outcome in opposite ways, as shown in Table 8. In the R01 program, both the mid- and high-funding groups have significantly lower keyword-matching ratios than that of the Group 1. As compared to low-funded projects, the keyword-matching ratio for mid-funded projects is 1.3% lower (CI

[7] We did not consider R03 program in most analyses, because of its low goodness-of-fit.

Table 8
Statistical results concerning funding size Group 2 (mid-level) and Group 3 (high-level) as compared to Group 1.

Funding program Group	Funding size group	Est. Coefficient	Std. Error	*t*-value	*p*-value
R01	Group 2	−0.013	0.004	−2.951	0.003**
	Group 3	−0.019	0.004	−4.358	0.000***
R21	Group 2	0.036	0.008	4.282	0.000***
	Group 3	0.031	0.01	3.184	0.002**

*** $p < .001$; ** $p < .01$; * $p < .05$.

95% [0.9%, 1.7%]) and that for highly-funded projects is 1.9% lower (CI 95% [1.5%, 2.3%]). On the other hand, both of these groups for the R21 program have significantly higher keyword-matching ratios than the low-funded group. The different ways in which funding size affects the outcome might be an important reason why it fails to show any significant impact in the general model and needs to be explained in the future.

6. Conclusions

In this paper, we reported our analysis on the keyword matchedness between NIH research grants and their funded publications. Matches were assessed by applying the RAKE keyword-extraction program to more than 500,000 grant and publication abstracts. Using linear regression method, we tested how the matchedness ratio was influenced by the funding rate of the NIH program in a specific year, the year gap between grants and publications, and the funding size per year of NIH grants.

Our analysis quantifies, for the first time, the evolution of research topics within the same research cycle, as represented in the abstracts of a grant and its linked publications. This is an important first step in understanding how research concepts shift within specific research contexts, which could in turn shed light on how science is conducted under local social and economic conditions. We specifically proposed three hypotheses based on the literature: one that focused on the relationship between ratio of keyword matchedness and funding rate in a given program in a specific year, one concerning the relationship between matchedness ratio and time from project to publication, and one which spoke to the relationship between keyword-matching rate and the funding size.

Even though the funding rate failed to show any significant impact on the outcome in the general model, it is a strong predictor in all three grouped regression models, which suggests its importance for each funding program. We believe this means that, with increased funding pressure, there are indeed faster shifts of research topics between the original grants and publications. On the other hand, our results also proved that there is a near-natural tendency for the conceptual connections between grants and publications to be increasingly distanced over time. However, contrary to our expectation, this trend was not proved in the group of R21 funding program, which needs to be further examined in the future. Similarly, our results also suggested that the impact of funding size on the keyword-matching ratio is also highly moderated by the funding group. For example, in the funding programs of R01 and R21, funding size creates opposite effects on the outcome. This could suggest that researchers being funded by these and other different types of grants have distinct behaviors in terms of how they pursue future research.

Given the exploratory nature of this study, a number of questions remain to be answered by future studies. First and foremost, our efforts are more exploratory than interpretive. Even though we tried to quantify the phenomenon of keyword-matching within the same research cycle in a very large dataset, out results do not point to any specific mechanism of the evolution of scientific vocabularies. The clear connections between increased funding pressure and faster keyword evolving speed seems to suggest that researchers resort to some sort of “bootlegging” strategies in such situations. But it is also possible that they are increasingly driven to find hotter and more innovative topics within the same conceptual framework of the research grant. Both possibilities cannot be ruled out by our results and should be analyzed in qualitative studies in the future.

Second, our results point to many differences between individual NIH funding programs, especially R01 and R21. It should be expected that their different nature (as R21 focusing on exploratory, developmental, and potentially smaller projects than those in R01) could explain part of the differences, such as the relationship between keyword-matching rate and funding size per year. However, finding out the reasons behind their differences would be a great supplement to the mechanisms as demonstrated in the present analysis.

Author contributions

Kai Li: Conceived and designed the analysis; Collected the data; Contributed data or analysis tools; Performed the analysis; Wrote the paper.

Erjia Yan: Conceived and designed the analysis; Wrote the paper.

Acknowledgement

This project was made possible in part by the Institute of Museum and Library Services (Grant Award Number: RE07-15-0060-15), for the project titled "Building an entitybased research framework to enhance digital services on knowledge discovery and delivery." The authors would also like to thank both anonymous reviewers for their help to improve this paper.

Appendix A. Top 30 keywords from paper keyword list (left panel) and grant keyword list (right panel)

Paper keyword	Count of keyword	Ratio among papers	Grant keyword	Count of keyword	Ratio among grants
present study	5179	1.14%	central hypothesis	3157	6.02%
critical role	4701	1.03%	molecular mechanism	3034	5.78%
mouse model	4655	1.02%	mouse model	2417	4.61%
gene expression	4302	0.94%	good understanding	2401	4.58%
high level	4268	0.94%	animal model	2384	4.54%
animal model	3869	0.85%	public health	1680	3.20%
wide range	3200	0.70%	clinical trial	1606	3.06%
cardiovascular disease	3061	0.67%	gene expression	1582	3.02%
molecular mechanism	3043	0.67%	preliminary study	1524	2.91%
oxidative stress	3008	0.66%	preliminary datum	1500	2.86%
risk factor	2931	0.64%	critical role	1498	2.86%
key role	2694	0.59%	cardiovascular disease	1390	2.65%
current study	2658	0.58%	human disease	1206	2.30%
clinical trial	2623	0.58%	large number	1162	2.22%
cancer cell	2566	0.56%	molecular basis	1073	2.05%
breast cancer	2550	0.56%	long term goal	1067	2.03%
high risk	2473	0.54%	immune response	1049	2.00%
central nervous system	2457	0.54%	therapeutic intervention	988	1.88%
body mass index	2328	0.51%	risk factor	982	1.87%
transcription factor	2310	0.51%	immune system	974	1.86%
immune response	2229	0.49%	ultimate goal	958	1.83%
endothelial cell	2166	0.48%	breast cancer	956	1.82%
good understanding	2034	0.45%	high level	948	1.81%
reactive oxygen specie	1975	0.43%	novel approach	931	1.77%
previous study	1966	0.43%	effective treatment	930	1.77%
heart failure	1963	0.43%	human health	928	1.77%
old adult	1913	0.42%	oxidative stress	852	1.62%
cell death	1866	0.41%	cancer cell	830	1.58%
cell type	1857	0.41%	major because	821	1.57%
cell proliferation	1836	0.40%	central nervous system	819	1.56%
early stage	1818	0.40%	early stage	773	1.47%
plasma membrane	1760	0.39%	model system	773	1.47%
central role	1758	0.39%	unique opportunity	730	1.39%
tumor cell	1751	0.38%	transcription factor	705	1.34%
prostate cancer	1729	0.38%	cell type	696	1.33%
extracellular matrix	1705	0.37%	tumor cell	695	1.32%
high rate	1701	0.37%	signaling pathway	694	1.32%
insulin resistance	1674	0.37%	urgent need	687	1.31%
immune system	1665	0.37%	current proposal	678	1.29%
protein level	1600	0.35%	disease progression	644	1.23%
human disease	1599	0.35%	new approach	642	1.22%
endoplasmic reticulum	1548	0.34%	pilot study	636	1.21%
blood pressure	1546	0.34%	molecular level	633	1.21%
physical activity	1509	0.33%	heart failure	624	1.19%
inflammatory response	1474	0.32%	stem cell	616	1.17%
murine model	1467	0.32%	wide variety	613	1.17%
control group	1450	0.32%	endothelial cell	607	1.16%
protein expression	1424	0.31%	general population	589	1.12%
disease progression	1417	0.31%	early detection	586	1.12%
cell line	1410	0.31%	neurodegenerative disease	579	1.10%

References

Ahlin, E. M., & Gibbs, J. C. (2012). The chicken or the egg? The relationship between COP funding and COP research. *Police Practice and Research*, *13*(6), 513–524. http://dx.doi.org/10.1080/15614263.2011.635482
Arora, A., & Gambardella, A. (2005). The impact of NSF support for basic research in economics. *Annales d'Economie et de Statistique*, 91–117.
Beaudry, C., & Allaoui, S. (2012). Impact of public and private research funding on scientific production: The case of nanotechnology. *Research Policy*, *41*(9), 1589–1606.
Ben-David, A. (2008). Comparison of classification accuracy using Cohen's weighted kappa. *Expert Systems with Applications*, *34*(2), 825–832. http://dx.doi.org/10.1016/j.eswa.2006.10.022
Bloch, C., & Sørensen, M. P. (2014). The size of research funding: Trends and implications. *Science & Public Policy*, *42*(1), 30–43.

Boyack, K. W. (2004). Mapping knowledge domains: Characterizing PNAS. *Proceedings of the National Academy of Sciences of the United States*, *101*(Suppl. 1), 5192–5199.
Boyack, K. W., Smith, C., & Klavans, R. (2018). Toward predicting research proposal success. *Scientometrics*, *114*(2), 449–461. http://dx.doi.org/10.1007/s11192-017-2609-2
Boyd, R. N. (1983). On the current status of the issue of scientific realism. *Erkenntnis*, *19*(1–3), 45–90.
Carter, G. M., Winkler, J. D., & Biddle-Zehnder, A. K. (1987). *An evaluation of the NIH research career development award*. Rand Corporation.
Colaco, M., Svider, P. F., Mauro, K. M., Eloy, J. A., & Jackson-Rosario, I. (2013). Is there a relationship between National Institutes of Health funding and research impact on academic urology? *The Journal of Urology*, *190*(3), 999–1003.
Etzkowitz, H. (2008). *The triple helix: University-industry-government innovation in action*. Routledge.
Etzkowitz, H., & Leydesdorff, L. (2000). The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university–industry–government relations. *Research Policy*, *29*(2), 109–123.
Fortin, J.-M., & Currie, D. J. (2013). Big science vs. little science: How scientific impact scales with funding. *PLoS One*, *8*(6), e65263.
Fujimura, J. H. (1987). Constructing "{Do}-{Able}" {Problems} in {Cancer} {Research}: {Articulating} {Alignment}. *Social Studies of Science*, *17*(2), 257–293. Retrieved from. http://www.jstor.org/stable/284951
Gläser, J., & Laudel, G. (2016). Governing science. *European Journal of Sociology*, *57*(01), 117–168. http://dx.doi.org/10.1017/S0003975616000047
Glaser, J., Lange, S., Laudel, G., Schimank, U., Whitley, R., & Engwall, L. (2010). The limits of universality: How field-specific epistemic conditions affect authority relations and their consequences. In *Reconfiguring knowledge production. Changing authority relationships in the sciences and their consequences for intellectual innovation*. pp. 291–324. Retrieved from. https://www.narcis.nl/publication/RecordID/oai:ris.utwente.nl:publications%2Fac2d26bd-ab34-4986-9c73-def848bd154a
Hackett, E. J. (1987). Funding and academic research in the life sciences: Results of an exploratory study. *Science & Technology Studies*, *5*(3/4), 134–147. Retrieved from. http://www.jstor.org/stable/690435
Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). *Diachronic word embeddings reveal statistical laws of semantic change* Retrieved from. http://arxiv.org/abs/1605.09096
Honnibal, M., & Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. pp. 1373–1378. Retrieved from. http://www.aclweb.org/anthology/D15-1162
Hörlesberger, M., Roche, I., Besagni, D., Scherngell, T., François, C., Cuxac, P., . . . & Holste, D. (2013). A concept for inferring 'frontier research' in grant proposals. *Scientometrics*, *97*(2), 129–148. http://dx.doi.org/10.1007/s11192-013-1008-6
Hottenrott, H., & Lawson, C. (2014). Research grants, sources of ideas and the effects on academic research. *Economics of Innovation and New Technology*, *23*(2), 109–133. http://dx.doi.org/10.1080/10438599.2013.814425
Jacob, B. A., & Lefgren, L. (2011a). The impact of research grant funding on scientific productivity. *Journal of Public Economics*, *95*(9), 1168–1177.
Jacob, B. A., & Lefgren, L. (2011b). The impact of NIH postdoctoral training grants on scientific productivity. *Research Policy*, *40*(6), 864–874. http://dx.doi.org/10.1016/j.respol.2011.04.003
Justeson, J. S., & Katz, S. M. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, *1*(1), 9–27. http://dx.doi.org/10.1017/S1351324900000048
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. Retrieved from. http://www.ncbi.nlm.nih.gov/pubmed/843571
Laudel, G. (2006). The art of getting funded: How scientists adapt to their funding conditions. *Science & Public Policy*, *33*(7), 489–504. http://dx.doi.org/10.3152/147154306781778777
Lewison, G. (1998). Gastroenterology research in the United Kingdom: Funding sources and impact. *Gut*, *43*(2), 288–293. Retrieved from. http://www.ncbi.nlm.nih.gov/pubmed/10189860
Lewison, G., & Dawson, G. (1998). The effect of funding on the outputs of biomedical research. *Scientometrics*, *41*(1–2), 17–27. http://dx.doi.org/10.1007/BF02457963
Lewison, G., Grant, J., & Jansen, P. (2001). International gastroenterology research: Subject areas, impact, and funding. *Gut*, *49*(2), 295–302. http://dx.doi.org/10.1136/GUT.49.2.295
Leydesdorff, L., & Etzkowitz, H. (1996). Emergence of a Triple Helix of university—industry—government relations. *Science & Public Policy*, *23*(5), 279–286. http://dx.doi.org/10.1093/spp/23.5.279
R Core Team. (2016). *R: {A} {Language} and {Environment} for {Statistical} {Computing}* Retrieved from. Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org/
Rezek, I., McDonald, R. J., & Kallmes, D. F. (2011). Is the h-index predictive of greater nih funding success among academic radiologists? *Academic Radiology*, *18*(11), 1337–1340. http://dx.doi.org/10.1016/J.ACRA.2011.06.017
Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. In *Text mining: Applications and theory*. pp. 1–20. http://dx.doi.org/10.1002/9780470689646.ch1
Shi, X., Nallapati, R., Lescovec, J., Mcfarl, D., & Jurafsky, D. (2010). Who leads whom: Topical lead-lag analysis across corpora. In *NIPS workshop on computational social science and wisdom of crowds*. pp. 1–5. Retrieved from. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.190.4612
Slaughter, S., & Leslie, L. L. (1997). *Academic capitalism: Politics, policies, and the entrepreneurial university* Retrieved from. Baltimore: Johns Hopkins University Press. https://www.worldcat.org/title/academic-capitalism-politics-policies-and-the-entrepreneurial-university/oclc/36017199
Slaughter, S., & Rhoades, G. (2004). *Academic capitalism and the new economy: Markets, state, and higher education* Retrieved from. Baltimore: Johns Hopkins University Press. https://www.worldcat.org/title/academic-capitalism-and-the-new-economy-markets-state-and-higher-education/oclc/53443009
Svider, P. F., Mauro, K. M., Sanghvi, S., Setzen, M., Baredes, S., & Eloy, J. A. (2013). Is NIH funding predictive of greater research productivity and impact among academic otolaryngologists? *The Laryngoscope*, *123*(1), 118–122. http://dx.doi.org/10.1002/lary.23659
Venable, G. T., Khan, N. R., Taylor, D. R., Thompson, C. J., Michael, L. M., & Klimo, P. (2014). A correlation between National Institutes of Health funding and bibliometrics in neurosurgery. *World Neurosurgery*, *81*(3–4), 468–472. http://dx.doi.org/10.1016/j.wneu.2013.11.013
Yan, E., Wu, C., & Song, M. (2018). The funding factor: A cross-disciplinary examination of the association between research funding and citation impact. *Scientometrics*, *115*(1), 369–384. http://dx.doi.org/10.1007/s11192-017-2583-8
Yan, E., & Zhu, Y. (2018). Tracking word semantic change in biomedical literature. *International Journal of Medical Informatics*, *109*(215), 76–86. http://dx.doi.org/10.1016/j.ijmedinf.2017.11.006