


Automated Research Impact Assessment: a new bibliometrics approach

Christina H. Drew¹  · Kristianna G. Pettibone¹ ·
Fallis Owen Finch III² · Douglas Giles² ·
Paul Jordan³

Received: 13 May 2015 / Published online: 30 January 2016
© Akadémiai Kiadó, Budapest, Hungary 2016

Abstract As federal programs are held more accountable for their research investments, The National Institute of Environmental Health Sciences (NIEHS) has developed a new method to quantify the impact of our funded research on the scientific and broader communities. In this article we review traditional bibliometric analyses, address challenges associated with them, and describe a new bibliometric analysis method, the Automated Research Impact Assessment (ARIA). ARIA taps into a resource that has only rarely been used for bibliometric analyses: references cited in “important” research artifacts, such as policies, regulations, clinical guidelines, and expert panel reports. The approach includes new statistics that science managers can use to benchmark contributions to research by funding source. This new method provides the ability to conduct automated impact analyses of federal research that can be incorporated in program evaluations. We apply this method to several case studies to examine the impact of NIEHS funded research.

Keywords Bibliometrics · Automated impact analysis · Research evaluation · Science of science management

JEL Classification I2 Education and Research Institutions

Electronic supplementary material The online version of this article (doi:[10.1007/s11192-015-1828-7](https://doi.org/10.1007/s11192-015-1828-7)) contains supplementary material, which is available to authorized users.

✉ Christina H. Drew
drewc@niehs.nih.gov

¹ Program Analysis Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, USA

² Open Intelligence, Inc., Raleigh, NC, USA

³ Office of Extramural Research, Office of Data Analysis Tools and Systems, National Institutes of Health, Bethesda, MD, USA

Introduction

Increasingly, federal agencies in the United States use program and research evaluations to inform program planning, identify areas for program improvement, document outputs and impacts, and justify the existence of programs and budgets (Haak et al. 2012; Hicks et al. 2004; Kostoff 1995; Milat et al. 2015). To address these needs, federal evaluators have developed strategies and methods for conducting portfolio and program evaluations (Hicks et al. 2004; Howell and Yemane 2006; Koplan et al. 1999; Lane 2010; Lane and Bertuzzi 2011; National Academy of Sciences (NAS) 2001, 2004; Quinlan et al. 2008; Srivastava et al. 2007; U.S. Government Accountability Office 2012). These strategies often rely on counts of activities conducted and outputs produced. Analyses of outputs often include counts of patents and bibliometric analyses of publications. In this paper, we describe a new method for automating bibliometric impact analysis, illustrating its utility with several case studies. We begin with a description of traditional bibliometric analyses and a discussion of the challenges associated with them. We also review examples of manual approaches that have been used to analyze the impact of publications, and describe strengths and challenges of the proposed approach.

Traditional bibliometric analyses

One of the primary methods for evaluating scientific research is bibliometric analysis. Evaluative bibliometrics is the use of publication and citation analyses as indicators of science output (Borgman and Furner 2002; Campbell et al. 2010; Moed 2005; Narin 1976). Citation counts tell us how many times other researchers cite a particular article (Martin 1996; Narin 1976; Reuters 1994), citer analysis provides an idea of how many researchers have cited the work, versus how many times the work has been cited (Ajiferuke and Wolfram 2010) and impact factors provide a proxy measure for the quality or impact of a journal on science by counting how often articles from that journal are cited by others (Campbell et al. 2010; Martin 1996; Reuters 1994).

Limitations of bibliometric analyses

Publishing an article in a peer-reviewed journal implies a certain level of quality; however, evaluators should use bibliometric analyses carefully because of the many factors that influence citation patterns (Borgman and Furner 2002). A number of authors have provided thoughtful reviews of the limitations of bibliometric analyses (Kostoff 1998; Lane 2010; Moed 2005; Phelan 1999). These limitations include the inability to compare bibliometric measures across fields, programs, or countries (Lane 2010); a tendency to favor older researchers (Lane 2010); political motivations (Kostoff 1998); and a susceptibility to bias and artificial influences, such as self-citations (Borgman and Furner 2002; Campbell et al. 2010; Kostoff 1998; Pasterkamp et al. 2007). Citation counts also include articles that are criticizing or discrediting research (Kostoff 1998; Lane 2010), so high citation counts may just reflect significant argument around a research idea, not necessarily a high impact scientific finding.

Many methods for analyzing bibliometric data emphasize the importance of linking electronic references to publications, people, groups, organizations, and nations (Borgman

and Furner 2002). However, these methods continue to treat the publication as a product or output of the research, rather than assessing the longer-term impact of the publication on a particular field.

Measuring impacts

For purposes of program and portfolio evaluation, we define impacts to be the benefits or changes resulting from scientific research, program activities or outputs (National Institute of Environmental Health Sciences (NIEHS) 2012). Impacts, often called outcomes, are the effects of the research on the research field or within society (The Kellogg Foundation 2004). Outputs are typically defined as the tangible products of research, such as publications and patents. Although identifying and reporting impacts can be crucial in making the case for program success, measuring impacts presents several challenges (Hughes et al. 2013; NIEHS 2012). Impacts are more difficult to measure than activities and outputs in part because it often takes several years for substantive changes to occur and the program may not be in existence by the time the impacts are realized (Guthrie et al. 2005; Orians et al. 2009; Teles and Schmitt 2011). Researchers themselves may have a limited interest in expending resources to track and measure impacts if funding for the project ends before the impacts are realized. The complexity of an impact also may make it difficult to attribute it to a specific grant or project (Guthrie et al. 2005; Stuart 2007; Teles and Schmitt 2011). In addition, data to evaluate long-term impacts are not easily available for quantitative analysis, and instead typically require intensive qualitative analysis methods (Orians et al. 2009).

Another challenge in tracking and assessing research impacts is that researchers may be hesitant to claim credit for impacts because other organizations or events may have contributed to the changes. While researchers may not be able to claim sole credit for these impacts, it is important to be able to track these broader changes and to document the contributions made by the project to achieving long-term impacts.

Precedents for analyzing the impact of publications

Early efforts to quantify the impact of basic science discoveries on technology and innovation provide the foundation for modern bibliometric analyses (Author unknown 1970; IIT Research Institute 1968). More recently, researchers developed methods for tracking the impact of articles in order to “track the documented flow and evolution of research over time until the linkages to far downstream products can be identified” (Kostoff 1998, p 29). However, Kostoff also warned that the process would be slow and laborious and would require users to make judgments about the appropriateness, quality and quantity of the impact. Kostoff concluded that traditional bibliometric measures can provide indicators of the productivity of a researcher and to some extent, an indicator of the quality of his or her research (Kostoff 1998).

A few researchers have used manual methods to measure the number of times a publication was cited by others in non-research products, and thus provide another precedent for our proposed methodology. In one approach the authors manually analyzed the research cited in patent applications (National Research Council (NRC) 1998; Roessner et al. 1998). In another approach authors assessed the number of times an agency or organization is

cited on a credible website (Hicks et al. 2004). Yet another group of researchers relied on resource intensive case study methods that coded the cited research for how important it was to the citing paper (Hanney et al. 2005; Jones et al. 2012; Wooding et al. 2005).

While these approaches lay the foundation for thinking of the downstream impact of the research presented in publications, they are resource intensive and require in-depth, time-consuming, qualitative assessment. Thus far, there has been little to no attempt to automate the process that links the research published in peer-reviewed literature to a funding source. Moreover, we find that only the few authors cited above have treated publications as impacts rather than outputs.

Automated research impacts assessment: a new approach

For many years, NIEHS has been interested in understanding the impacts of our investments in research grants.¹ Research program impacts typically include items such as improving health, reducing adverse environmental exposures or changes to environmental health regulations and policies. Early efforts in assessing impact led to the development of the Scientific Publication Information Retrieval and Evaluation System (SPIRES) (Boyack and Jordan 2011), which links peer-reviewed publications to the NIH-funded grants acknowledged in those articles, and is accessible to the public through the RePORTER website (NIH 2014a). This technology helped to enable bibliometric analyses at NIEHS and then advance those analyses across all of NIH. Formerly available tools such as the electronic Scientific Portfolio Assistant (eSPA) also linked downstream papers to grants supported by NIH (Haak et al. 2012).

Additionally, we have used logic models to conceptualize the full range of activities, outputs and impacts of our programs (Engel-Cox et al. 2008; Liebow et al. 2009; NIEHS 2012; Orians et al. 2009). However, in working with scientific program staff at NIEHS to develop logic models for scientific research programs, we found that it was particularly difficult to measure impacts. One key challenge was the lack of an automated method for tracking or analyzing impacts. To gain insights into the impacts of our asthma portfolio, for example, we found that we had to manually collect data on the impacts of asthma research from the researchers using resource intensive surveys and interviews (Orians et al. 2009). Other challenges include the length of time it takes to affect change after a research investment compared to the relatively short duration of a grant, and the increasing influence of contextual factors over time.

Given the expense of collecting and analyzing impact data, we sought a more automated approach to assess the impact of research investments in a specific area. Our first step was to identify an *important artifact* that relied on NIEHS' research in its conclusions or recommendations. For purposes of this discussion, we define important artifacts to be published materials that reflect high impact research, decisions or policies that have the ability to influence medicine and public health. Examples of important artifacts include documentation of policy and regulatory decisions, clinical and treatment guidelines, other major decision or guidance documents, or reference works from authoritative sources (such as the National Academies of Science or the Institute of Medicine) that can be used at a personal, community, regional, national, or international level to influence change.

¹ We use the word “grant” in this paper broadly, to include both projects that are conducted internally at NIH as well as “extramural” research that occurs beyond the walls of NIH.

With the rise of transparency and accountability, we observed that important artifacts are likely to have detailed lists or databases of references to authenticate the conclusions. Such databases yield a largely untapped resource for impact analysis. In 2008, Congress mandated that all papers reporting research supported by NIH-funds should acknowledge such funding and be made accessible to the public. The SPIRES tool links these peer-reviewed publications to NIH grants and thus provides us with a means to look at NIH grant support for virtually any list of publications. We propose in this paper that evaluating the funding sources for a list of references from an important artifact will yield useful insights into the contribution of NIH supported research to that artifact. And since a typical grant number includes information about which NIH Institute, Center or Office (ICO) has provided the primary funding, we can dig even deeper to look at the relative contributions of various ICOs to that artifact. The approach described below builds on the literature that uses bibliometric analyses to analyze the impact of research on important artifacts (Lewison et al. 2005; Leyedesdorff 1998; Jones et al. 2012; Wooding et al. 2005) and uses the existing NIH SPIRES bibliometric tool to automate the process.

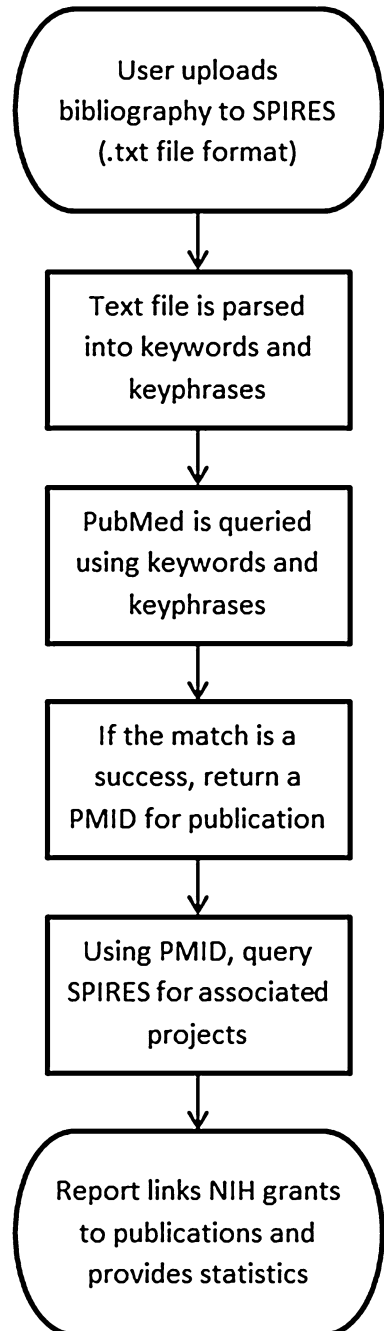
The Automated Research Impact Assessment² (ARIA) method proposed here leverages existing bibliometric tools (SPIRES) that link publications to NIH research grants in order to analyze the peer reviewed literature referenced in important artifacts. As part of the method, we developed a new parsing interface in SPIRES called the Reference Parsing and Retrieval Service (RePARS) as well as a number of novel bibliometric statistics that quantify the influence of NIH- and NIEHS-funded research on selected impacts. For example, we can use the ARIA method to review the references listed in a key piece of environmental health policy, identify those that acknowledge NIEHS funding support for that research, and compare them to the number of references that acknowledge other NIH ICOs.

Methods

Once an important artifact is identified, we employ a six-stage process to assign funding sources to each reference included in the data set (Fig. 1).

- (1) The user creates a text (.txt) file from the bibliography of the artifact to upload into SPIRES. The.txt file does not have to be formatted or ordered in a particular way. The only requirement is that it is machine readable text. Special characters (e.g. von B | ddingen vs. von Büdingen) can affect the accuracy of parsing and PubMed matching.
- (2) Text files are parsed into component parts (i.e., extracted into structured data fields) using two open source tools—Biblio::Citation::Parser from ParaTools, and ParsCit (Kan 2010; ParaTools 2004) as well as a custom script (written in Perl). Each reference is parsed by all three parsers and the most complete results are selected for use in the rest of the process.

² Kostoff used the term research impact assessment to describe a broad range of methods to evaluate research, including peer review, retrospective methods, bibliometrics, co-occurrence, cost-benefit and economic analyses, and network analyses (Kostoff 1995). We coined the phrase automated research impact assessment without knowing that the term was already in use, but feel that there is a good fit between our new bibliometric method and Kostoff's vision for broader research impact assessment activities.

Fig. 1 ARIA methodology

The following fields are extracted, when possible, from each reference:

- Publication title
- Publication year
- Authors
- Journal name
- Volume
- Pages

A reference is considered “parsable” if the publication title, publication year and authors can be identified. Currently we are not using the journal name, volume or page values that are parsed from the references to identify or exclude publications in the set that is analyzed by the RePARS tool, but future iterations of the tool may expand to use these fields.

- (3) Publication title and publication year are used to find the publication in PubMed using ESearch and EFetch (Sayers 2009). We found that using more information (journal name, publication year, volume, first page and author) with ECITMatch (Sayers 2009) reduced the accuracy of the records identified. Not all references from the artifact’s bibliography can be included in the analysis and are actively removed from subsequent statistical calculations under the following scenarios:
 - (a) The parser does not identify a match in PubMed to the title or author of the reference. Not all references (e.g. material printed in books or non-biomedical journals) are included in the PubMed database.
 - (b) No PubMed ID is assigned to the reference. SPIRES requires a PMID in order to link to NIH grants.
 - (c) The year of the reference is before 1980. These are removed because the PubMed data prior to 1980, and in particular the links to NIH grants during that time frame, are perceived to be of questionable quality.
- (4) If the match is a success, the query returns the PubMed ID number for the publication.
- (5) The PubMed ID number is searched in SPIRES for associated NIH grants.
- (6) A report is generated (in MS Excel format) providing both raw data and a series of statistics that summarizes how the references in the list were matched to PMIDs and linked to NIH grants. Separate worksheets within the export file provide statistics for each NIH ICO that is included in the list of matched publications. We use NIEHS below as an example.
 - A list of the references, grant numbers associated with each reference, and NIH Institutes or Offices associated with each reference³
 - Total number of references submitted, the initial sample size
 - Total number of references not parsed, those that are removed from the sample
 - Number with unmatched title and author data
 - Number with no PMID
 - Number with publication dates before 1980 (publication data before 1980 were considered too poor a quality to analyze)

³ See supplemental materials for an example of the raw data generated automatically in this report.

- Total number of references that are analyzable, which becomes a key denominator for the remaining statistics

$$\text{Total \# of references submitted} - \text{Total \# not parsed}$$

- Total number of references that acknowledge an NIH grant
- Total number of references that acknowledge an NIEHS grant
- % of references that acknowledge an NIH grant, a measure of NIH's impact on the artifact:

$$\frac{\text{Total \# of references that acknowledge an NIH grant}}{\text{Total \# of references that are analyzable}}$$

- % of references that acknowledge an NIEHS grant, a measure of NIEHS's impact on the artifact:

$$\frac{\text{Total \# of references that acknowledge an NIEHS grant}}{\text{Total \# of references that are analyzable}}$$

- % of NIH-funded references that acknowledge NIEHS funding, a measure of the relative contribution of NIEHS funding, compared to the NIH for this artifact:

$$\frac{\text{Total \# of references that acknowledge NIEHS support}}{\text{Total \# of references that acknowledge NIH support}}$$

Case study examples

To demonstrate the utility of this new bibliometric method, we analyzed six important artifacts associated with NIEHS research that were identified by NIEHS scientific program staff as pilot case studies. These artifacts were reviewed for three criteria: plausibility, credibility and importance, which we propose as essential features of a reasonable analysis. To meet the condition of plausibility, the important artifact had to be reasonably connected to NIEHS funded research. In other words, it occurred after a known NIEHS investment in the scientific topic area and is relevant to the NIEHS mission. To meet the condition of credibility, the artifact had to be produced or published by an organization deemed to have a widely accepted authoritative presence in the fields. Examples could include regulatory agencies like the U.S. Environmental Protection Agency (EPA) or the U.S. Food and Drug Administration (FDA) or renowned research organizations such as the World Health Organization (WHO), the Institute of Medicine, the International Agency for Research on Cancer, or the Pew Charitable Trust. To meet the condition of importance, the artifact had to make a significant contribution to the field of environmental health science. Below, we describe each of the artifacts and present the bibliometrics resulting from the ARIA.

Case study selections

We selected two different types of important artifacts for the case studies presented in this paper. First, we chose to look at national and international artifacts about arsenic, an environmental contaminant that is internationally recognized as a human health concern. Arsenic is a naturally occurring element and can be found in groundwater and surface water, as well as in many foods. Exposure to arsenic is associated with cancer, cardiovascular problems, and neurological effects. Several organizations have released findings

and recommendations on maximum contaminant levels for arsenic, including the NRC and the WHO. The NRC aims to improve government decision making and public policy, increase public understanding, and promote the acquisition and dissemination of knowledge in matters involving science, engineering, technology, and health. The WHO has a goal of ensuring that all people have the right to have access to an adequate supply of safe drinking water. We selected two key reports from NRC (1999, 2001) and another from WHO (2011) that were specifically related to arsenic in drinking water. The 1999 report was drafted to “independently review the arsenic toxicity data base and evaluate the scientific validity of EPA’s 1988 risk assessment for arsenic in drinking water” (NRC 1999, pg 1). The 2001 report updated “the scientific analyses, uncertainties, and findings of the 1999 report on the basis of relevant toxicological and health-effects studies published and relevant data developed since the 1999 NRC report.” It also evaluated “the analyses subsequently conducted by EPA in support of its regulatory decision-making for arsenic in drinking water” (NRC 2001, p. 2). The 2011 report from WHO provided a fourth update to the Guidelines for Drinking Water, incorporating new data generated 2008–2010. This case is plausible because there is significant evidence that arsenic in drinking water affects public health, credible because the artifacts are from prominent internationally recognized research and health organizations, and important because the artifacts summarize the state of the science at three different time periods.

We selected our second set of artifacts from the U.S. EPA (a credible source). This US regulatory agency employs standardized integrated science assessments (ISAs) to review, synthesize and evaluate the relevant science to establish environmental policies and regulations (important). ISAs are based on literature reviews of published research. The EPA conducts the ISAs and provides a publicly available database of references that are cited in the ISA, on their Health and Environmental Research Online website (www.hero.epa.gov).

Table 1 ARIA metrics for arsenic artifacts

	NRC Arsenic Report, 1999	NRC Arsenic Report, 2001	WHO Arsenic Report, 2011
Total # of references submitted	813	304	71
Total # of references that could not be analyzed			
Title, author or publication year could not be determined	11	20	3
Published before 1980	197	14	7
PMID could not be determined	221	52	25
	429	86	35
Total # of references that are analyzable	384	218	36
Total # of references that acknowledge an NIH grant	66	56	0
Total # of references that acknowledge an NIEHS grant	48	49	0
% of references that acknowledge an NIH grant	(66/384) 17 %	(56/218) 26 %	(0/36) 0 %
% of references that acknowledge an NIEHS grant	(48/384) 13 %	(49/218) 23 %	(0/36) 0 %
% of NIH references from NIEHS	(48/66) 73 %	(49/56) 88 %	(0/0) 0 %

The full output for the NRC 2001 Arsenic report can be found online in the supplemental material file. Other outputs can be made available upon request to the authors

We analyzed the impact of NIEHS-funded research on the ISAs for carbon monoxide, lead, and particulate matter, three environmental exposures that have been linked to health outcomes (plausible) (U.S. EPA 2009, 2010, 2012).

Case study results

The ARIA metrics for the case studies (Tables 1, 2) invite an array of exploratory observations. First, the number of references cited in each of the artifacts ranges widely. For example, the total number references for artifacts for the first case are 813, 304 and 71 (Table 1), and the references for the second case are 4686, 179 and 625 (Table 2). This raises questions about the comparability of the references across artifacts. In addition, there is quite a bit a variance in the number of references that cannot be parsed (from about 2 % to more than 50 %). From these examples, it seems common to remove about a third to half the references from the analysis. These un-parsable references may include books, conference proceedings, reports, or other documents from federal agencies such as EPA or from organizations such as the WHO. See further discussion about unparsed references below (Table 3).

With regard to specific funding sources displayed in Table 1, the two NRC reports have relatively similar percentages of references that cite NIH funds (17 and 26 %, respectively). Interestingly, the NRC 2001 report indicates that a greater percentage of the total number of references were funded by NIEHS (nearly double—from 13 to 23 %) and a higher proportion of the NIH funded references were from NIEHS (from 73 to 88 %), suggesting that the influence of NIEHS research in these documents has increased somewhat over time. On the other hand, the WHO report that we reviewed had many fewer

Table 2 ARIA metrics for EPA ISAs

	EPA Particulate Matter ISA, 2009	EPA Carbon Monoxide ISA, 2010	EPA Lead (Pb) ISA, 2012
Total # of references submitted	4686	179	625
Total # of references that could not be analyzed			
Title, author or publication year could not be determined	12	0	25
Published before 1980	58	5	8
PMID could not be determined	1778	0	195
	1848	5	228
Total # of references that are analyzable	2838	174	397
Total # of references that acknowledge an NIH grant	810	31	9
Total # of references that acknowledge an NIEHS grant	663	4	9
% of references that acknowledge an NIH grant	(810/2838) 29%	(31/174) 18%	(9/397) 2%
% of references that acknowledge an NIEHS grant	(663/2838) 23%	(4/174) 2%	(9/397) 2%
% of NIH references from NIEHS	(663/810) 82%	(4/31) 13%	(9/9) 100%

Table 3 Breakdown of references that could not be analyzed for the NRC arsenic report (2001)

		References in the NRC Arsenic Report (2001)	
Total # of references submitted to ARIA		304	
Published before 1980		14	
Never in PubMed			
	Gray literature	35	
	Book	13	
	Thesis	3	
	Foreign journal or language	3	
	Database	2	
Total		56	
Errors			
	Read errors	5	
	Parsing errors	11	
	Parsing errors fixed by analyst during data cleaning	38	
Total		54	
Total # of references that are analyzable		180	
		Before Data Cleaning	After Data Cleaning
Total # of references that could not be analyzed		124	86
Percentage of un parsed references that are parsing errors		(49/124) 40%	(11/86) 13%
Percentage of all references that are parsing errors		(49/304) 16%	(11/304) 4%

After data cleaning, only 4 % of references cited in one of the arsenic reports (NRC 2001) were found to be parsing errors

references overall and no references cited funding by NIH or NIEHS. We also observe that the NRC 2001 report builds on the earlier report from 1999, and the latter report does not cite every reference that was included in the former. The WHO 2011 report lists only 72 references, but includes reference to both the NRC arsenic reports. This poses a difficult analytical challenge: Should all the references from the NRC reports be “included” in the counts for the WHO 2011 report? Additional technologies that electronically link reference citations over time could be helpful in answering such questions.

An examination of the ISA results from EPA (Table 2) indicates relatively few references from NIH (results range from 2 to 29 %). We find this somewhat surprising, given that we would generally expect that NIEHS-funded research would be cited in EPA assessments. We also see that the relative contributions from a particular NIH ICO, in this case NIEHS, also varies. We see, for example, NIEHS’s contribution to the particulate matter and lead ISAs (at 82 and 100 % respectively) are much higher than the ISA for

carbon monoxide (just 13 %). These differences may indeed reflect differences in research investments, or in the impacts of those investments. The variability across similar artifacts from the same agency raises many questions that will guide future inquiries. For example, is there a critical mass of references that are needed in order to have a credible analysis? Is there a discernible pattern in terms of which studies were cited? Is it possible that the automated tool missed key publications? Future studies will no doubt explore these questions in more detail.

ARIA's strengths outweigh challenges

The ARIA approach provides a novel way to automate the quantitative analysis of research impacts, and can be particularly effective when used as part of a mixed method approach. ARIA can be used for those starting from an agnostic perspective, looking to explore possible relationships. Alternatively, the ARIA method can also be used to test hypotheses and confirm a relationship between publications, agency research funding and a specific impact. Before using the ARIA method, we encourage researchers to ensure that the artifacts are relevant, important and from a credible source.

Strengths of the ARIA method include a limited opportunity for bias, the ability to examine long-term impacts, and significant cost savings through automation. Challenges include typical bibliometric limitations, issues with NIH funding acknowledgements, capturing the right references using the RePARS tool, and limits to the PubMed dataset. A host of questions have yet to be explored.

Limited bias

Rather than relying on researcher recollection or burdening researchers with tracking their impacts and making judgment calls about what impact their research has had, the process of reviewing the sources cited by a key piece of research or policy is largely unbiased and easily quantifiable. It is possible that the assessments could be biased by the artifacts chosen. Some agencies, for example, might choose to select artifacts that are systemically either more or less likely to reference their work. However, making the RePARS tool available to all NIH users may encourage others to run similar analyses, potentially fostering a self-regulating culture, such as that found on Wikipedia (where multiple perspectives entered by multiple authors keep the tool accurate).

Analysis of long-term impacts

The ARIA method also provides an opportunity to analyze the long-term impacts of NIH funding. Quantifying the contribution of NIH research to changes in health outcomes, such as reductions in cancer mortality or asthma can be challenging. ARIA provides a strategy for assessing the role of NIH research funding in the development of important artifacts that modify or quantify those health outcomes, such as policies, regulations, clinical guidelines or reference documents. As we move forward into the future, better documentation will improve our ability to link research over longer and longer periods of time.

Cost savings through automation

The resources needed to conduct the ARIA analysis are minimal. The user is not required to know anything about the PubMed system in order to obtain results. The greatest expense is often associated with cleaning the dataset before it is parsed. Once the set is complete, the system automatically parses results, matches references to PubMed, identifies results, calculates statistics and summarizes the results in a series of tables. This is all done in a matter of hours, not days or weeks as would be required with a manual assessment. The automated feature fosters scalability, allowing analysts to obtain results for 20 or 2000 references. Cost savings of ARIA, compared to other methods that require manual data collection or analysis, are thus substantial.

General bibliometric challenges

Although ARIA is uniquely suited to efficiently assess the impact of federal research funding (biomedical), there are several limitations users would need to consider. First, as discussed above, there are many known limitations with bibliometric methods and these all apply to ARIA as well. A more unique challenge is that this method is only applicable for important artifacts that cite the evidence for their decisions by including a bibliography or database of the sources used to arrive at that decision. Many important artifacts, such as laws, proposed bills and other policies do not typically include references. Furthermore, those artifacts that do include references are not provided specifically for the purpose of assessing the value of relevance of these references.

Additionally, if the artifacts list references by chapter, users will need to manually review the citation list to remove duplicate citations. Our experience found that electronically removing duplicates using Excel or similar programs did not identify all duplicates, as often there were minor typographical errors that would prevent a machine from identifying a duplicate citation.

NIH funding acknowledgements are imperfect

The ARIA method requires that the references properly acknowledge the source of funding and that these data are stored in PubMed. Historically this has been a challenge. Boyack and Jordan (2011) document 16 different ways a single NIH project was referenced in peer reviewed journal articles over the 20-year life of that grant. SPIRES and RePARS account for many of these anomalies, but obviously cannot address an author's omission of any acknowledgment to a particular NIH-funded grant. In 2009, however, the U.S. Congress mandated that federally funded research be properly acknowledged and made publically available within 1 year of publication in PubMed Central (NIH 2014b). This mandate has likely reduced the frequency that grant numbers are not cited properly and increased the overall percentage of publications that cite grants. We expect funding source data to further improve as NIH strengthens its compliance monitoring, develops tools to correct improper linkages, and begins to withhold funds for non-compliance; and as journals increasingly capture sources of support in structured data fields. Although the funding source data is imperfect, given the size of the resource (SPIRES indexes over 2 million publications from PubMed and PubMed Central) and the availability of RePARS to NIH personnel through centralized web-based applications, we believe that it is worth exploring. It is possible to make useful conclusions from imperfect data.

Capturing the right references

RePARS (the parsing tool at the core of the ARIA method) also poses several challenges. RePARS relies on SPIRES to capture publication data. SPIRES may miss publications that it should catch or it may include publications that are not correctly matched to the reference. As Digital Object Identifiers (DOI) and PubMed IDs are increasingly used by journals and as part of citations, the match rates are likely to improve. Future enhancements to RePARS could potentially include DOI as part of the parsing strategy.

The potential for incorrectly matching references listed in a source file with actual papers available in PubMed also exists. We manually reviewed (side-by-side on screen comparisons) the list of references from the original artifacts to the corresponding output from RePARS and are reasonably convinced that the PubMed matches are accurate. In all the RePARS output examples we have worked with (going well beyond the 6 documents parsed for this paper), we have never observed any cases where a matched reference is pointing to an entirely different title or list of authors.

It is possible that some of the references removed from the analytical calculations within RePARS also have matches and excluding them from the analysis could skew the results. A benefit of matching references against the PubMed database, is that it includes all PubMed Central publications as well (PubMed Central is the online repository that enables the public access required Congress). This means that more recent articles published in non-biomedical journals are included in the matching algorithm because they are part of PubMed Central.

A related challenge is that the RePARS process removes publications that cannot be matched. This sometimes results in removing a large percentage of the references cited in an important artifact from the analysis, and perhaps introducing bias. The wide variability in the number of references in each case that can't be parsed also causes some concern. To better understand why some references are not analyzed, we manually coded the 86 references that were excluded from the NRC 2001 Arsenic Report output (Table 3 and supplemental file). Note that we began this assessment on a data set that had already been partially cleaned by the analyst. In 38 cases, PubMed IDs were substituted for the original text and by the time Table 3 was generated, these 38 references were correctly analyzed by the RePARS tool.

Most of the 86 references that the tool marked as “not analyzed” (after data cleaning) were published before 1980 or were considered grey literature (mainly documents produced by federal agencies), book chapters, theses, or articles published in international journals or languages. We considered a total of 16 of the 86 exclusions to be “errors.” Two types of errors were observed: read errors (5) and parsing errors (11). Read errors occur if references are not perceived by the parser as a single entity, in other words if partial references were imported. This happened in five cases for this particular NRC report—and is likely due to extra carriage returns in the imported data file. The analyst identified issues as parsing errors when she could see that (a) the full reference had been imported properly, (b) it did not fit into one of the other categories, and (c) there was reasonable expectation that the journal in the reference would be indexed in PubMed. Eleven instances of “references not analyzed” (just 14 % of the references not matched and only 4 % of the total number of references in the set) were identified as parsing errors. Even if we assume that all the references cleaned by the analyst early in the process were parsing errors, the parsing error rate is still relatively low (16 %). Ideally, as more NIH analysts use the tool we expect to be able to improve the quality of the parsing as well.

Going beyond PubMed

The first version of the RePARS tool has focused solely on searching PubMed and PubMed Central data for funding source information. Thus, impacts that are not noted in publically available documentation will be excluded from the analysis. In the future, additional data sets, such as iEdison (for patents) (NIH 2015a), the NIH Clinical Trials database (NIH 2015b), Health Services/Technology Assessment Texts (National Library of Medicine 2015), and others, could be added to the searching algorithms to strengthen the tool. Several international efforts to assess impact (IMPACT-EV 2014; Jump 2013; Research Excellence Framework 2014; Researchfish 2014) have been developed and could eventually be added. Additionally, when resources allow, we recommend mixing the ARIA approach with other research assessment methods like interviewing researchers and collecting case studies of research impact.

Many questions are yet to be explored

While ARIA is a useful new method for analyzing impacts, we expect that the findings from these analyses will serve to generate even more questions and hypotheses. The analysis method seems to invite as many questions as it might answer. Interpreting the meaning of the results may be difficult until we begin using ARIA more broadly across programs and NIH institutes/offices. We have yet to explore how the method does or does not contribute to the challenging issue of the “counterfactual,” or establishing what would have happened in the absence of the research. Furthermore, additional enhancements to the method are easy to envision. For example, we may explore how to analyze the relationship between a funded research portfolio and its contribution to a large dataset such as the Epigenomics Roadmap Consortium (Bernstein et al. 2010). We would also like to explore the possibility of including inputs, such as financial investment, in the analyses to enable more direct assessments of costs and benefits. We expect to work with researchers to understand whether additional databases that include NIH funding acknowledgements could be added to the matching algorithms. Additionally, new strategies to classify references as support “for” or “against” a scientific finding could help strengthen the approach.

Finally, we anticipate exploring the feasibility and potential utility of connecting multiple generations of artifacts in future iterations of the RePARS tool. As the NRC examples seem to indicate (see discussion of Table 1 above), references build upon each other over time, and the funding source may be obscured in subsequent references. For example, when a reference citing NIEHS as a funding source appears in a 2004 report, and an update to that report 5 years later cites the 2004 report but not the original reference, the source for the support is lost. Automated bibliometrics may help us attribute that funding source through time.

On balance, we believe that these limitations are more likely to result in a conservative estimate of the impact of the research, rather than an exaggeration of the impact; and thus the strengths of this approach outweigh the limitations.

Conclusion

The ARIA method presents a new automated approach to evaluate research impact. The more strategies we have for evaluating impact, the better understanding we will have of the value of that research. This bibliometric impact tracking method provides an inexpensive, quantitative method that will add to the NIH's toolbox of approaches that together help demonstrate the value of scientific research investments. The approach includes new statistics that science managers can use to benchmark contributions to research by funding source. In contrast to many current methods that start with an NIH or ICO investment and make attempts to analyze impact after the passage of time, ARIA facilitates a retrospective approach, where we begin with a significant artifact and examine the evidence of NIH or ICO investments in the development of that artifact. Working the problem both backwards and forwards helps to ensure a more complete picture of research impacts. As with all new approaches, time is needed to determine its full utility. We look forward to working with the NIH analysis community and with researchers to monitor and track these new bibliometric measures and establish potential benchmarks for using them into evaluate research portfolios.

Acknowledgments This work was supported by the National Institute of Environmental Health Sciences (NIEHS). The authors would like to acknowledge the work of Sheila Newton and Raymond Grissom, Jr., of the NIEHS Office of Planning and Policy Evaluation, who conducted an early manual review of an EPA Ozone Regulation as a test of this new bibliometric research method. Many thanks also to James Corrigan (National Cancer Institute) who provided comments on early versions of this paper; and to James Onken and Brian Haugen in the NIH Office of Extramural Research, Office of Data Analysis Tools and Systems, who currently manage the SPIRES database (and RePARS tool) for the National Institutes of Health.

References

- Ajiferuke, I., & Wolfram, D. (2010). Citer analysis as a measure of research impact: Library and information science as a case study. *Scientometrics*, *83*, 623–638. doi:10.1007/s11192-009-0127-6.
- Author Unknown. (1970). A trace of "traces". *Mosaic Magazine Science Articles Archive*, *1*(1), 14–19. Retrieved from http://www.mosaicsciencemagazine.org/index.php?mode=article&pk_magazine=109. Accessed 19 March 2015.
- Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., et al. (2010). The NIH roadmap epigenomics mapping consortium. [opinion and comment]. *Nature Biotechnology*, *28*(10), 1045–1048. doi:10.1038/nbt1010-1045. Retrieved from <http://www.nature.com/nbt/journal/v28/n10/full/nbt1010-1045.html>. Accessed 17 March 2015.
- Borgman, C. L., & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology*, *36*, 3–72.
- Boyack, K. W., & Jordan, P. (2011). Metrics associated with NIH funding: A high-level view. *Journal of the American Medical Informatics Association*, *18*(4), 423–431. doi:10.1136/amiajnl-2011-000213.
- Campbell, D., Picard-Aitken, M., Cote, G., Caruso, J., Valentim, R., Edmonds, S., et al. (2010). Bibliometrics as a performance measurement tool for research evaluation: The case of research funded by the National Cancer Institute of Canada. *American Journal of Evaluation*, *31*(1), 66–83. doi:10.1177/1098214009354774.
- Engel-Cox, J. A., Van Houten, B., Phelps, J., & Rose, S. W. (2008). Conceptual model of comprehensive research metrics for improved human health and environment. *Environmental Health Perspectives*, *116*(5), 583–592. doi:10.1289/ehp.10925. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/18470312>.
- Guthrie, K., Louie, J., David, T., & Foster, C. C. (2005). *The challenge of assessing policy and advocacy activities: Strategies for a prospective evaluation approach*. Los Angeles: The California Endowment.
- Haak, L., Ferriss, W., Wright, K., Pollard, M., Barden, K., Probus, M., et al. (2012). The electronic Scientific Portfolio Assistant: Integrating scientific knowledge databases to support program impact assessment. *Science and Public Policy*, *39*, 464–475. doi:10.1093/scipol/scs030.

- Hanney, S., Home, P., Frame, I., Grant, J., Green, P., & Buxton, M. (2005). Identifying the impact of diabetes research. *Diabetic Medicine*, 23, 176–184. doi:10.1111/j.1464-5491.2005.01753.x.
- Hicks, D., Tomizawa, H., Saitoh, Y., & Kobayashi, S. (2004). Bibliometric techniques in the evaluation of federally funded research in the United States. *Research Evaluation*, 13(2), 76–86. doi:10.3152/147154404781776446. Retrieved from <http://rev.oxfordjournals.org/content/13/2/7>. Accessed 19 March 2015.
- Howell, E. M., & Yemane, A. (2006). An assessment of evaluation designs: Case studies of 12 large federal evaluations. *American Journal of Evaluation*, 27(2), 219–236. doi:10.1177/1098214006287557. Retrieved from <http://aje.sagepub.com/content/27/2/219.abstract>.
- Hughes, D., Docto, L., Peters, J., Lamb, A. L., & Brindis, C. (2013). Swimming upstream: The challenges and rewards of evaluating efforts to address inequities and reduce health disparities. *Evaluation and Program Planning*, 38, 1–12. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0149718913000050>.
- IIT Research Institute. (1968). *Technology in retrospect and critical events in science: Prepared for the national science foundation* (Vol. Volume 1 & 2). Chicago: Illinois Institute of Technology Research Institute under Contract NSF-C535.
- Impact-EV. (2014). IMPACT-EV. <http://impact-ev.eu/>. Accessed 19 Aug 2015.
- Jones, T., Donovan, C., & Hanney, S. (2012). Tracing the wider impact of biomedical research: A literature search to develop a novel citation categorisation technique. *Scientometrics*, 93, 125–134. doi:10.007/s11192-012-0642-8.
- Jump, P. (2013). Australia prepares for (research) impact. *Times Higher Education*. June 22. Retrieved from <http://www.timeshighereducation.co.uk/news/australia-prepares-for-research-impact/2005011.article>.
- Kan, M. (2010). ParsCit. https://www.comp.nus.edu.sg/entrepreneurship/ParsCit_kanmy.html. Accessed 19 Aug 2015.
- Koplan, J. P., Milstein, R., & Wetterhall, S. (1999). Framework for program evaluation in public health. *MMWR: Recommendations and Reports*, 48, 1–40.
- Kostoff, R. N. (1995). Federal research impact assessment: Axioms, approaches, applications. *Scientometrics*, 34(2), 163–206.
- Kostoff, R. N. (1998). The use and misuse of citation analysis in research evaluation. *Scientometrics*, 43(1), 27–43.
- Lane, J. (2010). Let's make science metrics more scientific. *Nature*, 464(7288), 488–489. doi:10.1038/464488a.
- Lane, J., & Bertuzzi, S. (2011). Measuring the results of science Investments. *Science*, 331(6018), 678–680. doi:10.1126/science.1201865. Retrieved from <http://www.sciencemag.org/content/331/6018/678.short>.
- Lewison, G., Rippon, I., & Wooding, S. (2005). Tracking knowledge diffusion through citations. *Research Evaluation*, 14(1), 5–14. doi:10.3152/147154405781776319.
- Leydesdorff, L. (1998). Theories of citation? *Scientometrics*, 43(1), 5–25.
- Liebow, E., Phelps, J., Van Houten, B., Rose, S., Orians, C., Cohen, J., et al. (2009). Toward the assessment of scientific and public health impacts of the National Institute of Environmental Health Sciences Extramural Asthma Research Program using available data. *Environmental Health Perspectives*, 117(7), 1147–1154. doi:10.1289/ehp.0800476. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/19654926>.
- Martin, B. R. (1996). The use of multiple indicators in the assessment of basic research. *Scientometrics*, 36(3), 343–362. doi:10.1007/BF02129599. Retrieved from <http://sci2s.ugr.es/hindex/pdf/Martin1996.pdf>.
- Milat, A. J., Bauman, A. E., & Redman, S. (2015). A narrative review of research impact assessment models and methods. *Health Research Policy and Systems*, 13(18). doi:10.1186/s12961-015-0003-1. Retrieved from <http://www.health-policy-systems.com/content/pdf/s12961-015-0003-1.pdf>.
- Moed, H. F. (Ed.). (2005). *Citation analysis in research evaluation (information science and knowledge management)*. Netherlands Springer: Dordrecht.
- Narin, F. (1976). *Evaluative bibliometrics*. Cherry Hill, NJ: Computer Horizons.
- National Academy of Sciences (NAS). (2001). *Implementing the government performance and results act for research: A status report*. Washington, DC: National Academy of Sciences, National Academy of Engineering, Institute of Medicine.
- National Academy of Sciences (NAS). (2004). *NIH Extramural center programs: Criteria for initiation and evaluation*. Washington, DC: National Academy of Science.
- National Institute of Environmental Health Sciences (NIEHS). (2012). Partnerships for environmental public health evaluation metrics manual (NIH publication no. 12-7825). Durham, NC. Retrieved from http://www.niehs.nih.gov/research/supported/assets/docs/a_c/complete_peph_evaluation_metrics_manual_508.pdf. Accessed 13 March 2015.

- National Institutes of Health (NIH). (2014a). NIH RePORTER. <http://projectreporter.nih.gov/reporter.cfm>. Accessed 11 March 2015.
- National Institutes of Health (NIH). (2014b). NIH public access policy website. <http://publicaccess.nih.gov/policy.htm>. Accessed 20 March 2015.
- National Institutes of Health (NIH). (2015a). iEdison. <https://public.era.nih.gov/iedison>. Accessed 19 Aug 2015.
- National Institutes of Health (NIH). (2015b). NIH clinical trials database. <https://clinicaltrials.gov/>. Accessed 19 Aug 2015.
- National Library of Medicine (NLM). (2015). Health services/technology assessment texts <http://www.ncbi.nlm.nih.gov/books/NBK16710/>. Accessed 19 Aug 2015.
- National Research Council (NRC). (1999). *Arsenic in drinking water*. Washington, D.C.: National Research Council (NRC). doi:10.17226/6444.
- National Research Council (NRC). (2001). *Arsenic in drinking water: 2001 update*. Washington, D.C.: National Research Council (NRC). doi:10.17226/10194.
- National Research Council (NRC). (1998). Assessing the value of research in the chemical sciences: Report of a workshop. In 6. Patents and publicly funded research. Washington (DC): National Academies Press. doi:10.17226/6200.
- Orians, C., Abed, J., Drew, C., Rose, S. W., Cohen, J., & Phelps, J. (2009). Scientific and public health impacts of the NIEHS Extramural Asthma Research Program—Insights from primary data. *Research Evaluation*, 18(5), 375–385. doi:10.3152/095820209X480698. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/21921976>.
- ParaTools. (2004). ParaTools. <http://paracite.eprints.org/developers/>. Accessed 19 Aug 2015.
- Pasterkamp, G., Rotmas, J., de Kleijn, D., & Borst, C. (2007). Citation frequency: A biased measure of research impact significantly influenced by the geographical origin of research articles. *Scientometrics*, 70(1), 153–165.
- Phelan, T. J. (1999). A compendium of issues for citation analysis. *Scientometrics*, 45(1), 117–136. doi:10.1007/BF02458472.
- Quinlan, K. M., Kane, M., & Trochim, W. M. K. (2008). Evaluation of large research initiatives: Outcomes, challenges, and methodological considerations. *New Directions for Evaluation*, 2008(118), 61–72. doi:10.1002/ev.261.
- Research Excellence Framework. (2014). <http://www.ref.ac.uk/>. Accessed 19 Aug 2015.
- Researchfish. (2014). <https://www.researchfish.com/>. Accessed 19 Aug 2015.
- Reuters, T. (1994). The Thomson reuters impact factor. http://thomsonreuters.com/products_services/science/free/essays/impact_factor/. Accessed 19 March 2015.
- Roessner, D., Carr, R., Feller, I., McGeary, M., & Newman, N. (1998). *The role of NSF's support of engineering in enabling technological innovation: Phase II, final report to National Science Foundation*. Arlington, VA: SRI International.
- Sayers, E. (2009). The E-utilities in-depth: Parameters, syntax and more. In Entrez programming utilities help. Bethesda, MD: National Center for Biotechnology Information, U.S. National Library of Medicine. Retrieved from ESearch: http://www.ncbi.nlm.nih.gov/books/NBK25499/#_chapter4_ESearch_.EFetch: http://www.ncbi.nlm.nih.gov/books/NBK25499/#_chapter4_EFetch_. ECitMatch: http://www.ncbi.nlm.nih.gov/books/NBK25499/#_chapter4_ECitMatch_.
- Srivastava, C. V., Towery, N. D., & Zuckerman, B. (2007). Challenges and opportunities for research portfolio analysis, management, and evaluation. *Research Evaluation*, 16(3), 152–156. doi:10.3152/095820207x236385. Retrieved from <http://rev.oxfordjournals.org/content/16/3/152.abstract>.
- Stuart, J. (2007). Necessity leads to innovative evaluation approach and practice. *Evaluation Exchange*, XII (1), 2.
- Teles, S., & Schmitt, M. (2011). The elusive craft of evaluating advocacy. *Stanford Social Innovation Review*, 4. Retrieved from http://www.ssireview.org/images/digital_edition/2011SU_Feature_TelesSchmitt.pdf.
- The Kellogg Foundation. (2004). Logic model development guide. Battle Creek, MI. Retrieved from <http://www.wkcf.org/knowledge-center/resources/2006/02/wk-kellogg-foundation-logic-model-development-guide.aspx>. Accessed 13 March 2015.
- U.S. Environmental Protection Agency (EPA). (2009). Integrated science assessment for particulate Matter (EPA/600/R-08/139F). Research Triangle Park, NC. Retrieved from <http://cfpub.epa.gov/ncea/cfm/recorddisplay.cfm?deid=216546>. Accessed 13 March 2015.
- U.S. Environmental Protection Agency (EPA). (2010). Integrated science assessment for carbon monoxide (EPA/600/R-09/019F). Research Triangle Park, NC. Retrieved from <http://cfpub.epa.gov/ncea/cfm/recorddisplay.cfm?deid=218686>. Accessed 13 March 2015.

- U.S. Environmental Protection Agency (EPA). (2012). Integrated science assessment for lead (EPA/600/R-10/075B). Research Triangle Park, NC. Retrieved from <http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=255721>. Accessed 13 March 2015.
- U.S. Government Accountability Office. (2012). Designing evaluations (Publication No. GAO-12-208G). Washington, DC. Retrieved from <http://www.gao.gov/products/GAO-12-208G>. Accessed 11 March 2015.
- Wooding, S., Hanney, S., Buxton, M., & Grant, J. (2005). Payback arising from research funding: Evaluation of the Arthritis Research Campaign. *Rheumatology*, *44*, 1145–1156. doi:10.1093/rheumatology/keh708.
- World Health Organization (WHO). (2011). *Arsenic in drinking water*. Geneva: World Health Organization.