METHOD ARTICLE

Understanding the funding characteristics of research impact: A proof-of-concept study linking REF 2014 impact case studies with Researchfish grant agreements [version 1; peer review: 2 approved with reservations]

Gavin Reddick ¹, Dmitry Malkov^{1,2}, Beverley Sherbon ¹, Jonathan Grant^{3,4}

¹Interfolio UK, Cambridge, UK

²Science Policy Research Unit, University of Sussex Business School, Brighton, UK

³Different Angles, Cambridge, UK

⁴The Bennett Institute for Public Policy, University of Cambridge, Cambridge, UK

V1 First published: 17 Dec 2021, 10:1291
<https://doi.org/10.12688/f1000research.74374.1>
 Second version: 17 May 2022, 10:1291
<https://doi.org/10.12688/f1000research.74374.2>
 Latest published: 20 Sep 2022, 10:1291
<https://doi.org/10.12688/f1000research.74374.3>

Abstract

Background: All parts of the research community have an interest in understanding research impact whether that is around the pathways to impact, processes around impact, methods for measurement, describing impact and so on. This proof of concept study explored the relationship between research funding and research impact using the case studies submitted to the UK Research Excellence Framework (REF) exercise in 2014 as a proxy for impact.

Methods: The paper describes an approach to link the REF impact case studies with the underpinning research grants present in the Researchfish dataset, primarily using the publications captured in both datasets. Where possible the methodology utilised unique identifiers such as Digital Object Identifiers and PubMed ID's, and where this was not possible the funding information within each publication was used.

Results: Through this automated approach 21% of the non-redacted case studies could be linked to a specific research grant. Additional qualitative analysis was then done for unlinked REF impact case studies, which involved reading the document to identify additional information to make the linkage. This approach was taken on 100 REF impact case studies selected at random and resulted in only seven having no identifiable research grants funding associated. The linked research grants were analysed to identify characteristics that are more frequently associated with these grants, than non-linked ones.

Conclusions: This analysis did point to some interesting observations such as the grant funding linked to REF impact case studies are more

Open Peer Review

Approval Status

	1	2
version 3 (revision) 20 Sep 2022		 view
version 2 (revision) 17 May 2022	 view ↑	↑
version 1 17 Dec 2021	 view	 view

1. **Adam Kamenetzky** , National Institute for Health Research Central Commissioning Facility, Twickenham, UK
2. **Daniele Rotolo** , University of Sussex, Brighton, UK

Any reports and responses or comments on the article can be found at the end of the article.

likely to be longer, higher financial value, have more publications and be more collaborative (amongst other characteristics). These findings should be used with caution at present and not be over interpreted, this is due to the sample size for this proof of concept study and some potential limitations on the data which were not addressed at this stage.

Keywords

Researchfish, REF, digital-data, linkage, impact, research characteristics

This article is included in the [Research on Research, Policy & Culture](#) gateway.

Corresponding author: Beverley Sherbon (beverley.sherbon@interfolio.com)

Author roles: **Reddick G:** Conceptualization, Methodology, Supervision, Visualization, Writing – Original Draft Preparation; **Malkov D:** Data Curation, Formal Analysis; **Sherbon B:** Funding Acquisition, Project Administration, Validation, Writing – Review & Editing; **Grant J:** Conceptualization, Funding Acquisition, Methodology, Writing – Original Draft Preparation

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2021 Reddick G *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Reddick G, Malkov D, Sherbon B and Grant J. **Understanding the funding characteristics of research impact: A proof-of-concept study linking REF 2014 impact case studies with Researchfish grant agreements [version 1; peer review: 2 approved with reservations]** F1000Research 2021, **10**:1291 <https://doi.org/10.12688/f1000research.74374.1>

First published: 17 Dec 2021, **10**:1291 <https://doi.org/10.12688/f1000research.74374.1>

Introduction

The purpose of this proof-of-concept study was to explore the relationship between research funding and research impact by linking Research Excellence Framework (REF) 2014 impact cases studies (ICS) with Researchfish Grant Agreements (GA). As such it builds on a long history of studies investigating factors associated with research impact (Marjanovic *et al.*, 2009). For example, in the 1960s and 1980s there were a series of studies that examined the contributions research makes to society and what were the characteristics of that research. Some studies looked at the genesis of individual innovations (Jewkes *et al.*, 1958; Sherwin and Isenson, 1967; Illinois Institute of Technology, 1968; Comroe and Dripps, 1976; Battelle Laboratories, 1973), whilst others focused on better understanding the process through which research contributes to innovation i.e. research translation pathways and variables (Evered *et al.*, 1987; Narin, 1989; Arundel *et al.*, 1995). In the 1990s and 2000s, the theme of measuring research impact – both quantitatively through economic analysis and qualitatively through case studies – began to dominate the scholarly literature (e.g. Mansfield, 1991; Herbertz and Müller-Hill, 1995; Buxton and Hanney, 1996; Grant *et al.*, 2000; Grant and Buxton, 2018; Hanney *et al.*, 2003a, 2003b; Wooding *et al.*, 2004). By the 2010s some of these approaches began to be operationalised into national assessment through, for example, the introduction of impact into the UK's REF and to a lesser extent the Australian Engagement and Impact Assessment (Williams and Grant, 2018).

Bozeman *et al.* (1999) explained how these studies had moved through four incremental phases: 1) *historical descriptions* – tracing innovations back to their fundamental supporting inventions; 2) *'research event' based case-studies* – building a family tree of research events that led to an innovation; 3) *matched comparisons* – taking matched successful and unsuccessful innovations, tracing and comparing their development; and 4) *conventional case studies* – using action research, interviews, surveys, narrative descriptions – complemented with economic and bibliometric techniques in an attempt to increase methodological rigour and objectivity (Grant and Wooding, 2010). Today we can perhaps add a fifth phase that is associated data linkage and datamining, facilitated by access to digital data (King's College London and Digital Science, 2015; Onken *et al.*, 2020). One of the best proponents of this is a recent study by Onken *et al.* (2020) that traced the long-term impact of research funded by the National Institute of General Medical Science by linking grant data with primary publications and associated citations (over a number of generations) with patents, and drug products approved by the US Food and Drug Administration.

Building on the opportunity presented by digital data, in the proof-of-concept study reported here we examined whether it was possible to link REF 2014 ICS with Researchfish GA, and where that occurs what are the characteristics of linked GA versus non linked GAs. We were motivated to undertake the study while an eye was kept on the impending outcomes of REF2021 and the anticipated publication of a further set of *circa* 7000 case studies. As described below, the first iteration of the study resulted in relatively low levels of linkage so it was not known whether the 'unlinked' case studies were 'real' i.e. do not have underpinning research grants associated with them or were an 'artefact', either of (i) the linkage algorithm created by the authors to help the manual coding task or (ii) have associated underpinning research grants that are not indexed on Researchfish. To test this a random sample of 100 ICS were selected to see whether they could be linked to GA though more in depth quantitative and qualitative approaches, that is either through semi-automated algorithms or by hand. Based on this in-depth assessment a detailed comparison of the GAs that were linked to REF ICS vs all GAs in the Researchfish database was undertaken. This elucidated a number of interesting observations about the characteristics between research funding and research impact, although it must be stressed that these observations need to be validated and thus should be treated with caution.

Data sources

The two key data sources for this study were the REF 2014 ICS and Researchfish GA. The REF reviews the research quality of UK universities every 5-6 years. It matters not only as a signal of the reputation of an institution, but also because it determines the allocation of government block grant funding to universities, known as 'QR funding' (quality related research funding). The REF has been running in various iterations since 1986, but critically in the 2014 exercise (and the current 2021 iteration) the assessment of societal impact was included. REF is organised around four main panels (A to D) representing broad cognate disciplines (such as Arts and Humanities, Panel D) and 36 units of assessment (UOA, or sub panels) for specific disciplines (such as History, UOA 30; REF2014).

Impact was assessed through 6,975 ICS that were 4-5 page summaries of the contribution research had made on society over a 20 year period (King's College London and Digital Science, 2015). The ICS are published through the online REF2014 database, which includes an API allowing for data extraction, linkage and analysis. The database only contains ICS that were not redacted and where the submitting university had given permission for them to be published, resulting in 6637 ICS that could be analysed for the purpose of this study. One section of the ICS was the 'underpinning research' which typically contained citations to publications in the (peer reviewed) literature, which included where available digital object identifiers (DOI) which could facilitate data linkage.

Researchfish is an online platform designed to enable researchers to report the outcomes of their work across multiple funders, to re-use their data for their own purposes and to have control over who sees and accesses the data. Researchfish is essentially a data collection tool and supporting service for organisations to track research and evidence impact. Research outputs (and outcomes and impact) are gathered through a standard 'question set' initially developed by funding institutions through a consultative process, within subsequent ongoing governance from the Researchfish Question Set Subgroup which is comprised of stakeholders from funders and research organisations that use the system. This question set has 16 main outcome types e.g. publications, collaborations, IP, engagement activities and so on with each being broken down into sub-types, of which there are 103 in total. A researcher, or one of their delegates, can add, edit and delete entries, and crucially, attribute entries to research grants and awards (GAs). This collation and attribution of research outputs and outcomes serves a number of purposes. Research funders can capture a range of data that have been submitted by the researchers they fund – from publications, policy impact to products and interventions – enabling them to evaluate the impact of their research funding by various units of assessment (e.g., disciplinary focus, research funding mechanism, host institution etc). Research publications are automatically populated using web scraping technologies and the researcher or delegate confirms whether the publication is associated with the research grant. Where that automation occurs the DOI is also captured thus facilitating linking with other external datasets including potentially the REF 2014 ICS.

Methods

As illustrated in Figure 1, and described below, a four-step approach was adopted for this proof of concept study. In this paper the goal of the code was to enable manual time intensive tasks to be automated making a broader analysis of REF2014 more feasible. All linking was first attempted with code, validated and then supplemented by manual coding.

Step 1: Linking ICS with Researchfish GA

At the outset we tested whether it was possible to link REF ICS with Researchfish GAs using DOIs captured in both datasets. DOIs are persistent identifiers that remain fixed for the lifetime of a document and are widely used to identify academic, professional and government information such as journal articles and research reports. As such they occur in both REF ICS and Researchfish GAs providing a theoretical mechanism to link both datasets. However, linkage is complicated by varied and different approaches to indexing research publications. For example, in ICS researchers may use PubMed identifiers as well as both short and long forms of DOIs some may even provide no identifier at all. To take into account this variance a process was developed to clean and standardise DOIs to bibliographic information in the REF2014 ICS (Figure 2).

Step 2: Improving data linkage for a randomly selected group of 100 ICS

A significant limitation of the first step was that only 21% of the ICS could be linked to GAs. The aim of step two was to assess, using the random number generator in Excel a sample of 100 case studies, whether the 79% of 'unlinked' case studies were 'real' i.e., do not have underpinning research grants associated with them or are an 'artefact', either of (i) the linkage algorithm developed for the feasibility study (Step 1) or (ii) have associated underpinning research grants that are

Figure 1. Schematic approach of project methodology. A four-step approach was adopted for this proof of concept study to test whether it was possible to link Impact Case Studies (ICS) from the 2014 Research Excellence Framework (REF) exercise, to Researchfish Grant Agreements (GAs), and then to investigate the characteristics of the grants links to case studies compared to those that were not linked.

Figure 2. Process for cleaning and standardising digital object identifiers (DOIs) in Research Excellence Framework (REF) 2014 impact case studies (ICS). The process for matching bibliographic references from REF ICS needed to allow for variable types of persistent identifiers, namely DOIs and PubMed IDs, and then convert them all into valid DOIs for a consistent dataset to work on. This figure explains the process for cleaning, deduplicating and standardising the DOIs used for the study.

not indexed on Researchfish. This is illustrated in Panel A in Figure 3. On the horizontal axis is whether there is a Researchfish GA and on the vertical axis whether the ICS can be linked or not to the GA. The bottom left-hand box (I) indicates those 21% of ICS that could be linked to the GA in Step 1. The top left-hand box (II) are those GAs that do actually underpin an ICS but the linkage algorithm in the feasibility study failed to make the match (that is they are an

Figure 3. Conceptual overview for linking Research Excellence Framework (REF) 2014 impact case studies (ICS) with Researchfish grant agreements (GAs). This figure illustrates the next step of the process which aimed to assess the unlinked ICS (79%) taken from the Research REF 2014 dataset, and investigate whether they really did not have any underpinning research grant associated with them or are an 'artefact', either of (i) the linkage algorithm developed for the feasibility study (Step 1) or (ii) have associated underpinning research grants that are not indexed on Researchfish. The box on the left represents the full set of case studies (panel A), and the different possibilities for each, and then the box on the right (panel B) represents 100 randomly selected case studies that could not be linked in step 1, and then results of further investigation on each. Box I: Linked in feasibility study (i.e. 21%). Box II: GA underpins REF case study but not identified with DOI linkage. Box III: Funding underpins REF case study, but no GA i.e. on Researchfish. Box IV: By inference, REF case studies not underpinned by grant funding.

‘artefact’ of the approach adopted). Similarly, the bottom right-hand box (III) have associated underpinning research grants but they are not indexed on Researchfish e.g. it may be an ICS that is underpinned by National Institutes of Health funding from the US or by a funder using Researchfish but they have chosen not to track that specific grant in the system for some reason. The final box (IV) in the top right-hand corner are those inferred ICS that have no underpinning research grants (whether indexed on Researchfish or from another non-indexed research funder).

The aim of this second step was in effect to populate this 2×2 matrix with 100 randomly selected case studies that could not be linked from Step 1. This involved developing and running semi-automated algorithms to improve data matching and reading the case studies to identify additional information. Overall, four specific approaches were used. The first was enhanced DOI matching, which was effectively applying improvements to the algorithm that had occurred since it was originally applied in the feasibility study. The second approach involved extracting funding information from papers that were cited in the underpinning research on the ICS and then seeing whether that information could be matched to a GA. Typically, this involved a grant identifier in the paper and matching it with Researchfish. The third approach was using the structured funding information in the ICS and again seeing whether that could be matched to the Researchfish GA. The structured funding information included in the ICS database is limited to a small number ($n = 16$) of funders that were supported through the UK Science Budget disbursed by the Department of Business, Innovation & Skills (BIS) bodies and the Wellcome Trust (that co-funded the development of the ICS database). After this, qualitative judgement was used to compare, for example, the topic of the case study with titles and abstracts of GAs using keyword searches.

Step 3: Additional qualitative analysis for unlinked ICS

The third step was based on qualitative analysis and involved reading the ICS to identify additional information to link to GA data and/or funding and once that was exhausted following up with telephone or email interviews with the authors of the remaining ICS to see whether the underpinning research was funded or not, and if so who funded it. Each of the ICS were read by three of the authors (DM, GR and JG) who met on a weekly basis to review their findings and ensure consistency in coding. The interviews were conducted by one author (GR).

Step 4: Comparing the characteristics of linked ICS with GA with all GA

The final step involved comparing the linked ICS with GA to all GA, using a number of metrics derived from Researchfish output data. The purpose of this approach was to test whether such comparisons could be made and whether, in principle, they could provide interesting information for understanding the relationship between research funding and research impact. For this set we looked at both the original linked GA (i.e. the 21%) and those 55 ICS that we manage to link through the qualitative (Step 3) assessment.

Results

The initial scraping of bibliographic information in the ICS (Step 1), resulted in 13,708 complete DOIs being identified. Of the 13,708 DOIs, 2805 (or 20%) could be matched to equivalent DOIs maintained in the Researchfish GA data. These GA DOIs are captured by a research object (i.e. a paper) either directly reported and attributed to a specific GA by the researchers (or their delegates) or automatically harvested based on funding acknowledgements in the papers themselves, and then subsequently confirmed by the researcher. This meant that 1383 of 6637 (i.e. 21%) non redacted case studies that can be downloaded from the REF impact case study database could be linked to specific research grants using this automated approach. As illustrated in [Figures 4 and 5](#), the distribution of DOIs scrapped from ICS varied by the UOA, with greater numbers in Panels A and B than C and D, as was the number of linked GAs per ICS.

[Table 1](#) summarises the results of Step 2 of developing semi-automated linkage by focusing on the randomly selected 100 case studies. As illustrated in this table the majority (57) of the ICS could be linked to GAs through these enhanced semi-automated approaches. The enhanced DOI matching included one ICS that would have been picked up in the feasibility study (Step 1) due to an update in the data within Researchfish (the publication had been entered manually but a DOI for the publication was subsequently identified). DOIs for the remaining nine ICS were identified by extracting the bibliographic data from the case study, and using [Crossref](#), to identify likely DOIs before validating and then discovering matches to GAs. Extracting the funding data from publications cited in the ICS and matching that to the GAs resulted in a further six linkages but the most significant addition was made through the use of the structured funding information captured in the ICS database, resulting in further 41 ICS being linked to GAs.

The remaining 43 ICS were then read by three of the authors. This resulted in the identification of 34 ICS that had some form of underpinning research grant funding, but from a funder not indexed on Researchfish. For the remaining nine ICS, the authors were identified and contacted via email seeking information on any underpinning research funding and offering a response either by return via email or to arrange a telephone interview. Of the nine ICS, responses were received from six, and non-responses from three. Of the five ICS with additional information, two confirmed that they had some

Figure 4. Distribution of digital object identifier (DOIs) in impact case studies (ICS) for each for each units of assessment (UoA). The figure shows the distribution of the number of extracted and validated publication DOIs for each of the ICS within each of the Research Excellence Framework UoA as a box plot.

Figure 5. Distribution of grant agreements (GAs) per impact case studies (ICS) for each units of assessment (UoA). The figure shows the distribution of the number of GAs in Researchfish that were able to be linked to each of the ICS within each of the Research Excellence Framework UoA as a box plot.

sort of research funding and were therefore allocated to Box III, with the remaining seven to Box IV of which five we confirmed no underpinning research funding.

As illustrated in Panel B in [Figure 1](#), based on this analysis the 2×2 matrix for the 100 randomly allocated case studies could be repopulated. This resulted in the majority of ICS (55 i.e. 10 Box I and 45 Box II) being linked to Researchfish

Table 1. Results of Phase 1, semi-automated linkage.

	Panel A (Life sciences)	Panel B (Engineering and Physical Sciences)	Panel C (Social Sciences)	Panel D (Arts and Humanities)	Total
Remaining cases	25	25	25	25	100
Enhanced DOI matching	23	22	22	23	90
Publication funding extraction	21	20	21	22	84
Structured funding information	12	9	12	10	43

The table shows the number of unmatched impact case studies (ICS) drawn from each of the four panels at each stage of the process. For example, in the case of Panel A the enhanced DOI (Digital Object Identifier) matching reduced the number of unmatched ICS in the sample from 25 to 23. This was further reduced to 21 unmatched case studies after using the publication funding extraction, and then finally reduced to 12 after extracting structured funding information.

Table 2. Differences between impact case study (ICS) with an underpinning grant agreement (GA) and all GAs.

Characteristics (Unit)	Updated ICS with underpinning GA identified in scoping study (n = 25)	All GA (n = 82603, as of 31/12/2013)	Original GA that were linked to ICS in feasibility study (n = 1383)
Median length of grant agreement (years)	3.7 years	3 years	3.5 years
Median value of grant agreement (£)	£270,345	£98,217	£320,132
Median number of publications	165	1	16
Median time for a policy influence	5 years	3 years	4 years
Median time for IP registration	4 years	3 years	3 years
Median number of collaborations	1	0	1
Median value of further funding	£125,000	£130,000	£170,000

The table shows the difference in characteristics between GAs in Researchfish depending on whether they were not linked to an ICS, were linked to an ICS as part of the original match, or were linked as part of the scoping study. The characteristics presented are: Median length of grant agreement (years), Median value of grant agreement (Great British Pounds £), Median number of publications, Median time for a policy influence, Median time for IP registration, Median number of collaborations, Median value of further funding (Great British Pounds £).

GAs, and a further 38 having some form of underpinning research from funders not indexed on Researchfish. Only 7 of the 100 case studies seemed to have no identifiable external research grant funding associated with them and for three of this information was not able to be definitively confirmed.

Finally, and as illustrated in [Table 2](#), the characteristics 1383 of 6637 (i.e. 21%) non redacted ICS and the additional 55 ICS that could be were subsequently linked through the more in-depth assessment were compared to the 82,603 GA in Researchfish (as of the 31/12/2013 i.e. at a similar time the ICS were submitted). Although these indicative results should be treated with considerable caution they do through up a number of interesting observations. For example, it would seem that grant funding linked to REF impact case studies are more likely to: be longer in duration; be larger in value; have more publications; have policy influence appear sooner; have more collaborations; have higher levels of further funding; have more intellectual property.

Conclusions

The primary objective of this study was to test the feasibility and utility of linking REF ICS with Researchfish GAs to assess whether there is an opportunity to contribute to the broad literature on factors associated with research success. At its simplest, the answer is yes to both elements of this question. It proved feasible to link the two independent datasets and when linked generated interesting observations that *could* make an important contribution to the literature.

However, this conclusion should not be over interpreted as there are four of significant limitations with this proof of concept study. First, only a small proportion (21%) of the ICS could be linked using fully automated algorithms, however

the more in-depth qualitative investigation showed that proportion could be significantly increased. Reassuringly, and as noted in [Table 2](#), this increase did not alter the initial policy findings of the study i.e. there were differences between ICS that could be linked to GAs, vis-à-vis those that could not.

The second caveat is the data quality in both the ICS and the GAs, and particular the use of linkable identifiers such as DOIs. This issue may resolve itself as the proportion of publications reported within Researchfish that have DOIs has increased from *circa* 80% in 2006 to *circa* 92% in 2020. Similarly, the use of DOIs was automated in REF 2021 with case study authors having to confirm the details of underpinning research publications through third-party database when submitting ICS. This would suggest that in REF 2021 the number publications reported in ICS with DOIs will increase significantly (from around 26,000 reported by the 6,637 ICS in 2014).

The third caveat is that the Researchfish GA data is limited to those funders who use the platform (and which awards fit within the funders inclusion criteria for tracking in the Researchfish platform). Whilst this is the majority of UK funders it is notable in the list of non-indexed funders that there are a number of international funders who contributed the research that underpins ICS but the nature and characteristics of this research funding is excluded from the analysis. There is no *a priori* reason to think that their characteristics would necessarily be different to the funders indexed on Researchfish but that is an untested assumption that needs to be considered when interpreting the data from the two studies.

The final caveat is that we analysed the linkages between ICS and GA and we have yet to assess the size of those GAs, the number of GAs per case study or the nature of the GA funding beyond that presented in [Table 2](#). All of this data is potentially available and something that could be examined in detail in a larger scaled up study, either of REF 2014 ICS or those from REF 2021.

The publication of the REF2021 ICS presents an opportunity to further develop this approach. Assuming a higher rate of automated linkage, say of around 60%, between the ICS and GAs (due to better use of DOIs) the application of the semi-automated and qualitative approaches developed here could be applied across the remaining *circa* 3000 case studies at not too great a cost. Back of the envelope calculation suggests that about 100 case studies could be processed a day. This means it would be practicable to scale up work presented in this paper, with the opportunity to make a significant contribution to our understanding of the characteristics of research funding underpinning societal impact.

Data availability

Source data

The [REF 2014 data](#) is publicly available for download, and is available for reuse as described by the REF 2014 data [terms of use](#).

The publication information used was gathered from [Crossref](#) and [PubMed](#).

Attribution information was used from [Researchfish](#). This is not publicly available for reuse, but requests can be made to individual organisation listed at <https://researchfish.com/the-members/>. A large amount of the data collected via Researchfish and used in this study is publicly available for reuse via the [Gateway to Research](#) at <https://gtr.ukri.org/>.

Acknowledgements

The authors wish to thank Steven Hill (Director of Research), Lewis Dean (Head of Research Funding) and Nelly Wung (Senior Policy Advisor) from Research England for the useful and supportive discussion of early findings. The authors also wish to acknowledge that Dmitry Malkov undertook some of the preliminary work for this study during an internship with Interfolio UK as part of his Science and Technology Policy MSc, at the Science Policy Research Unit, University of Sussex Business School.

References

Arundel A, van de Pall G, Soete L: *PACE Report: Innovation Strategies of Europe's Largest Industrial Firms: Results of the PACE Survey for Information Sources, Public Research, Protection of Innovations and Government Programmes* Maastricht: Limburg University Press; 1995.

Battelle Laboratories: *Interactions of Science and Technology in the Innovative Process: Some Case Studies*. National Science Foundation Report NSF C667 Columbus, OH: Battelle Columbus Laboratories; 1973.

Bozeman B, Rogers J, Dietz JS, et al.: *The Research Value Mapping Project: Qualitative-quantitative Case Studies of Research Projects Funded by the Office of Basic Energy Sciences* Atlanta, GA: Georgia Institute of Technology; 1999.

Buxton M, Hanney S: **How Can Payback from Health Services Research Be Assessed?**. *Journal of Health Service Research and Policy*. 1996; **1**(1): 35–43.

[PubMed Abstract](#) | [Publisher Full Text](#)

Comroe JH, Dripps RD: **Scientific Basis for the Support of Biomedical Science**. *Science*. 1976; **192**(4235): 105–111.

[Publisher Full Text](#)

Evered DC, Anderson J, Griggs P, et al.: **The Correlates of Research Success**. *BMJ*. 1987; **295**(6592): 241–246.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Grant J, Buxton MJ: **Economic returns to medical research funding**. *BMJ Open*. 2018; **8**: e022131.

[PubMed Abstract](#) | [Publisher Full Text](#)

Grant J, Wooding S: *In Search of the Holy Grail: Understanding Research Success*. Cambridge: RAND Europe; 2010.

Grant J, Cottrell R, Cluzeau F, et al.: **Evaluating the 'payback' on biomedical research by characterising papers cited on clinical guidelines: An applied bibliometric study**. *BMJ*. 2000; **320**: 1107–1111.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Hanney S, Gonzalez-Block M, Buxton M, Kogan M: **The Utilisation of Health Research in Policy-making: Concepts, Examples and Methods of Assessment**. *Health Research Policy and Systems*. 2003a; **1**(2): 1–28. 44.

[Publisher Full Text](#)

Hanney S, Frame I, Grant J, et al.: *From Bench to Bedside: Tracing the Payback Forwards from Basic or Early Clinical Research – A Preliminary Exercise and Proposals for a Future Study*. HERG Research Report. Uxbridge: Health Economics Research Group, Brunel University; 2003b.

Herbert H, Müller-Hill B: **Quality and Efficiency of Basic Research in Molecular Biology: A Bibliometric Analysis of Thirteen Excellent Research Institutes**. *Research Policy*. 1995; **24**(6): 959–979.

[Publisher Full Text](#)

Illinois Institute of Technology: *Technology in Retrospect and Critical Events in Science (TRACES)*. Washington, DC: National Science Foundation; 1968.

Jewkes J, Sawers D, Stillerman R: *The Sources of Invention*. London: Macmillan; 1958.

King's College London & Digital Science: *The nature, scale and beneficiaries of research impact: An initial analysis of Research Excellence Framework (REF) 2014 impact case studies*. London: The Policy Institute, King's College London; 2015.

Mansfield E: **Academic Research and Industrial Innovation**. *Research Policy*. 1991; **20**(1): 1–12.

[Publisher Full Text](#)

Marjanovic S, Hanney S, Wooding S: *A historical reflection on research evaluation studies, their recurrent themes and challenges*. Cambridge: RAND Europe; 2009.

Narin F: **The Impact of Different Modes of Research Funding**. Evered D, Harnett S, editors. *The Evaluation of Scientific Research*. Chichester: John Wiley and Sons; 1989.

Onken J, Miklos AC, Aragon R: **Tracing Long-Term Outcomes of Basic Research Using Citation Networks**. *Frontiers in Research Metrics and Analytics*. 2020; **5**: 5.

[PubMed Abstract](#) | [Publisher Full Text](#)

Sherwin CW, Isenson RS: **Project Hindsight. A Defense Department study on the Utility of Research**. *Science*. 1967; **156**(3782): 1571–1577.

[PubMed Abstract](#) | [Publisher Full Text](#)

Williams K, Grant J: **A comparative review of how the policy and procedures to assess research impact evolved in Australia and the UK**. *Research Evaluation*. 2018; **27**(2): 93–105.

[Publisher Full Text](#)

Wooding S, Hanney S, Buxton M, et al.: *The Returns from Arthritis Research: Volume 1: Approach, Analysis and Recommendations*. Cambridge: RAND Europe; 2004.

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 02 February 2022

<https://doi.org/10.5256/f1000research.78121.r116005>

© 2022 Rotolo D. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Daniele Rotolo

SPRU (Science Policy Research Unit), University of Sussex, Brighton, UK

Many thanks for the opportunity to read this interesting proof of concept. The challenges of generating data that integrate funding and research output (including impact) remains relatively unaddressed despite the importance of such type of data source to inform policymaking. Hence, this paper provides an interesting and promising contribution in this direction. The main argument of the paper is also clear to follow. I have provided below some suggestions that I hope are helpful to strengthen some aspects of the paper.

First, depending on what could be disclosed, the paper would benefit of a more detailed description of ResearchFish GA data and of their coverage. This would allow the reader to reach a better understanding of what could explain the “missing links” (e.g. in terms of which UK and non-UK funders are not included in the data).

Second, the analysis is focussed on how many ICSs could be linked to ResearchFish GAs. However, it is unclear whether the matching was also assessed in terms of proportion of ResearchFish GAs that could be linked in the case of ICSs with multiple funding sources. How does the matching perform in these cases?

Finally, as you also argued, the results reported in Table 2 should be cautiously interpreted (I suggest adding the word “exploratory” in the caption of the table). This is particularly true since, as discussed above, an ICS could be supported by more than one funding within and/or outside ResearchFish GA data. This seems an important point to clarify in the paper.

Minor comments:

- Figure 1 would be easier to read if the four circles were aligned in sequence from left to right (similar to Figure 2).
- Panel A and B in Figure 3 could be combined.

Thanks again for the opportunity to read your paper.

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: innovation studies, emerging technologies, scientometrics, corporate science

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 06 May 2022

Beverley Sherbon

Comment 1:

Many thanks for the opportunity to read this interesting proof of concept. The challenges of generating data that integrate funding and research output (including impact) remains relatively unaddressed despite the importance of such type of data source to inform policymaking. Hence, this paper provides an interesting and promising contribution in this direction. The main argument of the paper is also clear to follow. I have provided below some suggestions that I hope are helpful to strengthen some aspects of the paper.

First, depending on what could be disclosed, the paper would benefit of a more detailed description of ResearchFish GA data and of their coverage. This would allow the reader to reach a better understanding of what could explain the “missing links” (e.g. in terms of which UK and non-UK funders are not included in the data).

Response to comment 1:

A paragraph has been added describing the coverage of Researchfish data.

Currently Researchfish has data on over 195,000 Grant Agreements, with over 80% of them

from the UK. These UK data report on 268,000 different outputs, outcomes or impact before December 31st 2013 (the cut off period for REF 2014). All the major funders in the UK (i.e. UKRI, Wellcome Trust and other medical research charities) use Researchfish and over the period 2006-2013 this accounted for between £2.5-£4.0 billion of research funding each year. It should, however be noted that Researchfish does not cover research that is funded by other means, for example block grants to universities (QR funding), direct donation from philanthropists and other self initiated research.

Comment 2:

Second, the analysis is focussed on how many ICSs could be linked to ResearchFish GAs. However, it is unclear whether the matching was also assessed in terms of proportion of ResearchFish GAs that could be linked in the case of ICSs with multiple funding sources. How does the matching perform in these cases?

Response to comment 2:

This is a really good question but one we consider to be out of scope of the current paper. We are currently working on a project that tries to estimate what proportion of grant funding leads to some form of impact (of which being linked to an ICS could be a proxy indicator). The literature on this topic is very sparse and thus we do believe it is a fruitful area to explore but, as noted, is probably beyond the remit of the current work.

Comment 3:

Finally, as you also argued, the results reported in Table 2 should be cautiously interpreted (I suggest adding the word “exploratory” in the caption of the table). This is particularly true since, as discussed above, an ICS could be supported by more than one funding within and/or outside ResearchFish GA data. This seems an important point to clarify in the paper.

Response to comment 3:

This change has been made

Comment 4:

Minor comments:

- Figure 1 would be easier to read if the four circles were aligned in sequence from left to right (similar to Figure 2).
- Panel A and B in Figure 3 could be combined.

Response to comment 4:

Both figures have been slightly altered in the next version of the paper. Another reviewer commented on our language of Panel A and Panel B in Figure 3 runs the risk of being confused with the REF panels, so we have renamed these.

Competing Interests: None

© 2022 Kamenetzky A. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Adam Kamenetzky

National Institute for Health Research Central Commissioning Facility, Twickenham, UK

The authors present a novel and timely proof-of-concept study to investigate linkages between research funding and the reporting of subsequent research impact narratives, using two large data sources - Researchfish (an online platform for reporting research outcomes to funders), and the Research Excellence Framework (a national research impact assessment exercise in the UK) 2014 impact case study database.

The introduction traces a brief history of studies of research impact and in doing so provides a helpful context to the study.

Comment #1: *[Intro, para 3]* The sentence beginning “As described below...” and the remainder of this paragraph starts to describe the study methods and (final sentence) goes on to note “interesting observations...” and study limitations. Rather than noting these here partially, I would suggest these aspects might be more clearly and fully described in later sections of the article? Instead, it may be helpful for the authors to describe here in a sentence or two their specific rationale for setting out to link research funding and research impact data (e.g. reflections on why and for whom this particular analysis might be useful).

The authors clearly describe the two data sources that form the basis for the study.

The article methods and results describe and present insights from a four-step process undertaken by the authors to link Researchfish Grant Awards (GA) with REF impact case studies (ICS), primarily using commonly-reported publications and associated funding information as the locus for further analysis. Overall, I found these sections to be detailed and well described. However I do have a couple of queries relating to the availability of source data and software used for the analysis, and would recommend that the authors consult relevant sections of the journal's data guidelines (noted with reference to specific sub-sections in comments #2 & #5 below) to ensure that the journal's open data policy is being met to the fullest extent possible.

Comment #2: *[Methods, para 1]* The authors refer both to “the code” and (subsequently) “the [linkage] algorithm”, but use these terms somewhat interchangeably and do not provide details. I would suggest that the authors a) clearly describe the software used for analysis (see data guidelines, section 1.1), and b) make available the source code/algorithm that they developed, as part of the study data (see data guidelines, section 2.1.2) or alternatively provide a rationale for why this cannot be shared as part of the data availability statement (see data guidelines, section 2.3).

Comment #3: *[Methods, figure 3]* It may be helpful for the authors to use alternative terms than “Panel A” / “Panel B” for the two boxes presented in this figure, to avoid any confusion with REF review panels. Also, both here and at other points in the article the authors refer to a “feasibility study” and “Step 1” somewhat interchangeably, when describing the initial semi-automated step

to link REF ICS and Researchfish GA via commonly-reported publications; it would be helpful to be consistent when describing these as “Step 1”, throughout. Lastly, both in the left-hand box and at various points in the main body of the article the authors refer to “[the] 21%”; it would be helpful to clarify here the absolute numbers of ICS / GA that were linked / unlinked in each quadrant (as per the right-hand box), as well as the total numbers of ICS being analysed (as I understand it, n=6,637 at step 1, and n=100 at step 2?)

Comment #4: *[Methods, Step 3]* The authors describe a process to interview (via telephone or email) original authors of the REF ICS, to ascertain whether underpinning research was funded, and who funded it. While appreciably much of this information may already be in the public domain (though not necessarily referred to explicitly in REF ICS), this process could be considered primary data collection for the purposes of the research study. Please could the authors either a) provide information around ethical considerations and approvals (which would ordinarily include informed consent processes undertaken with interviewees upfront, provision of copies of interview protocols that were used, etc.) for the purposes of this study, or b) if ethical approvals were not sought, consider this data as appropriate for scoping purposes only and not appropriate for inclusion as research data as part of the qualitative analysis undertaken in Step 3? If the latter, it may be helpful for the authors to reflect on the scalability of the method outlined in this study and the relative benefits of this additional interview step (e.g. in terms of the number of additional matches it provided) versus the additional requirements (e.g. in seeking appropriate ethical approvals, informed consent, opportunity costs for interviewers/interviewees etc.), were this analysis to be carried out across a larger sample of ICS?

Comment #5: *[Results, para 1]* The authors report a total of 1,383 research grants as successfully matched to impact case studies via automated methods. These linkages would seem novel and the underlying data of value to a range of audiences (not least the research funding organisations themselves). Similarly to comment #2, above, and in line with the journal's data guidelines, I would suggest that the authors consider publishing these data, at least to include a set of searchable GA-DOI-ICS matches? Alternatively, it would be helpful to understand which specific elements of these data are considered unsuitable or unavailable to publish, if they are not already in the public domain (e.g. as the authors note via the UKRI Gateway to Research, or via other funders' open publication of grant award data). As appreciably this may not be trivial given the number of funders whose grantees' data are held in Researchfish, the authors might give an indication of any data protection issues that could arise from any potential extension or scaling up of this method, and in seeking to publish - as would seem appropriate - the underlying matched (e.g. GA-DOI-ICS) data.

Comment #6: *[Results, Table 2 & final para]* The authors refer to a “scoping study” however I am a bit unclear as to the nature or sequencing of this in relation to the 4-step process previously described in the methods and results. Additionally, the median number of publications associated with impact case studies in this category (= “Updated ICS...”) noted as 165, would seem rather discrepant with the equivalent figure for Step 1 (= “Original GA...”) noted as 16. Perhaps both these aspects could be clarified and/or explained?

The authors conclude that the method is both feasible and useful as a means to link two independent datasets with information on the progress of research towards wider societal benefits, and I would agree that there is broad value in efforts to explore such linkages further. In particular, as outlined above, I would recommend to the authors that to the greatest degree

possible, such data are made publicly available to encourage further analysis and ensure reproducibility of results. With the REF2021 exercise mandating the use of unique publication IDs (via DOI), funder IDs (via GRID) and grant award reference numbers, this kind of linkage analysis ought to become increasingly possible using publicly-available data. The authors' efforts to show proof of concept in this regard are thus particularly timely.

I thank the authors for the opportunity to review this study and would be happy to review any revised version or findings from any extension of the method, as appropriate.

Is the rationale for developing the new method (or application) clearly explained?

Partly

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

No

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Research impact assessment, research evaluation, research on research, studies of research funding organisations, science policy research & analysis.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 06 May 2022

Beverley Sherbon

Thank you for reviewing our paper and the considered comments.

Comment #1: *[Intro, para 3]* The sentence beginning "As described below..." and the remainder of this paragraph starts to describe the study methods and (final sentence) goes on to note "interesting observations..." and study limitations. Rather than noting these here partially, I would suggest these aspects might be more clearly and fully described in later sections of the article? Instead, it may be helpful for the authors to describe here in a

sentence or two their specific rationale for setting out to link research funding and research impact data (e.g. reflections on why and for whom this particular analysis might be useful).

Response to comment 1:

We have reflected on this comment but have decided against making any changes. We think this section provides a useful route map for the reader in navigating the rest of the paper. We have, however, noted one type which we have corrected in this paragraph (changing '...study while an eye ...' to '... study with an eye ...')

Comment #2: *[Methods, para 1]* The authors refer both to "the code" and (subsequently) "the [linkage] algorithm", but use these terms somewhat interchangeably and do not provide details. I would suggest that the authors a) clearly describe the software used for analysis (see data guidelines, section 1.1), and b) make available the source code/algorithm that they developed, as part of the study data (see data guidelines, section 2.1.2) or alternatively provide a rationale for why this cannot be shared as part of the data availability statement (see data guidelines, section 2.3).

Response to comment 2:

We agree with the author that there is some confusion in the language and we have therefore made changes throughout the manuscript avoid the specific and perhaps misleading use of 'code' and 'algorithm'. In the study we were basically running iteratively different searchers terms and processes and in the text we have clarified that.

Comment #3: *[Methods, figure 3]* It may be helpful for the authors to use alternative terms than "Panel A" / "Panel B" for the two boxes presented in this figure, to avoid any confusion with REF review panels. Also, both here and at other points in the article the authors refer to a "feasibility study" and "Step 1" somewhat interchangeably, when describing the initial semi-automated step to link REF ICS and Researchfish GA via commonly-reported publications; it would be helpful to be consistent when describing these as "Step 1", throughout. Lastly, both in the left-hand box and at various points in the main body of the article the authors refer to "[the] 21%"; it would be helpful to clarify here the absolute numbers of ICS / GA that were linked / unlinked in each quadrant (as per the right-hand box), as well as the total numbers of ICS being analysed (as I understand it, n=6,637 at step 1, and n=100 at step 2?)

Response to comment 3:

We have changed 'Panel A' to Figure 3a and 'Panel B' to Figure 3b in throughout. We have also added 'n=1383' in Figure 3a for the ICS that could be linked to Researchfish GA. Note the numbers in the other boxes are unknown, hence Step 2 in the study. Finally, we have amended the text to talk solely about Step 1 (as opposed to feasibility study) as suggested.

Comment #4: *[Methods, Step 3]* The authors describe a process to interview (via telephone or email) original authors of the REF ICS, to ascertain whether underpinning research was funded, and who funded it. While appreciably much of this information may already be in the public domain (though not necessarily referred to explicitly in REF ICS), this process could be considered primary data collection for the purposes of the research study. Please could the authors either a) provide information around ethical considerations and approvals

(which would ordinarily include informed consent processes undertaken with interviewees upfront, provision of copies of interview protocols that were used, etc.) for the purposes of this study, or b) if ethical approvals were not sought, consider this data as appropriate for scoping purposes only and not appropriate for inclusion as research data as part of the qualitative analysis undertaken in Step 3? If the latter, it may be helpful for the authors to reflect on the scalability of the method outlined in this study and the relative benefits of this additional interview step (e.g. in terms of the number of additional matches it provided) versus the additional requirements (e.g. in seeking appropriate ethical approvals, informed consent, opportunity costs for interviewers/interviewees etc.), were this analysis to be carried out across a larger sample of ICS?

Response to comment 4:

Given the emergent nature of this proof-of-concept study ethics approval was not sought for the email interviews. It should be noted that the information used to approach the nine interviewees was all in the public domain. The template email that we sent interviewees is available from the authors. The main idea of this proof-of-concept study was to see whether such data linkage could occur within the need to approach individuals. The findings here, as discussed in the conclusion, suggest this may be the case with improved DOIs linkage in more contemporary Researchfish GA data and the REF 2021 case studies. In other words, it is not envisaged in a future study that interviews will be needed and thus the last point is redundant.

Comment #5: *[Results, para 1]* The authors report a total of 1,383 research grants as successfully matched to impact case studies via automated methods. These linkages would seem novel and the underlying data of value to a range of audiences (not least the research funding organisations themselves). Similarly to comment #2, above, and in line with the journal's data guidelines, I would suggest that the authors consider publishing these data, at least to include a set of searchable GA-DOI-ICS matches? Alternatively, it would be helpful to understand which specific elements of these data are considered unsuitable or unavailable to publish, if they are not already in the public domain (e.g. as the authors note via the UKRI Gateway to Research, or via other funders' open publication of grant award data). As appreciably this may not be trivial given the number of funders whose grantees' data are held in Researchfish, the authors might give an indication of any data protection issues that could arise from any potential extension or scaling up of this method, and in seeking to publish - as would seem appropriate - the underlying matched (e.g. GA-DOI-ICS) data.

Response to comment 5:

We have added some data on Researchfish, but for the reasons acknowledged have developed this point.

Comment #6: *[Results, Table 2 & final para]* The authors refer to a "scoping study" however I am a bit unclear as to the nature or sequencing of this in relation to the 4-step process previously described in the methods and results. Additionally, the median number of publications associated with impact case studies in this category (= "Updated ICS...") noted as 165, would seem rather discrepant with the equivalent figure for Step 1 (= "Original GA...") noted as 16. Perhaps both these aspects could be clarified and/or explained?

Response to comment 6:

We have updated Table 2 to refer to 'Steps 2-4' instead of 'scoping study' and corrected the n to 55 (an error in the original draft). The text already notes the caution needed in interpreting the table but we have amplified that with an additional sentence.

Comment #7:

In particular, as outlined above, I would recommend to the authors that to the greatest degree possible, such data are made publicly available to encourage further analysis and ensure reproducibility of results.

Response to comment 7:

The authors of the paper fully agree with the reviewer, and data is available wherever possible in line with the necessary terms of collection and data regulations.

Competing Interests: None

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research