

A peer-reviewed version of this preprint was published in PeerJ on 1 October 2013.

[View the peer-reviewed version](https://peerj.com/articles/175) (peerj.com/articles/175), which is the preferred citable publication unless you specifically need to cite this preprint.

Piwowar HA, Vision TJ. 2013. Data reuse and the open data citation advantage. PeerJ 1:e175 <https://doi.org/10.7717/peerj.175>

Data reuse and the open data citation advantage

Heather Piwowar, Todd J Vision

BACKGROUND: Attribution to the original contributor upon reuse of published data is important both as a reward for data creators and to document the provenance of research findings. Previous studies have found that papers with publicly available datasets receive a higher number of citations than similar studies without available data. However, few previous analyses have had the statistical power to control for the many variables known to predict citation rate, which has led to uncertain estimates of the “citation boost”. Furthermore, little is known about patterns in data reuse over time and across datasets.

METHOD AND RESULTS: Here, we look at citation rates while controlling for many known citation predictors, and investigate the variability of data reuse. In a multivariate regression on 10,555 studies that created gene expression microarray data, we found that studies that made data available in a public repository received 9% (95% confidence interval: 5% to 13%) more citations than similar studies for which the data was not made available. Date of publication, journal impact factor, open access status, number of authors, first and last author publication history, corresponding author country, institution citation history, and study topic were included as covariates. The citation boost varied with date of dataset deposition: a citation boost was most clear for papers published in 2004 and 2005, at about 30%. Authors published most papers using their own datasets within two years of their first publication on the dataset, whereas data reuse papers published by third-party investigators continued to accumulate for at least six years. To study patterns of data reuse directly, we compiled 9,724 instances of third party data reuse via mention of GEO or ArrayExpress accession numbers in the full text of papers. The level of third-party data use was high: for 100 datasets deposited in year 0, we estimated that 40 papers in PubMed reused a dataset by year 2, 100 by year 4, and more than 150 data reuse papers had been published by year 5. Data reuse was distributed across a broad base of datasets: a very conservative estimate found that 20% of the datasets deposited between 2003 and 2007 had been reused at least once by third parties.

CONCLUSION: After accounting for other factors affecting citation rate, we find a robust citation benefit from open data, although a smaller one than previously reported. We conclude there is a direct effect of third-party data reuse that persists for years beyond the time when researchers have published most of the papers reusing their own data. Other factors that may also contribute to the citation boost are considered. We further conclude that, at least for gene expression microarray data, a substantial fraction of archived datasets are reused, and that the intensity of dataset reuse has been steadily increasing since 2003.

1 Introduction

2 "Sharing information facilitates science. Publicly sharing detailed research data—sample
3 attributes, clinical factors, patient outcomes, DNA sequences, raw mRNA microarray
4 measurements—with other researchers allows these valuable resources to contribute far beyond
5 their original analysis. In addition to being used to confirm original results, raw data can be used
6 to explore related or new hypotheses, particularly when combined with other publicly available
7 data sets. Real data is indispensable when investigating and developing study methods, analysis
8 techniques, and software implementations. The larger scientific community also benefits: sharing
9 data encourages multiple perspectives, helps to identify errors, discourages fraud, is useful for
10 training new researchers, and increases efficient use of funding and patient population resources
11 by avoiding duplicate data collection." ([Piwowar et. al. 2007](#))

12 Making research data publicly available also has costs. Some of these costs are borne by society:
13 For example, data archives must be created and maintained. Many costs, however, are borne by
14 the data-collecting investigators: Data must be documented, formatted, and uploaded.
15 Investigators may be afraid that other researchers will find errors in their results, or "scoop"
16 additional analyses they have planned for the future.

17 Personal incentives are important to balance these personal costs. Scientists report that receiving
18 additional citations is an important motivator for publicly archiving their data ([Tenopir et. al.
19 2011](#)).

20 There is evidence that studies that make their data available do indeed receive more citations than
21 similar studies that do not ([Gleditsch & Strand, 2003](#) ; [Piwowar et. al. 2007](#) ; [Ioannidis et. al. 2009](#) ;
22 [Pienta et. al. 2010](#) ; [Henneken & Accomazzi, 2011](#) ; [Sears, 2011](#) ; [Dorch, 2012](#)). These findings have
23 been referenced by new policies that encourage and require data archiving (e.g. ([Rausher et. al.
24 2010](#))), demonstrating the appetite for evidence of personal benefit.

25 In order for journals, institutions and funders to craft good data archiving policy, it is important to
26 have an accurate estimate of the citation differential. Estimating an accurate citation differential is
27 made difficult by the many confounding factors that influence citation rate. In past studies, it has
28 seldom been possible to adequately control these confounders statistically, much less
29 experimentally. Here, we perform a large multivariate analysis of the citation differential for
30 studies in which gene expression microarray data either was or was not made available in a public
31 repository.

32 We seek to improve on prior work in two key ways. First, the sample size of this analysis is large
33 – over two orders of magnitude larger than the first citation study of gene expression microarray
34 data ([Piwowar et. al. 2007](#)), which gives us the statistical power to account for a larger number of
35 cofactors in the analyses. The resulting estimates thus isolate the association between data
36 availability and citation rate with more accuracy. Second, this report goes beyond citation
37 analysis to include analysis of data reuse attribution directly. We explore how data reuse patterns
38 change over both the lifespan of a data repository and the lifespan of a dataset, as well as looking
39 at the distribution of reuse across datasets in a repository.

40

41

1 **Materials and Methods**

2 The main analysis in this paper examines the citation count of a gene expression microarray
3 experiment relative to availability of the experiment's data, accounting for a large number of
4 potential confounders.

5 **Relationship between data availability and citation**

6 *Data collection*

7 To begin, we needed to identify a sample of studies that had generated gene expression
8 microarray data in their experimental methods. We used a sample that had been collected
9 previously ([Piwowar, 2011d](#) ; [Piwowar, 2011c](#)); briefly, a full-text query uncovered papers that
10 described wet-lab methods related to gene expression microarray data collection. The full-text
11 query had been characterized as having high precision (90%, with a 95% CI of 86%
12 to 93%) and moderate recall (56%, CI of 52% to 61%) for this task. Running the query in
13 PubMed Central, HighWire Press, and Google Scholar identified 11603 distinct gene expression
14 microarray papers published between 2000 and 2009.

15 Citation counts for 10,555 of these papers were found in Scopus and exported in November 2011.
16 Although Scopus now has an API that would facilitate easy programmatic access to citation
17 counts, at the time of data collection the authors were not aware of any way to query and export
18 data other than through the Scopus website. The Scopus website had a limit to the length of query
19 and the number of citations that could be exported at once. To work within these restrictions we
20 concatenated 500 PubMed IDs at a time into 22 queries, each of the form "**PMID(1234) OR**
21 **PMID(5678) OR ...**".

22 The independent variable of interest was the availability of gene expression microarray data. Data
23 availability had been previously determined for our sample articles in ([Piwowar, 2011d](#)), so we
24 directly reused that dataset. Datasets were considered to be publicly available if they were
25 discoverable in either of the two most widely-used gene expression microarray repositories:
26 NCBI's Gene Expression Omnibus (GEO), and EBI's ArrayExpress. GEO was queried for links to
27 the PubMed identifiers in the analysis sample using "**pubmed_gds [filter]**" and ArrayExpress
28 was queried by searching for each PubMed identifier in a downloaded copy of the ArrayExpress
29 database. An evaluation of this method found that querying GEO and ArrayExpress with PubMed
30 article identifiers recovered 77% of the associated publicly available datasets ([Piwowar &](#)
31 [Chapman, 2010](#)).

32 *Primary analysis*

33 The core of our analysis is a set of multivariate linear regressions to evaluate the association
34 between the public availability of a study's microarray data and the number of citations received
35 by the study. To explore what variables to include in these regressions, we first looked at
36 correlations between the number of citations and a set of candidate variables, using Pearson
37 correlations for numeric variables and polyserial correlations for binary and categorical variables.
38 We also calculated correlations amongst all variables to investigate collinearity.

39 Citation counts for 10555 papers were exported from Scopus in November 2011.

1 We used a subset of the 124 attributes from (Piwowar, 2011d) previously shown or suspected to
2 correlate with citation rate (Table 1). The main analysis was run across all papers in the sample
3 with covariates found to have significant pairwise correlation with citation rate. These included:
4 the date of publication, the journal which published the study, the journal impact factor, the
5 journal citation half-life, the number of articles published by the journal, the journal's open access
6 policy, whether the journal is considered a core clinical journal by MEDLINE, the number of
7 authors of the study, the country of the corresponding author, the citation score of the institution
8 of the corresponding author, the publishing experience of the first and last author, and the subject
9 of the study itself.

10 Publishing experience was characterized by the number of years since the author's first paper in
11 PubMed, the number of papers the author has published, and the number of citations the author
12 has received from PubMed Central, estimated using Author-ity Clusters. The topic of the study
13 was characterized by whether the MeSH terms classified it as related to cancer, animals, or plants.
14 For more information on study attributes see (Piwowar, 2011d). Citation count was log transformed
15 to be consistent with prior literature. Other count variables were square-root transformed.
16 Continuous variables were represented with 3-part spines in the regression, using the rcs function
17 in the R rms library.

18 The independent variable of data availability was represented as 0 or 1 in the regression,
19 describing whether or not associated data had been found in either of the two data repositories.
20 Because citation counts were log transformed, the relationship of data availability to citation
21 count was described with 95% confidence intervals after raising the regression coefficient to the
22 power of e.

23 *Comparison to 2007 study*

24 We ran two modified analyses to attempt to reproduce the findings of (Piwowar *et. al.* 2007) with
25 the larger dataset of the current study. First, we used a subset of studies with roughly the same
26 inclusion criteria as the earlier paper -- studies on cancer, with humans, published prior to 2003 --
27 and the same regression coefficients: publication date, impact factor, and whether the
28 corresponding author's address is in the USA. We followed that with a second regression that
29 included several additional important covariates: number of authors and number of previous
30 citations by the last author.

31 *Stratification by year*

32 Because publication date is such a strong correlate with both citation rate and data availability, we
33 performed a separate analysis stratifying the sample by publication year, in addition to including
34 publication date as a covariate. Fewer covariates could be included in these yearly regressions
35 because included fewer datapoints than the full regression. The yearly regressions included date
36 of publication, the journal which published the study, the journal impact factor, the journal's open
37 access policy, the number of authors of the study, the citation score of the institution of the
38 corresponding author, the previous number of PubMed Central citations received by the first and
39 last author, whether the study was on the topic of cancer, and whether it used animals.

40 *Manual review of citation context*

41 We manually reviewed the context of citations to data collection papers to estimate how many
42 citations to data collection papers were made in the context of data reuse. We (Jonathan Carlton,

1 in acknowledgements) reviewed 50 citations chosen randomly from the set of all citations to 100
2 data collection papers. Specifically, we randomly selected 100 datasets deposited in GEO in
3 2005. For each dataset, we located the data collection article within ISI Web of Science based on
4 its title and authors, and exported the list of all articles that cited this data collection article. From
5 this, we selected 50 random citations stratified by the total number of times the data collection
6 article had been cited. By manual review of the relevant full-text of each paper, we determined if
7 the data from the associated dataset had been reused within the study.

8 **Data reuse patterns from accession number attribution**

9 A second, independent dataset was collected to correlate with reuse attributions made through
10 mentions of accession numbers, rather than formal citations.

11 ***Data collection***

12 Datasets are sometimes attributed directly through mention of the dataset identifier (or accession
13 number) in the full-text, in which case the reuse may not contribute to the citation count of the
14 original paper. To capture these instances of reuse, we collected a separate dataset to study reuse
15 patterns based on direct data attribution. We used the NCBI eUtils library and custom Python
16 code to obtain a list of all datasets deposited into the Gene Expression Omnibus data repository,
17 then searched PubMed Central for each of these dataset identifiers (using queries of the form
18 "'GSEnnnn' OR 'GSE nnnn'"). For each hit we recorded the PubMed Central ID of the paper that
19 mentioned the accession number, the year of paper publication, and the author surnames. We also
20 recorded the dataset accession number, the year of dataset publication, and the investigator names
21 associated with the dataset record.

22 ***Statistical analysis***

23 To focus on data reuse by third party investigators (rather than authors attributing datasets they
24 had collected themselves), we excluded papers with author surnames in common with those
25 authors who deposited the original dataset, as in (Piwowar *et. al.* 2011a). PubMed Central contains
26 only a subset of papers recorded in PubMed. As described in (Piwowar *et. al.* 2011a), to extrapolate
27 from the number of data reuses in PubMed Central to all possible data reuses in PubMed, we
28 divided the yearly number of hits by the ratio of papers in PMC to papers in PubMed for this
29 domain (domain was measured as the number of articles indexed with the MeSH term “gene
30 expression profiling”).

31 We retained papers published between 2001 and 2010 as reuse candidates. We excluded 2011
32 because it had a dramatically lower proportion of papers in PubMed Central at the time of our
33 data collection: the NIH requirement to deposit a paper into PMC permits a 12 month embargo.

34 To understand our findings on a per-dataset basis, we stratified reuse estimates by year of dataset
35 submission and normalized our reuse findings by the number of datasets deposited that year.

36 **Data and script availability**

37 Statistical analyses were last run on Wed Apr 3 03:54:52 2013 with R version 2.15.1 (2012-06-22).
38 Packages used included reshape2 (Wickham, 2007), plyr (Wickham, 2011), rms (Harrell, 2012),
39 polycor (Fox, 2010), ascii (Hajage, 2011), ggplot2 (Wickham, 2009), gplots (Bolker *et. al.* 2012), knitr
40 (Xie, 2012), and knitcitations (Boettiger, 2013). P-values were two-tailed.

1 Raw data and statistical scripts are available in the Dryad data repository at [data uploaded to
2 Dryad at the time of article acceptance; citation will be included once known]. Data collection
3 scripts are on GitHub at <https://github.com/hpiwowar/georeuse> and
4 <https://github.com/hpiwowar/pypub>.

5 The Markdown version of this manuscript with interleaved statistical scripts (Xie, 2012) is on
6 GitHub <https://github.com/hpiwowar/citation11k>. Publication references are available in a
7 publicly-available [Mendeley group](#) to facilitate exploration.

8 **Results**

9 **Description of cohort**

10 We identified 10557 articles published between 2001 and 2009 as collecting gene expression
11 microarray data.

12 The papers were published in 667 journals, with the top 12 journals accounting for 30% of the
13 papers (Table 1).

14	Cancer Res	0.04	
15	Proc Natl Acad Sci U S A	0.04	
16	J Biol Chem	0.04	
17	BMC Genomics	0.03	
18	Physiol Genomics	0.03	
19	PLoS One	0.02	
20	J Bacteriol	0.02	
21	J Immunol	0.02	
22	Blood	0.02	
23	Clin Cancer Res	0.02	
24	Plant Physiol	0.02	
25	Mol Cell Biol	0.01	

26 *Table 1: Proportion of sample published in most common journals*

27 Microarray papers were published more frequently in later years: 2% of articles in our sample
28 were published in 2001, compared to 15 % in 2009 (Table 2).

29		2001	2002	2003	2004	2005	2006	2007	2008	2009	
30	---	-----	-----	-----	-----	-----	-----	-----	-----	-----	
31		0.02	0.05	0.08	0.11	0.13	0.12	0.17	0.18	0.15	

32 *Table 2: Proportion of sample published each year*

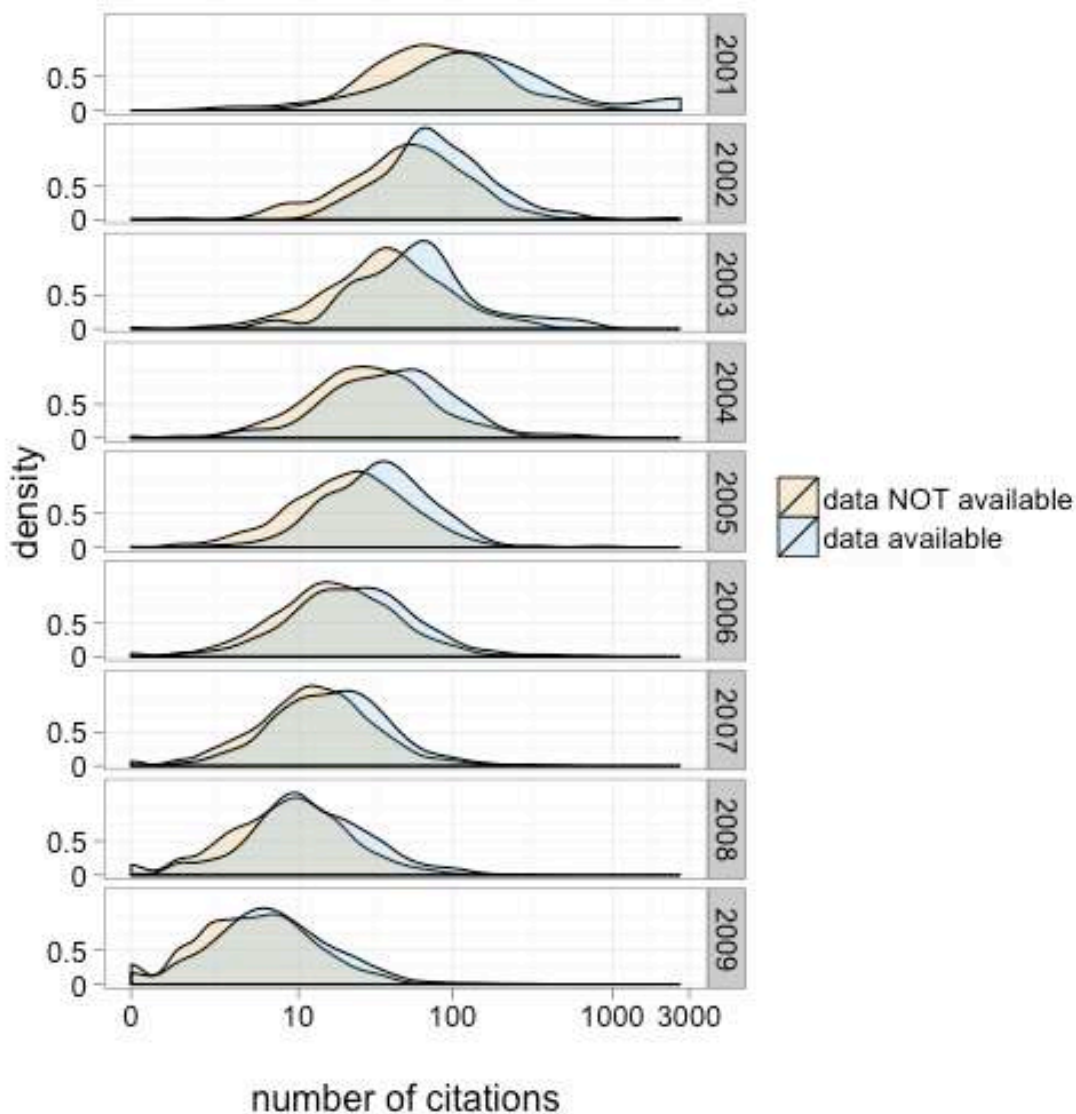
33 The papers were cited between 0 and 2643 times, with an average of 32 citations per paper and a
34 median of 16 citations.

35 The GEO and ArrayExpress repositories had links to associated datasets for 24.8% of these
36 papers.

37

1 **Data availability is associated with citation boost**

2 Without accounting for any confounding factors, the distribution of citations was similar for
3 papers with and without archived data. That said, we hasten to mention several strong
4 confounding factors. For example, the number of citations a paper has received is strongly
5 correlated to the date it was published: older papers have had more time to accumulate citations.
6 Furthermore, the probability of data archiving is also correlated with the age of an article -- more
7 recent articles are more likely to archive data (Piwowar, 2011d). Accounting for publication date,
8 the distribution of citations for papers with available data is right-shifted relative to the
9 distribution for those without, as seen in Figure 1.



10

11 *Figure 1: Citation density for papers with and without publicly available microarray data, by*
12 *year of study publication*

1 Other variables have been shown to correlate with citation rate (Fu & Aliferis, 2008). Because
2 single-variable correlations can be misleading, we performed multivariate regression to isolate the
3 relationship between data availability and citation rate from confounders.

4 The multivariate regression included attributes to represent an article's journal, journal impact
5 factor, date of publication, number of authors, number of previous citations of the first and last
6 author, number of previous publications of the last author, whether the paper was about animals
7 or plants, and whether the data was made publicly available. Citations were 9% higher for papers
8 with available data, independent of other variables ($p < 0.01$, 95% confidence intervals [5% to
9 13%]).

10 We also performed an analysis on a subset of manually curated articles. The findings were similar
11 to those of the whole sample, supporting our assumption that errors in automated inclusion
12 criteria determination did not have substantial influence on the estimate (see Supplementary
13 Article S1).

14 **More covariates led to a more conservative estimate**

15 Our estimate of citation boost, 9% as per the multivariate regression, is notably smaller than the
16 69% (95% confidence intervals of 18 to 143%) citation advantage found by (Piwowar *et. al.* 2007),
17 even though both studies looked at publicly available gene expression microarray data. There are
18 several possible reasons for this difference.

19 First, (Piwowar *et. al.* 2007) concentrated on datasets from high-impact studies: human cancer
20 microarray trials published in the early years of microarray analysis (between 1999 and 2003). By
21 contrast, the current study included gene expression microarray data studies on any subject
22 published between 2001 and 2009. Second, because the (Piwowar *et. al.* 2007) sample was small
23 (85 papers), the previous analysis included only a few covariates: publication date, journal impact
24 factor, and country of the corresponding author.

25 We attempted to reproduce the (Piwowar *et. al.* 2007) methods with the current sample. Limiting
26 the inclusion criteria to studies with MeSH terms "human" and "cancer", and to papers published
27 between 2001 and 2003, reduced the cohort to 308 papers. Running this subsample with the
28 covariates used in the (Piwowar *et. al.* 2007) paper resulted in a comparable estimate to the 2007
29 paper: a citation increase of 47% (95% confidence intervals of 6% to 103%).

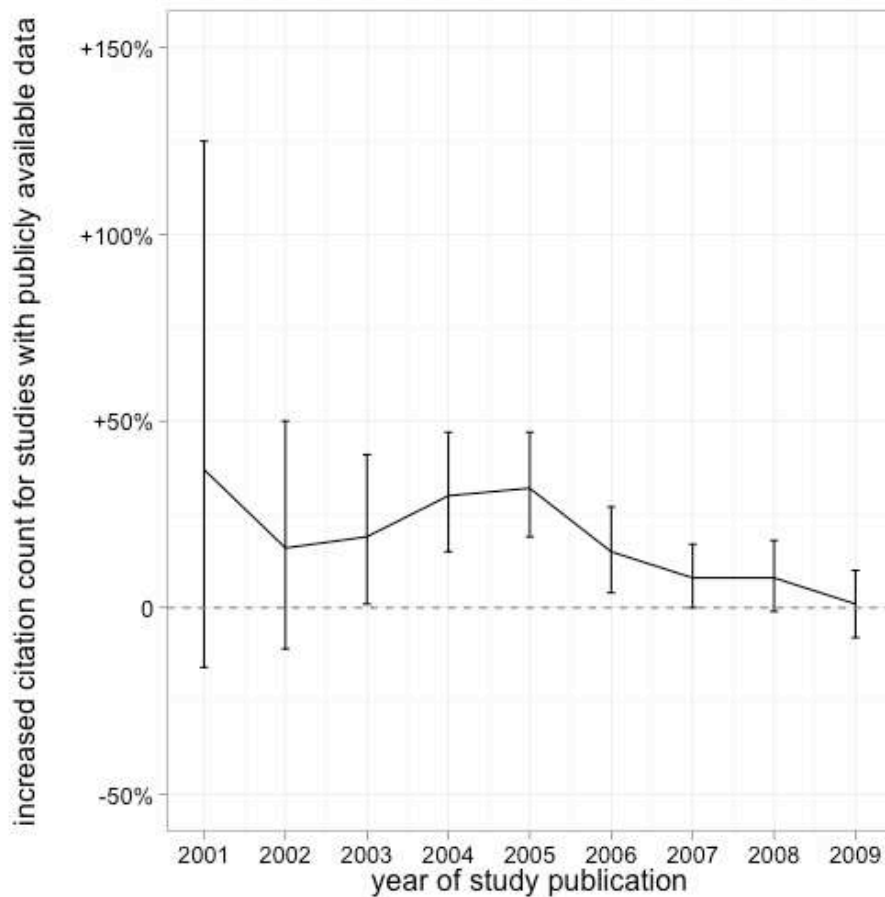
30 The subsample of 308 papers was large enough to include a few additional covariates: number of
31 authors and citation history of the last author. Including these important covariates decreased the
32 estimated effect to 18% with a confidence interval that spanned a loss of 17% citations to a boost
33 of 66%.

34 **Citation boost over time**

35 After completing our comparison to prior results, we returned to the whole sample. Because
36 publication date is such as strong correlate with both citation rate and data availability, we ran
37 regressions for each publication year individually. The estimate of citation boost varied by year of
38 publication. The citation boost was greatest for data published in 2004 and 2005, at about 30%.
39 Earlier years showed citation boosts with wider confidence intervals due to relatively small
40 sample sizes, while more recently published data showed a less pronounced citation boost (Table
41 3, Figure 2).

year	estimate	[95% confidence interval]
2001	1.37	[0.84, 2.25]
2002	1.16	[0.89, 1.5]
2003	1.19	[1.01, 1.41]
2004	1.3	[1.15, 1.47]
2005	1.32	[1.19, 1.47]
2006	1.15	[1.04, 1.27]
2007	1.08	[1, 1.17]
2008	1.08	[0.99, 1.18]
2009	1.01	[0.92, 1.1]

Table 3: Estimated citation boost multiplier for studies with publicly available data, by year of study publication



13

Figure 2: Increased citation count for studies with publicly available data, by year of publication. Estimates from multivariate analysis, lines indicate 95% confidence intervals.

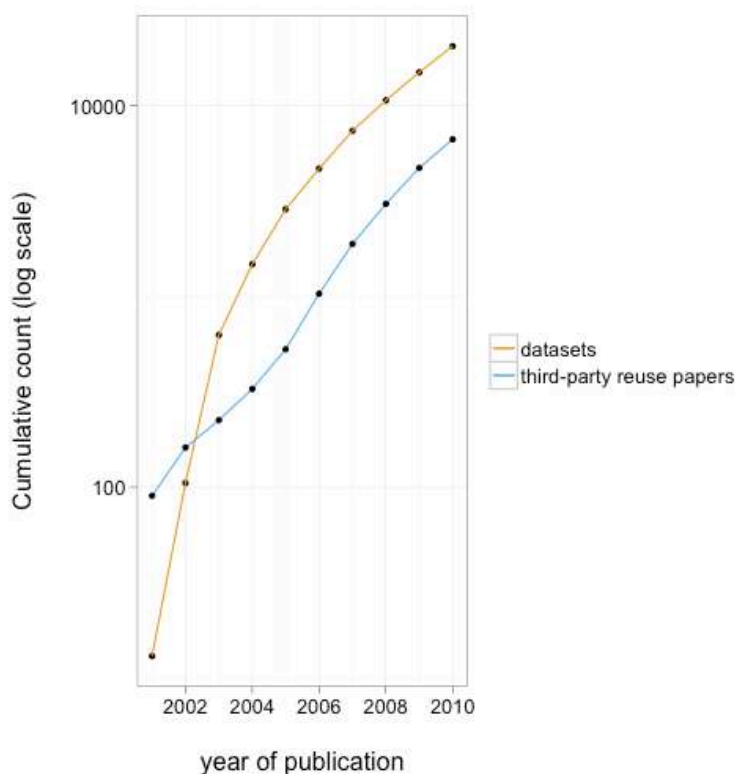
16 Data reuse is a demonstrable component of citation boost

17 To estimate the proportion of the citation boost directly attributable to data reuse, we randomly
 18 selected and manually reviewed 138 citations. We classified 8 (6%) of the citations as attributions
 19 for data reuse (95% CI: 3% to 11%).

1 **Evidence of reuse from mention of dataset identifiers in full text**

2 A complementary dataset was collected and analyzed to characterize data reuse: direct mention of
 3 dataset accession numbers in the full text of papers. In total there were 9274 mentions of GEO
 4 datasets in papers published between 2000 and 2010 within PubMed Central across 4543 papers
 5 written by author teams whose last names did not overlap those who deposited the data.
 6 Extrapolating this to all of PubMed, we estimate there may be about 1.4081×10^4 third-party
 7 reuses of GEO data attributed through accession numbers in all of PubMed for papers published
 8 between 2000 and 2010.

9 The number of reuse papers started to grow rapidly several years after data archiving rate started
 10 two grow. In recent years both the number of datasets and the number of reuse papers have been
 11 growing rapidly, at about the same rate, as seen in Figure 3.

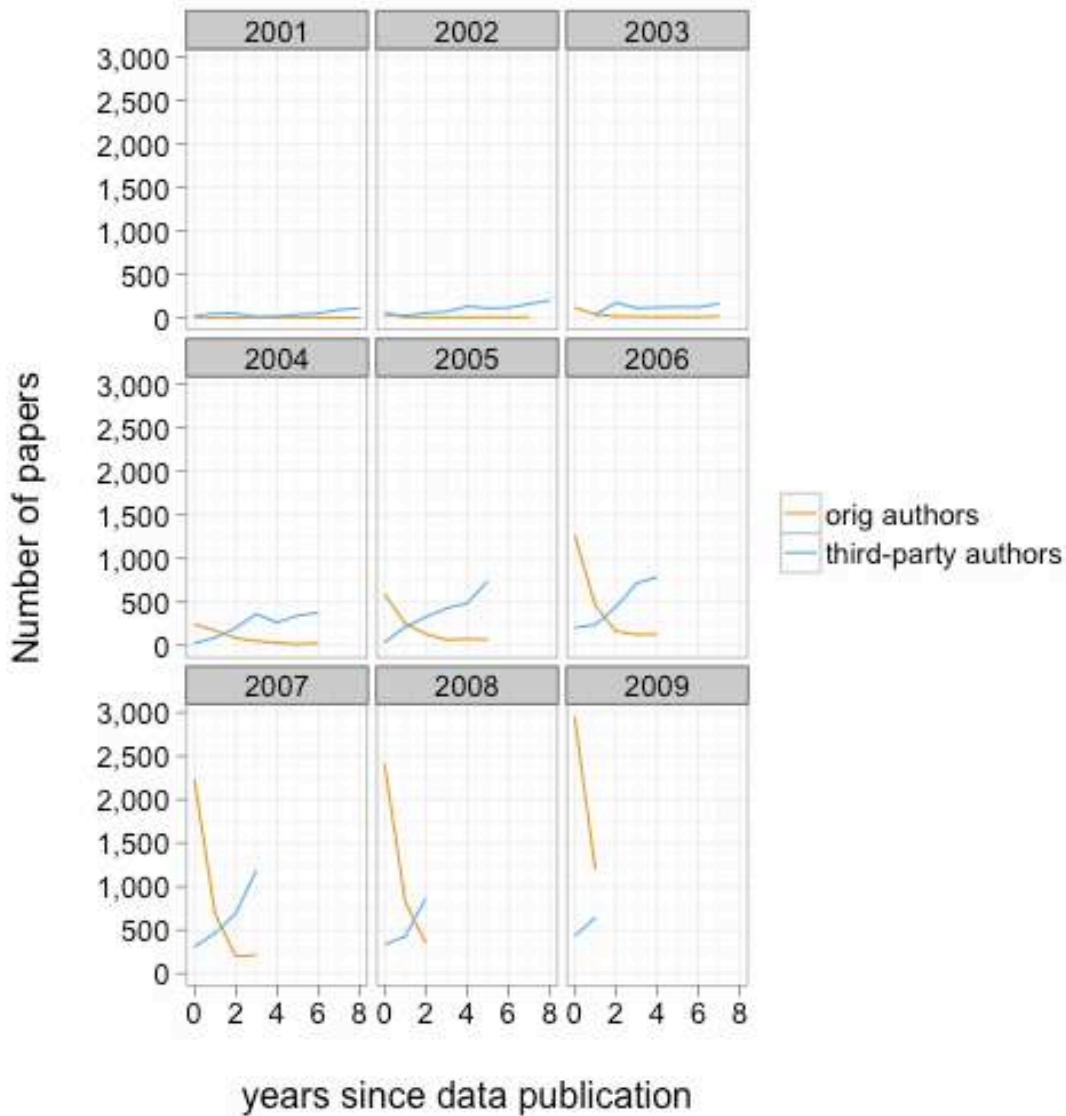


12

13 *Figure 3: Cumulative number of datasets deposited in GEO each year, and cumulative number of*
 14 *third-party reuse papers published that directly attribute GEO data published each year, log*
 15 *scale.*

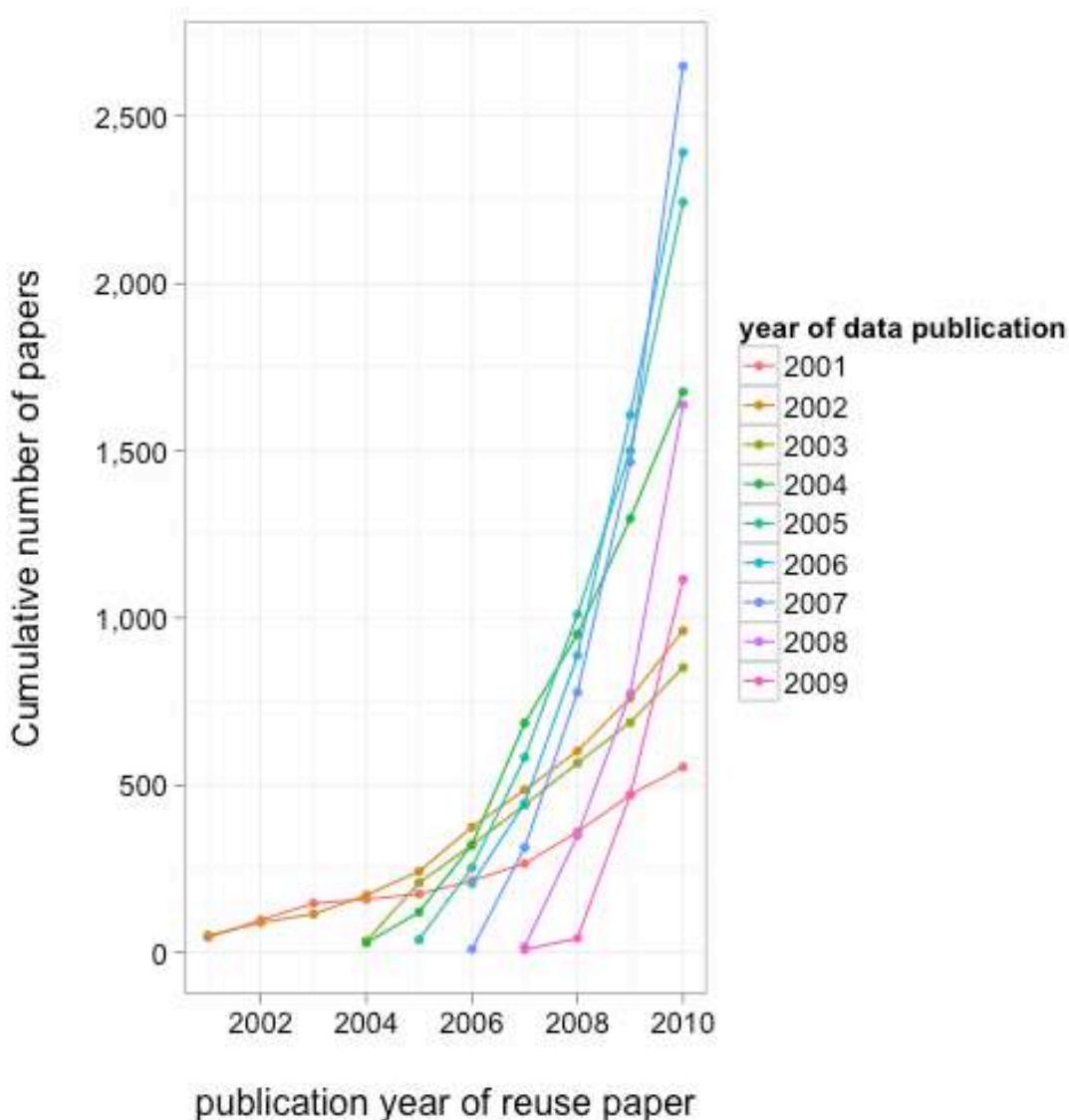
16 The level of third-party data use was high: for 100 datasets deposited in year 0, we estimate that
 17 40 papers in PubMed reused a dataset by year 2, 100 by year 4, and more than 150 by year 5. This
 18 data reuse curve is remarkably constant for data deposited between 2004 and 2009. The reuse
 19 growth trend for data deposited in 2003 has been slower, perhaps because 2003 data is not as
 20 ground-breaking as earlier data, and is likely less standards-compliant and technically relevant
 21 than later data.

1 We found that almost all instances of self reuse (identified by surname overlap with data
 2 submission record) were published within two years of dataset publication. This pattern contrasts
 3 sharply with third party data reuse, as seen in Figure 4.



4
 5 *Figure 4: Number of papers mentioning GEO accession numbers. Each panel represents reuse of*
 6 *a particular year of dataset submissions, with number of mentions on the y axis, years since the*
 7 *initial publication on the x axis, and a line for reuses by the data collection team and a line for*
 8 *third-party investigators*

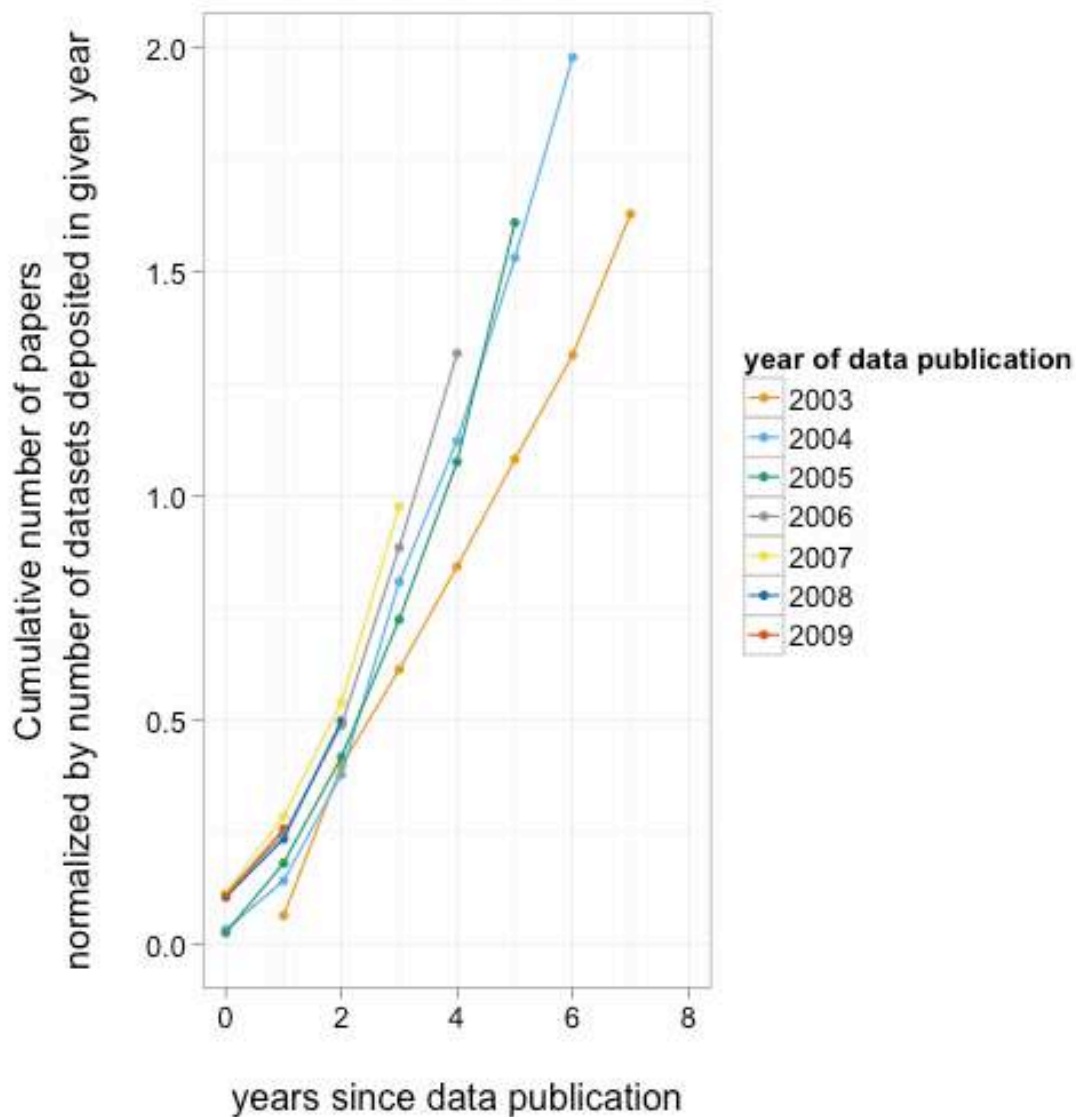
9 The cumulative number of third-party reuse papers is illustrated in Figure 5. Separate lines are
 10 displayed for different dataset publication years.



1

2 *Figure 5: Cumulative number of third-party reuse papers, by date of reuse paper publication.*
 3 *Separate lines are displayed for different dataset submission years*

4 Because the number of datasets published has grown dramatically with time, it is instructive to
 5 consider the cumulative number of third-party reuses normalized by the number of datasets
 6 deposited each year (Figure 6). In the earliest years for which data is available, 2001-2002, there
 7 were relatively few data deposits, but these datasets have been disproportionately reused. We
 8 exclude the early years from the plot to examine the pattern of data reuse once gene expression
 9 datasets became more common. Since 2003, the rate at which individual datasets are reused has
 10 increased with each year of data publication.

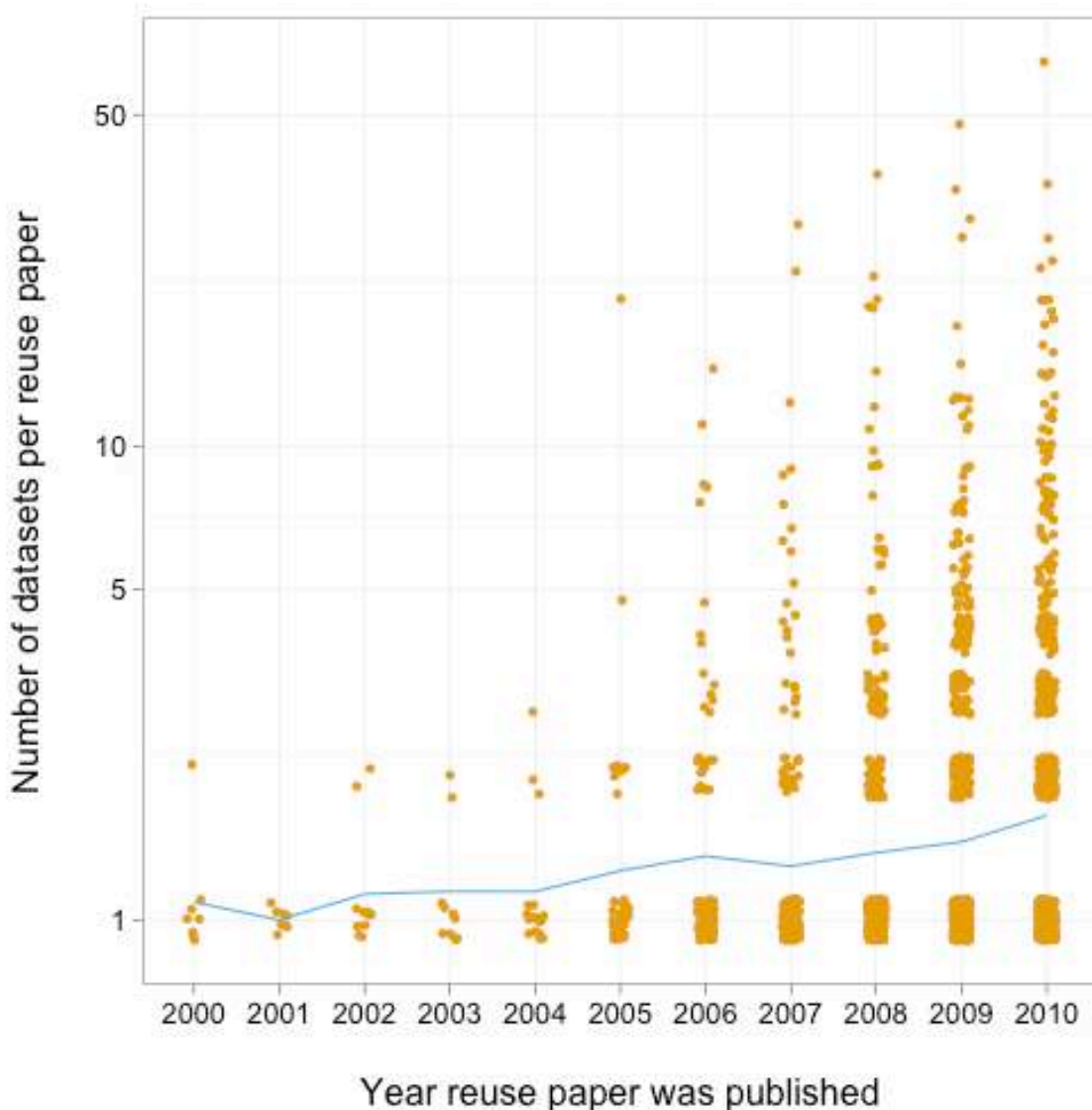


1

2 *Figure 6: Cumulative third-party reuse, normalized by number of datasets deposited each year,*
 3 *plotted as elapsed years since data publication*

4 **Growth in the number of datasets in each reuse paper over time**

5 The number of distinct datasets used in a reuse paper was found to increase over time (Figure 7).
 6 In 2002-2004 almost all reuse papers only used one or two datasets. By 2010, 25% of reuse
 7 papers used 3 or more datasets.

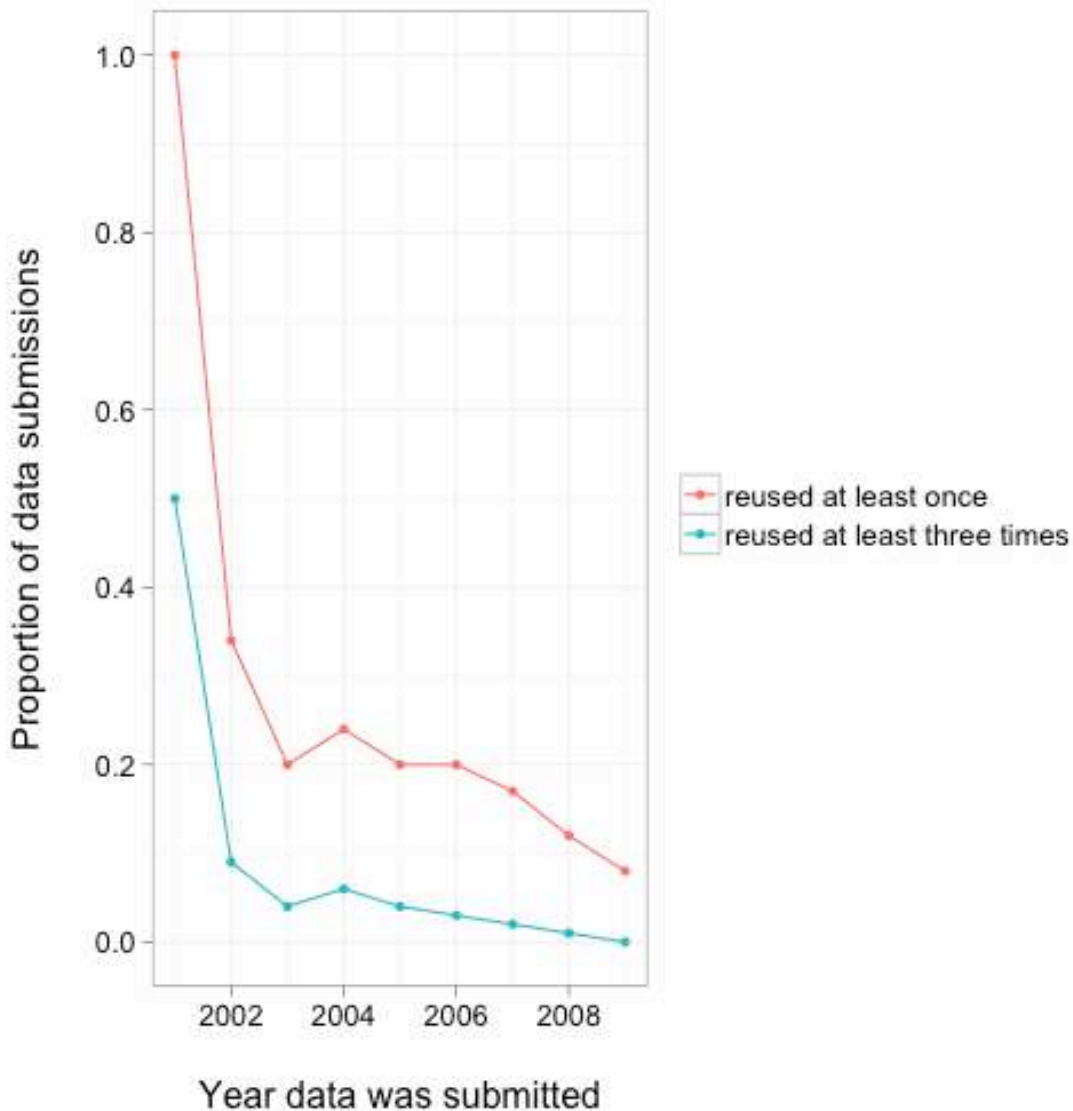


1

2 *Figure 7: Scatterplot of year of publication of third-party reuse paper (with jitter) vs number of*
 3 *GEO datasets mentioned in the paper (log scale). The line connects the mean number of datasets*
 4 *attributed in reuse papers vs publication year.*

5 **Distribution of reuse across datasets**

6 It is useful to know the distribution of reuse amongst datasets. Since our methods only detect
 7 reuse by papers in PubMed Central (a small proportion of the biomedical literature) and only
 8 when the accession number is given in the full text, our estimates of reuse are extremely
 9 conservative. Despite this, we found that reuse was not limited to just a few papers (Figure 8).
 10 Nearly all datasets published in 2001 were reused at least once. The proportion of reused datasets
 11 declined in subsequent years, with a plateau of about 20% for data deposited between 2003 and
 12 2007. The actual rate of reuse across all methods of attribution, and extrapolated to all of
 13 PubMed, is likely much higher

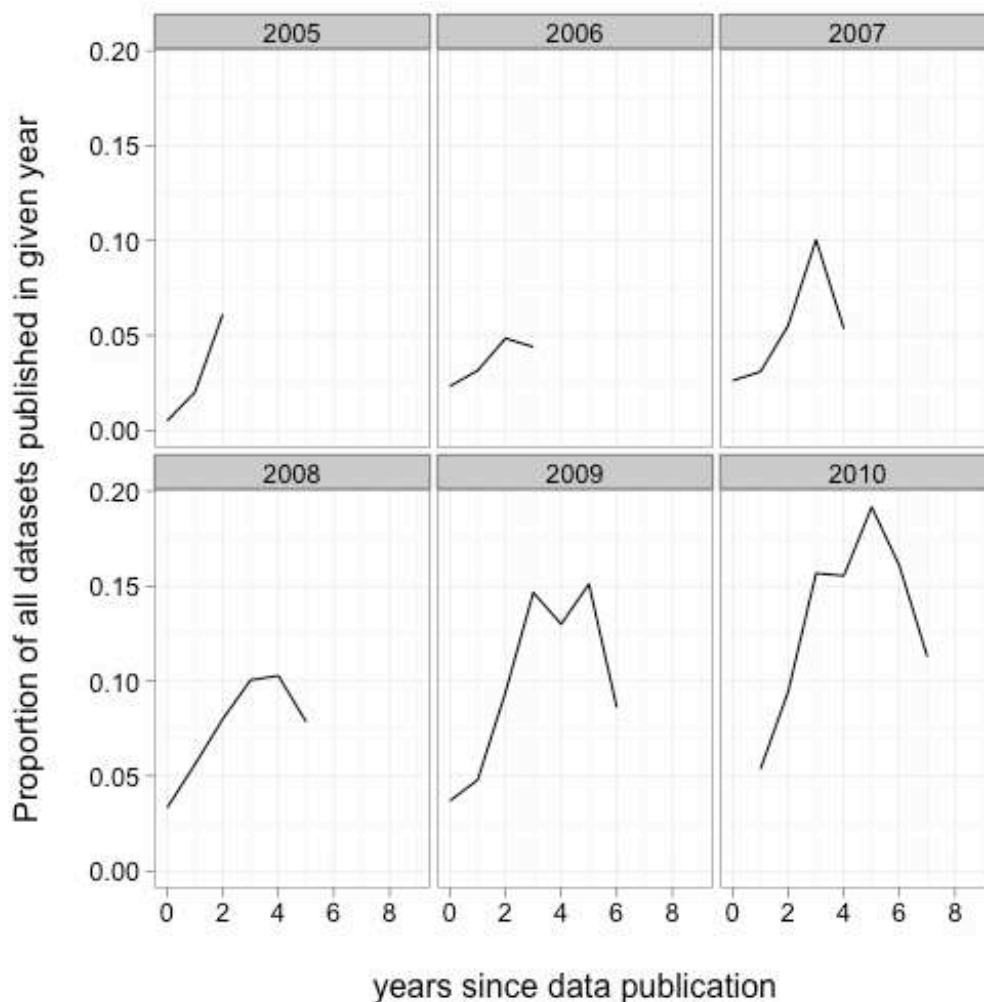


1

2 *Figure 8: Proportion of data reused by third-party papers vs year of data submission. Lower*
 3 *bound, because only considers reuse by papers in PubMed Central, and only when reuse is*
 4 *attributed through direct mention of a GEO accession number*

5 **Distribution of the age of reused data**

6 We found the authors of third-party data reuse papers were most likely to use data that was 3-6
 7 years old by the time their paper was published, normalized for how many datasets were
 8 deposited each year (Figure 9). For example, in aggregate, we found that microarray reuse papers
 9 from 2005 mentioned the accession numbers of more than 5% of all datasets that had been
 10 submitted two years earlier, in 2003. Reuse papers from 2008 mentioned about 7% of the datasets
 11 submitted two years prior (in 2006), more than 10% of the datasets submitted 3 and 4 years prior
 12 (2005 and 2004), and about 7% of the datasets submitted 5 years earlier, in 2003.



1

2 *Figure 9: Proportion of data submissions that contributed to data reuse papers, by year of reuse*
 3 *paper publication and dataset submission. Each panel includes a cohort of data reuse papers*
 4 *published in a given year. The lines indicate the proportion of datasets that were mentioned, in*
 5 *aggregate, by the data reuse papers, by the year of dataset publication. The proportion is relative*
 6 *to the total number of datasets submitted in a given year.*

7 **Discussion**

8 **The open data citation boost**

9 One of the primary findings of this analysis is that papers with publicly available microarray data
 10 received more citations than similar papers that did not make their data available, even after
 11 controlling for many variables known to influence citation rate. We found the open data citation
 12 boost for this sample to be 9% overall (95% confidence interval: 5% to 13%), but the boost
 13 depended heavily on the year the dataset was made available. Datasets deposited very recently
 14 have so far received no (or few) additional citations, while those deposited in 2004-2005 showed
 15 a clear boost of about 30% (confidence intervals 15% to 48%). Older datasets also appeared to
 16 receive a citation boost, but the estimate is less precise because relatively little microarray data
 17 was collected or archived in the early 2000s.

1 The citation boost reported here is smaller than that reported in the previous study by (Piwowar *et. al.* 2007), which estimated a citation boost of 69% for human cancer gene expression microarray studies published before 2003 (95% confidence intervals of 18 to 143%). Our attempt to replicate the (Piwowar *et. al.* 2007) study here suggests that aspects of both the data and analysis can help to explain the quantitatively different results. It appears that clinically relevant datasets released early in the history of microarray analysis were particularly impactful. Importantly, however, the new analysis also suggested that the previous estimate was confounded by significant citation correlates, including the total number of authors and the citation history of the last author. This finding reinforces the importance of accounting for covariates through multivariate analysis and the need for large samples to support full analysis: the 69% estimate is probably too high, even for its high-impact sample. Nonetheless, a 10-30% is citation boost may still be an effective motivator for data deposit, given that prestigious journals have been known advertise their impact factors to three decimal places (Smith, 2006).

14 A paper with open data may be cited for reasons other than data reuse, and open data may be reused without citation of the original paper. Ideally, we would like to separate these two phenomena (data reuse and paper citation) and measure how often the latter is driven by the former. In our manual analysis of 138 citations to papers with open data, we observed that 6% (95% CI: 3% to 11%) of citations were in the context of data reuse. While this methodology and sample size does not allow us to estimate with any precision the proportion of the data citation boost that can be attributed to data reuse, the result is consistent with data reuse being a major contributor.

22 Another result of importance from the citation analysis is that papers based on self data reuse dropped off steeply after two years, while data reuse papers by third-party authors continued to accumulate even after six years. This suggests that while researchers may have some incentive for protecting their own exclusive use of data close to the time of the initial publication, the equation changes dramatically after a short period. This provides some evidence to guide policy decisions regarding the length of data embargoes allowed by journal archiving policies such as the Joint Data Archiving Policy described by (Rausher *et. al.* 2010).

29 **Challenges collecting citation data**

30 This study required obtaining citation counts for thousands of articles using PubMed IDs. This was not supported at the time of data collection using either Thomson Reuter's Web of Science or Google Scholar. While this type of query was (and is) supported by Elsevier's Scopus database, we lacked institutional access to Scopus, individual subscriptions were not available, and attempts to request access through Scopus staff were unsuccessful. One of us (HP) attempted to use the British Library's walk-in access of Scopus while visiting the UK. Unfortunately, the British Library's policies did not permit any method of electronic input of the PubMed identifier list (the list is 10,000 elements long). HP eventually obtained official access to Scopus through a Research Worker agreement with Canada's National Research Library (NRC-CISTI), after being fingerprinted to obtain a police clearance certificate because she had recently lived in the United States.

41 Our understanding of research practice suffers because access to tools and data is so difficult.

42

43

1 **Patterns of data reuse**

2 To better understand patterns of data reuse, a larger sample of reuse instances is needed than can
3 easily be assembled through manual classification of citation context. To that end, we looked at a
4 complementary source of information about reuse of the same datasets: direct mention of GEO or
5 ArrayExpress accession numbers within the body of a full-text research article. The large number
6 of instances of reuse identified this way allowed us to ask questions about the distribution of
7 reuse over time and across datasets. The results indicate that dataset reuse has been increasing
8 over time (excluding the initial years of GEO and ArrayExpress when few datasets were
9 deposited and reuse appears to have been atypically broad). Recent reuse analyses include more
10 datasets, on average, than older reuse studies. Also, the fact that reuse was greatest for datasets
11 published between three and six years previously suggests that the lower citation boost we
12 observed for recent papers is due, at least in part, to a relatively short follow-up time.

13 Extrapolating to all of PubMed, we estimate the number of reuse papers published per year is on
14 the same order of magnitude, and likely greater, than the number of datasets made available. This
15 data reuse curve is remarkably constant for data deposited between 2004 and 2009. This
16 reinforces the conclusions of an earlier analysis: even modest data reuse can provide an
17 impressive return on investment for science funders ([Piwowar et. al. 2011b](#)).

18 We have observed a moderate proportion of datasets being reused by third parties (more than
19 20% of the datasets deposited between 2003 and 2007). It is important to recognize that this is
20 likely a gross underestimate. It includes only those instances of reuse that can be recognized
21 through the mention of accession number in PubMed Central. No attempt has been made to
22 extrapolate these distribution statistics to all of PubMed, nor to reflect additional attributions
23 through paper citations or mentions of the archive name alone. Further, many important instances
24 of data reuse leave no trace in the published literature, such as those in education and training.

25 **Reasons for the data citation boost**

26 While we cannot exclude that the open data citation boost is driven entirely by third-party data
27 reuse, there may be other factors contributing to the effect either directly or indirectly. The
28 literature on possible reasons for an "Open Access citation benefit" suggests a number of factors
29 that may also be relevant to open data ([Craig et. al. 2007](#)). Building upon this work, we suggest
30 several possible sources for an "Open Data citation benefit":

- 31 1. *Data Reuse*. Papers with available datasets can be used in ways that papers without data
32 cannot, and may receive additional citations as a result.
- 33 2. *Credibility Signalling*. The credibility of research findings may be higher for research
34 papers with available data. Such papers may be preferentially chosen as background
35 citations and/or the foundation of additional research.
- 36 3. *Increased Visibility*. Third party researchers may be more likely to encounter a paper that
37 has available data, either by a direct link from the data or indirectly due to cross-
38 promotion. For example, links from a data repository to a paper may increase the search
39 ranking of the research paper.
- 40 4. *Early View*. When data is made available before a paper is published, some citations may
41 accrue earlier than they would otherwise because of accelerated awareness of the
42 methods, findings, etc.

1 5. *Selection Bias*. Authors may be more likely to publish data for papers they judge to be
2 their best quality work, because they are particularly proud or confident in the results
3 ([Wicherts et. al. 2011](#)).

4 Importantly, almost all of these mechanisms are aligned with more efficient and effective
5 scientific progress: increased data use, facilitated credibility determination, earlier access,
6 improved discoverability, and a focus on best work through data availability are good for both
7 investigators and the science community as a whole. Working through the one area where
8 incentives between scientific good and author incentives conflict, finding weaknesses or faults in
9 published research, may require mandates. Or, instead, perhaps the research community will
10 eventually come to associate withheld data with poor quality research, as it does today for
11 findings that are not disclosed in a peer-reviewed paper ([Ware, 2008](#)).

12 The citation boost in the current study is consistent with data reuse observed in this study and the
13 small-scale annotation reported in ([Rung & Brazma, 2013](#)). Nonetheless, it is possible some of the
14 other sources postulated above may have contributed citations for the studies with available data.
15 Further work will be needed to understand the relative contributions from each source. For
16 example, in-depth analyses of all publications from a set of data-collecting authors could support
17 measurement of selection bias. Observing search behavior of researchers, and the returned search
18 hit results, could characterize increased visibility due to data availability. Hypothetical examples
19 could be provided to authors to determine whether they would be systematically more likely to
20 cite a paper with available data in situations where they are considering the credibility of research
21 findings.

22 **Future work**

23 Additional future work can improve on these results by considering and integrating all methods of
24 data use attribution. This holistic effort would include identifying citations to the paper that
25 describes the data collection, mentions of the dataset identifier itself -- whether in full text, the
26 references section, or supplementary information -- citations to the dataset as a first-class research
27 object, and even mentions of the data collection investigators in acknowledgement sections. The
28 citations and mentions would need classification based on context to ensure they are in the
29 context of data reuse.

30 The obstacles encountered in obtaining the citation data needed for this study, as described earlier
31 in the Discussion, demonstrate that improvements in tools and practice are needed to make
32 impact tracking easier and more accurate, for day-to-day analysis as well as studies for evidence-
33 based policy. Such research is hamstrung without programmatic access to the full-text of the
34 research literature and to the citation databases that underpin impact assessment. The lack of
35 conventions and tool support for data attribution ([Mooney & Newton, 2012](#)) is also a significant
36 obstacle, and undoubtedly led to undercounting in the present study. There is much room for
37 improvement, and we are hopeful about recent steps toward data citation standards taken by
38 initiatives such as DataCite.

39 Data from current and future studies can start to be used to estimate the impact of policy
40 decisions. For example, do embargo periods decrease the level of data reuse? Do restrictive or
41 poorly articulated licensing terms decrease data reuse? Which types of data reuse are facilitates
42 by robust data standards and which types are unaffected?

1 Qualitative assessment of data reuse is an essential complement to large-scale quantitative
2 analyses. Repeating and extending previous studies will help us to understand the potential of
3 data reuse, areas of progress, and remaining challenges (e.g. ([Zimmerman, 2003](#) ; [Wan & Pavlidis,](#)
4 [2007](#) ; [Wynholds et. al. 2012](#) ; [Rolland & Lee, 2013](#))). Usage statistics from primary data repositories
5 and value-added repositories are also useful sources of insight into reuse patterns ([Rung & Brazma,](#)
6 [2013](#)).

7 Citations are blind to many important types of data reuse. The impact of data on practitioners,
8 educators, data journalists, and industry researchers are not captured by attribution patterns in the
9 scientific literature. Altmetrics indicators uncover discussions in social social media, syllabi,
10 patents, and theses: analyzing such indicators for datasets would provide valuable evidence of
11 reuse beyond the scientific literature. As evaluators move away from assessing research based on
12 journal impact factor and toward article-level metrics, post-publication metrics rates will become
13 increasingly important indicators of research impact ([Piwowar, 2013](#)).

14 **Conclusions**

15 We find a robust citation benefit from open data, although a smaller one than previously reported.
16 We conclude there is a direct effect of third-party data reuse that persists for years beyond the
17 time when researchers have published most of the papers reusing their own data. We further
18 conclude that, at least for gene expression microarray data, a substantial fraction of archived
19 datasets are reused, and that the intensity of dataset reuse has been steadily increasing since 2003.

20 It is important to remember that the primary rationale for making research data available has
21 nothing to do with evaluation metrics or citation benefits: a full account of experimental process
22 and findings is a tenant of science, and publicly-funded science is a public resource ([Smith, 2006](#)).
23 Nonetheless, robust evidence of personal benefit will help as science transitions from "data not
24 shown" to a culture that simply expects data to be part of the published record.

25 **Acknowledgements**

26 The authors thank Angus Whyte for suggestions on study design. We thank Jonathan Carlson and
27 Estephania Sta. Maria for their hard work on data collection and annotation. Michael Whitlock
28 and the Biodiversity Research Centre at the University of British Columbia provided community
29 and resources. Finally, we are grateful to everyone who helped with access to Scopus, particularly
30 Andre Vellino, CISTI, and friends at the British Library.

31 **References**

32 Publication references are available in a publicly-available [Mendeley group](#) to facilitate
33 exploration.

- 34 • Carl Boettiger, (2013) knitcitations: Citations for knitr markdown files.
35 <https://github.com/cboettig/knitcitations>
- 36 • Gregory Bolker, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Liaw, Thomas
37 Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz, Bill Venables,
38 (2012) gplots: Various R programming tools for plotting data. [http://CRAN.R-](http://CRAN.R-project.org/package=gplots)
39 [project.org/package=gplots](http://CRAN.R-project.org/package=gplots)

- 1 • I Craig, A Plum, M McVeigh, J Pringle, M Amin, (2007) Do open access articles have
2 greater citation impact? A critical review of the literature. *Journal of Informetrics* 1 (3) 239-
3 248 [10.1016/j.joi.2007.04.001](https://doi.org/10.1016/j.joi.2007.04.001)
- 4 • Bertil Dorch, (2012) On the Citation Advantage of linking to data. *hprints*
5 <http://hprints.org/hprints-00714715>
- 6 • John Fox, (2010) polycor: Polychoric and Polyserial Correlations. [http://CRAN.R-](http://CRAN.R-project.org/package=polycor)
7 [project.org/package=polycor](http://CRAN.R-project.org/package=polycor)
- 8 • Lawrence Fu, Constantin Aliferis, (2008) Models for predicting and explaining citation count
9 of biomedical articles.. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA*
10 *Symposium* 222-6
11 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2656101&tool=pmcentrez&rende](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2656101&tool=pmcentrez&rendertype=abstract)
12 [rttype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2656101&tool=pmcentrez&rendertype=abstract)
- 13 • Nils Gleditsch, Havard Strand, (2003) Posting your data: will you be scooped or will you be
14 famous?. *International Studies Perspectives* 4 (1) 89-97 [http://www.prio.no/Research-and-](http://www.prio.no/Research-and-Publications/Publication/?oid=55406)
15 [Publications/Publication/?oid=55406](http://www.prio.no/Research-and-Publications/Publication/?oid=55406)
- 16 • David Hajage, (2011) ascii: Export R objects to several markup languages. [http://CRAN.R-](http://CRAN.R-project.org/package=ascii)
17 [project.org/package=ascii](http://CRAN.R-project.org/package=ascii)
- 18 • Edwin Henneken, Alberto Accomazzi, (2011) Linking to Data - Effect on Citation Rates in
19 Astronomy. *arXiv* 4-NA <http://arxiv.org/abs/1111.3618>
- 20 • John Ioannidis, David Allison, Catherine Ball, Issa Coulibaly, Xiangqin Cui, Aed'n Culhane,
21 Mario Falchi, Cesare Furlanello, Laurence Game, Giuseppe Jurman, Jon Mangion, Tapan
22 Mehta, Michael Nitzberg, Grier Page, Enrico Petretto, Vera Noort, (2009) Repeatability of
23 published microarray gene expression analyses.. *Nature genetics* 41 (2) 149-55
24 [10.1038/ng.295](https://doi.org/10.1038/ng.295)
- 25 • Hailey Mooney, Mark Newton, (2012) The Anatomy of a Data Citation: Discovery, Reuse,
26 and Credit. *Journal of Librarianship and Scholarly Communication* 1 (1) [http://jlscc-](http://jlscc-pub.org/jlsc/vol1/iss1/6)
27 [pub.org/jlsc/vol1/iss1/6](http://jlscc-pub.org/jlsc/vol1/iss1/6)
- 28 • Frank Harrell Jr, (2012) rms: Regression Modeling Strategies. [http://CRAN.R-](http://CRAN.R-project.org/package=rms)
29 [project.org/package=rms](http://CRAN.R-project.org/package=rms)
- 30 • Amy Pienta, George Alter, Jared Lyle, (2010) The Enduring Value of Social Science
31 Research: The Use and Reuse of Primary Research Data. *The Organisation, Economics and*
32 *Policy of Scientific Research workshop* <http://hdl.handle.net/2027.42/78307>
- 33 • Heather Piwowar, Roger Day, Douglas Fridsma, (2007) Sharing detailed research data is
34 associated with increased citation rate. *PLoS ONE* 2 (3)
35 <http://dx.doi.org/10.1371/journal.pone.0000308>
- 36 • Heather Piwowar, Wendy Chapman, (2010) Recall and bias of retrieving gene expression
37 microarray datasets through PubMed identifiers.. *Journal of biomedical discovery and*
38 *collaboration* 5 7-20 <http://www.ncbi.nlm.nih.gov/pubmed/20349403>
- 39 • Heather Piwowar, Jonathan Carlson, Todd Vision, (2011a) Beginning to track 1000 datasets
40 from public repositories into the published literature. *Proceedings of the American Society for*
41 *Information Science and Technology* 48 (1) 1-4 [10.1002/meet.2011.14504801337](https://doi.org/10.1002/meet.2011.14504801337)
- 42 • Heather Piwowar, Todd Vision, Michael Whitlock, (2011b) Data archiving is a good
43 investment.. *Nature* 473 (7347) 285-NA [10.1038/473285a](https://doi.org/10.1038/473285a)
- 44 • Heather Piwowar, (2011c) Data from: Who shares? Who doesn't? Factors associated with
45 openly archiving raw research data. *Dryad Digital Repository* [10.5061/dryad.mf1sd](https://doi.org/10.5061/dryad.mf1sd)
- 46 • Heather Piwowar, (2011d) Who Shares? Who Doesn't? Factors Associated with Openly
47 Archiving Raw Research Data. *PLoS ONE* 6 (7) e18657-NA [10.1371/journal.pone.0018657](https://doi.org/10.1371/journal.pone.0018657)
- 48 • Heather Piwowar, (2013) Value all research products.. *Nature* 493 (7431) 159-NA
49 [10.1038/493159a](https://doi.org/10.1038/493159a)

- 1 • Mark Rausher, Mark McPeck, Allen Moore, Loren Rieseberg, Michael Whitlock, (2010)
2 Data archiving.. *Evolution; international journal of organic evolution* 64 (3) 603-4
3 <http://www.ncbi.nlm.nih.gov/pubmed/20050907>
- 4 • Betsy Rolland, Charlotte Lee, (2013) Beyond trust and reliability. 435-NA
5 <http://dl.acm.org/citation.cfm?id=2441776.2441826>
- 6 • Johan Rung, Alvis Brazma, (2013) Reuse of public genome-wide gene expression data..
7 *Nature reviews. Genetics* 14 (2) 89-99 <http://www.ncbi.nlm.nih.gov/pubmed/23269463>
- 8 • Jon Sears, (2011) Data Sharing Effect on Article Citation Rate in Paleocyanography -
9 KomFor. <http://www.komfor.net/blog/unbenanntemitteilung>
- 10 • Richard Smith, (2006) Commentary: the power of the unrelenting impact factor--is it a force
11 for good or harm?. *International journal of epidemiology* 35 (5) 1129-30
12 <http://ije.oxfordjournals.org/content/35/5/1129.full>
- 13 • Carol Tenopir, Suzie Allard, Kimberly Douglass, Arsev Aydinoglu, Lei Wu, Eleanor Read,
14 Maribeth Manoff, Mike Frame, (2011) Data sharing by scientists: practices and perceptions..
15 *PLoS one* 6 (6) e21101-NA [10.1371/journal.pone.0021101](https://doi.org/10.1371/journal.pone.0021101)
- 16 • Xiang Wan, Paul Pavlidis, (2007) Sharing and reusing gene expression profiling data in
17 neuroscience.. *Neuroinformatics* 5 (3) 161-75
18 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2980754&tool=pmcentrez&rende](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2980754&tool=pmcentrez&rendertype=abstract)
19 [rtype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2980754&tool=pmcentrez&rendertype=abstract)
- 20 • Mark Ware, (2008) Peer review: benefits, perceptions and alternatives. *PRC Summary Papers*
21 *4* <http://www.publishingresearch.net/documents/PRCsummary4Warefinal.pdf>
- 22 • Jelte Wicherts, Marjan Bakker, Dylan Molenaar, (2011) Willingness to share research data is
23 related to the strength of the evidence and the quality of reporting of statistical results.. *PloS*
24 *one* 6 (11) e26828-NA [10.1371/journal.pone.0026828](https://doi.org/10.1371/journal.pone.0026828)
- 25 • Hadley Wickham, (2007) Reshaping Data with the reshape Package. *Journal of Statistical*
26 *Software* 21 (12) 1-20 <http://www.jstatsoft.org/v21/i12/>
- 27 • Hadley Wickham, (2009) ggplot2: elegant graphics for data analysis.
28 <http://had.co.nz/ggplot2/book>
- 29 • Hadley Wickham, (2011) The Split-Apply-Combine Strategy for Data Analysis. *Journal of*
30 *Statistical Software* 40 (1) 1-29 <http://www.jstatsoft.org/v40/i01/>
- 31 • Laura Wynholds, Jillian Wallis, Christine Borgman, Ashley Sands, Sharon Traweek, (2012)
32 Data, data use, and scientific inquiry. 19-NA [10.1145/2232817.2232822](https://doi.org/10.1145/2232817.2232822)
- 33 • Yihui Xie, (2012) knitr: A general-purpose package for dynamic report generation in R.
34 <http://CRAN.R-project.org/package=knitr>
- 35 • Ann Zimmerman, (2003) Data Sharing and Secondary Use of Scientific Data: Experiences of
36 Ecologists. *Dissertations and Theses (Ph.D. and Master's)*
37 <http://hdl.handle.net/2027.42/39373>