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.

PeerJ PrePrints | https://peerj.com/preprints/1 | v1 received: 3 Apr 2013, published: 4 Apr 2013, doi: 10.7287/peerj.preprints.1

Introduction

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

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

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

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

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

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

PeerJ PrePrints | https://peerj.com/preprints/1 | v1 received: 3 Apr 2013, published: 4 Apr 2013, doi: 10.7287/peerj.preprints.1

Materials and Methods

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

Relationship between data availability and citation

Data collection

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

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

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

Primary analysis

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

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

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

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

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

Comparison to 2007 study

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

Stratification by year

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

Manual review of citation context

We manually reviewed the context of citations to data collection papers to estimate how many citations to data collection papers were made in the context of data reuse. We (Jonathan Carlton, in acknowledgements) reviewed 50 citations chosen randomly from the set of all citations to 100 data collection papers. Specifically, we randomly selected 100 datasets deposited in GEO in 2005. For each dataset, we located the data collection article within ISI Web of Science based on its title and authors, and exported the list of all articles that cited this data collection article. From this, we selected 50 random citations stratified by the total number of times the data collection article had been cited. By manual review of the relevant full-text of each paper, we determined if the data from the associated dataset had been reused within the study.

Data reuse patterns from accession number attribution

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

Data collection

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

Statistical analysis

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

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

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

Data and script availability

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

PeerJ PrePrints | https://peerj.com/preprints/1 | v1 received: 3 Apr 2013, published: 4 Apr 2013, doi: 10.7287/peerj.preprints.1

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

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

Results

Description of cohort

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

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

Cancer Res	0.04
Proc Natl Acad Sci U S A	0.04
J Biol Chem	0.04
BMC Genomics	0.03
Physiol Genomics	0.03
PLoS One	0.02
J Bacteriol	0.02
J Immunol	0.02
Blood	0.02
Clin Cancer Res	0.02
Plant Physiol	0.02
Mol Cell Biol	0.01

Table 1: Proportion of sample published in most common journals

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

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

Table 2: Proportion of sample published each year

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

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

Data availability is associated with citation boost

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

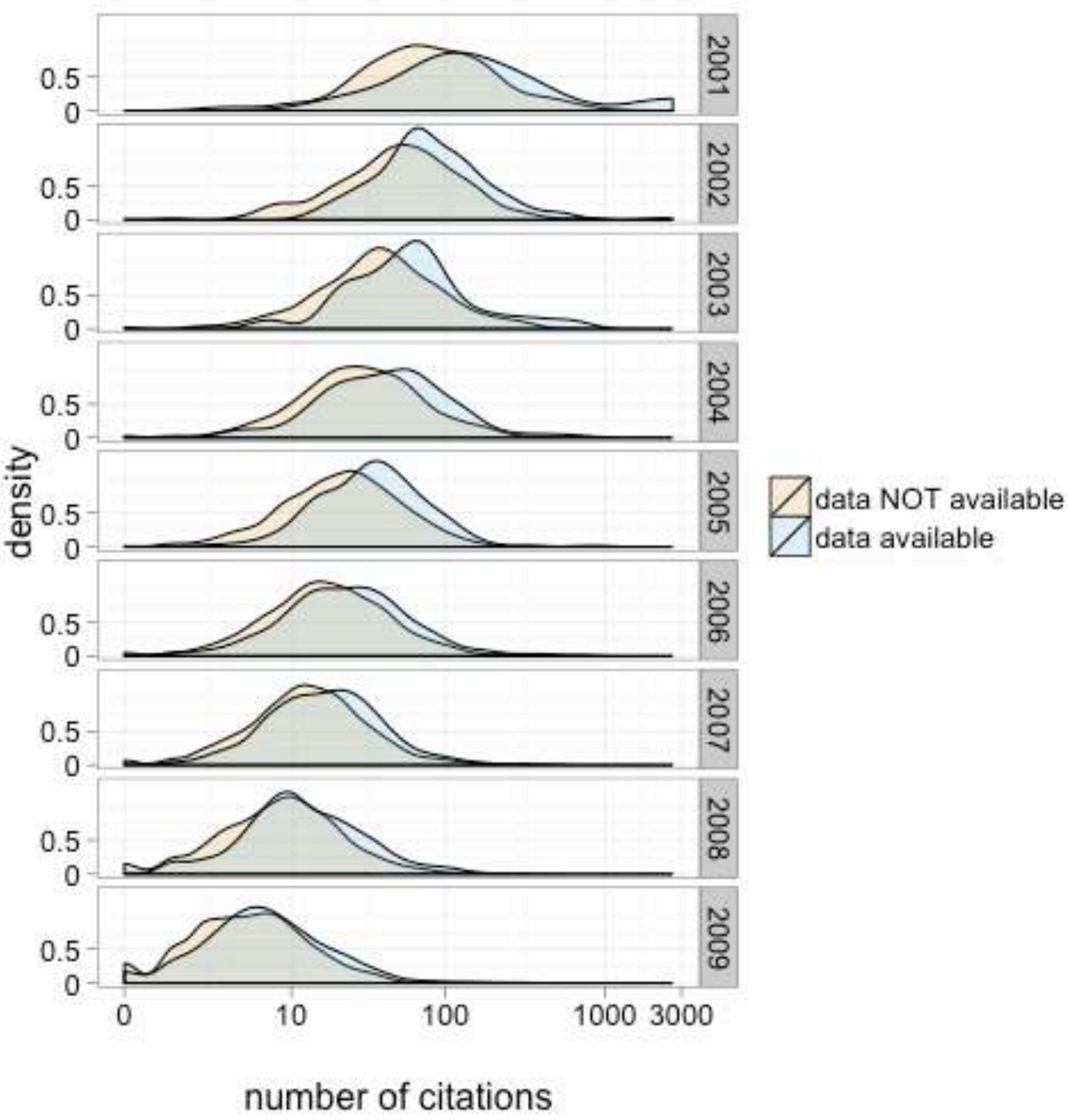

Figure 1: Citation density for papers with and without publicly available microarray data, by year of study publication

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

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

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

More covariates led to a more conservative estimate

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

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

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

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

Citation boost over time

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

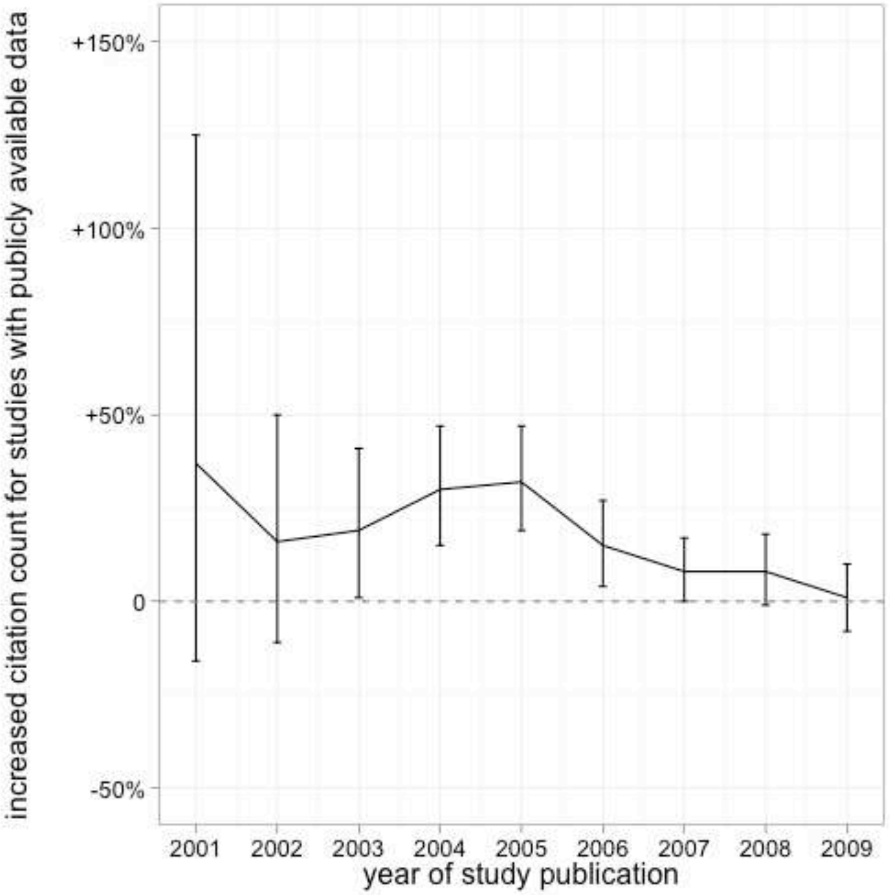

Figure 2: Increased citation count for studies with publicly available data, by year of publication. Estimates from multivariate analysis, lines indicate 95% confidence intervals.

Data reuse is a demonstrable component of citation boost

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

Evidence of reuse from mention of dataset identifiers in full text

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

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

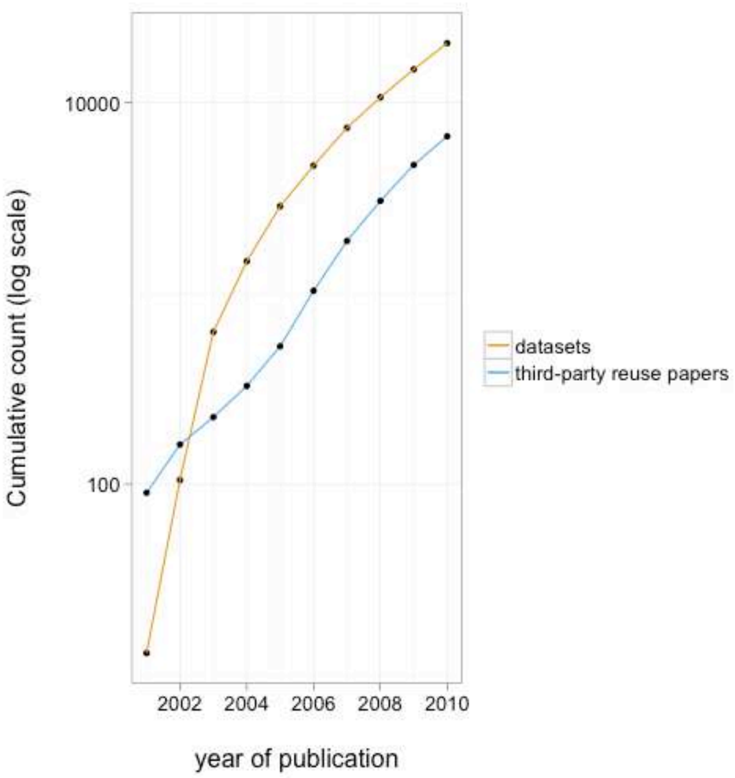

Figure 3: Cumulative number of datasets deposited in GEO each year, and cumulative number of third-party reuse papers published that directly attribute GEO data published each year, log scale.

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

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

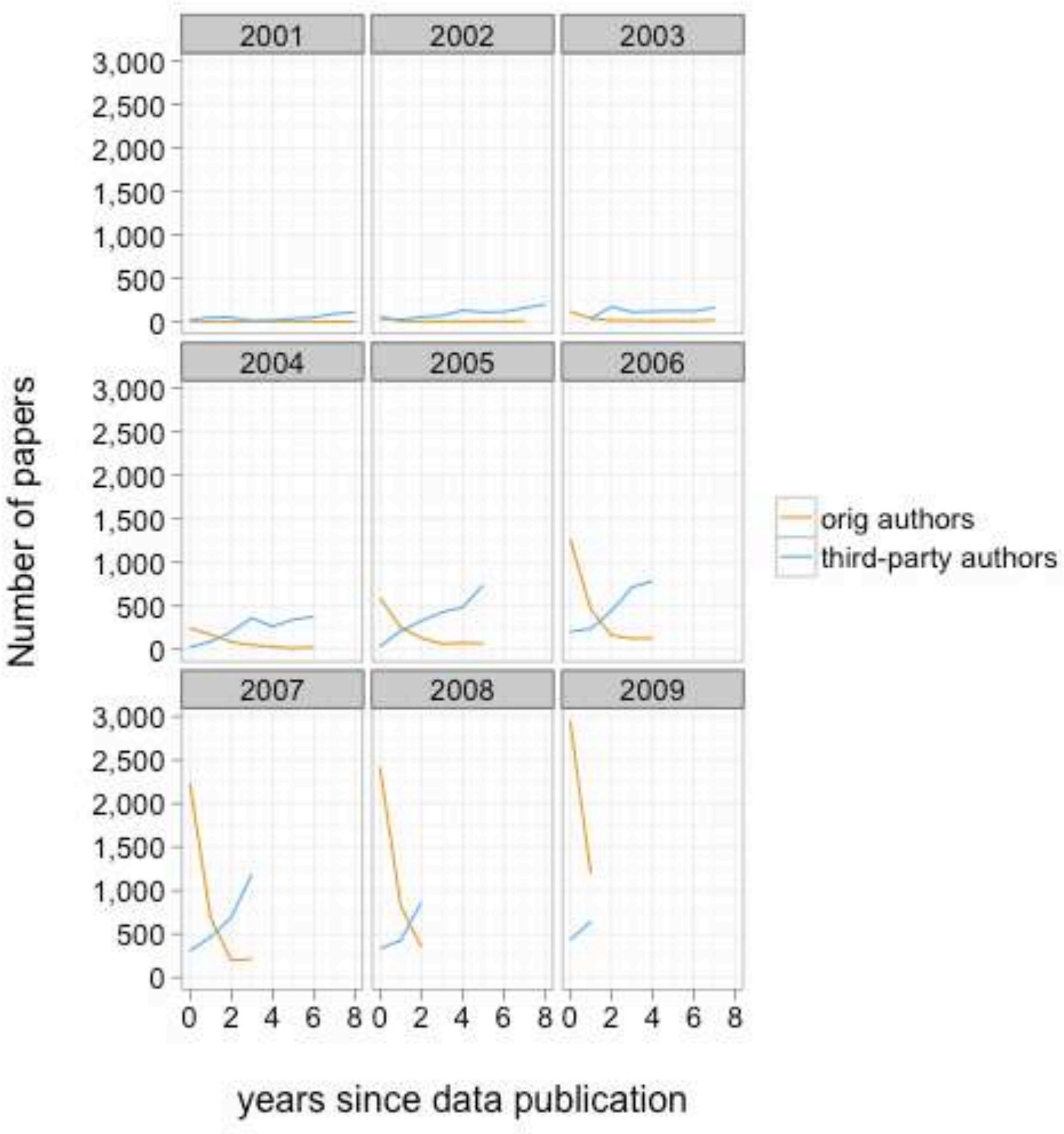

Figure 4: Number of papers mentioning GEO accession numbers. Each panel represents reuse of a particular year of dataset submissions, with number of mentions on the y axis, years since the initial publication on the x axis, and a line for reuses by the data collection team and a line for third-party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

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

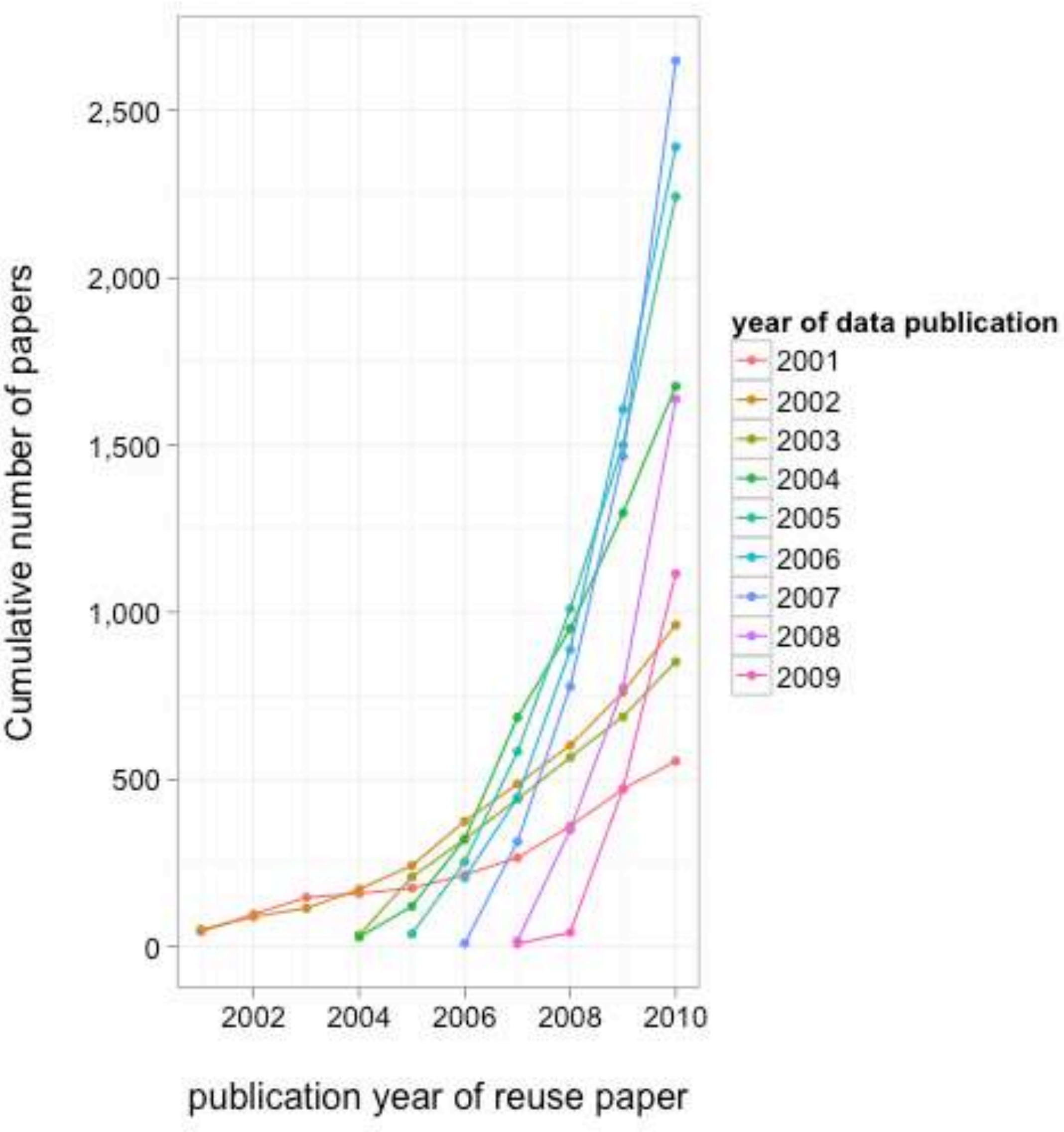

Figure 5: Cumulative number of third-party reuse papers, by date of reuse paper publication. Separate lines are displayed for different dataset submission years

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

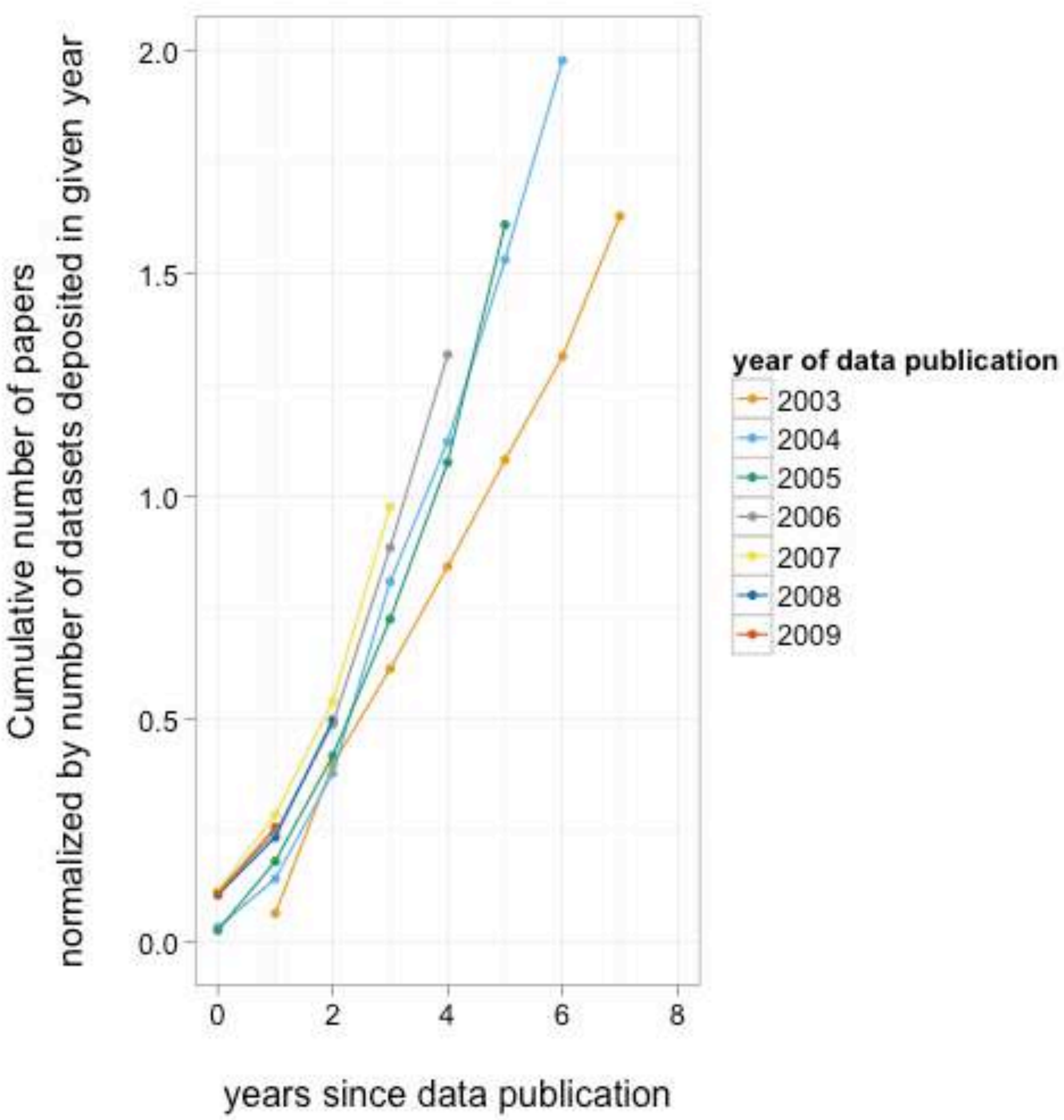

Figure 6: Cumulative third-party reuse, normalized by number of datasets deposited each year, plotted as elapsed years since data publication

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

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

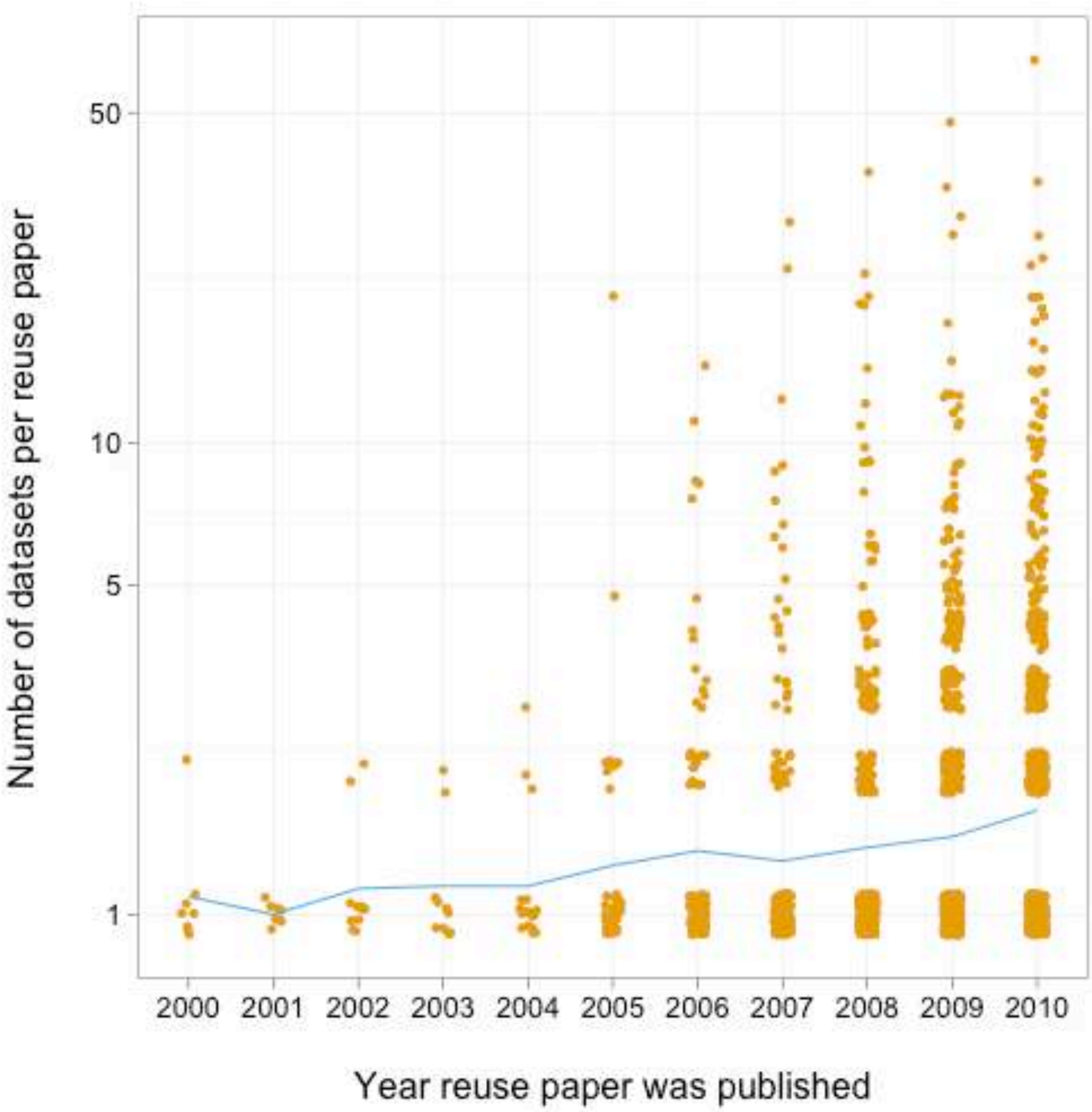

Figure 7: Scatterplot of year of publication of third-party reuse paper (with jitter) vs number of GEO datasets mentioned in the paper (log scale). The line connects the mean number of datasets attributed in reuse papers vs publication year.

Distribution of reuse across datasets

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

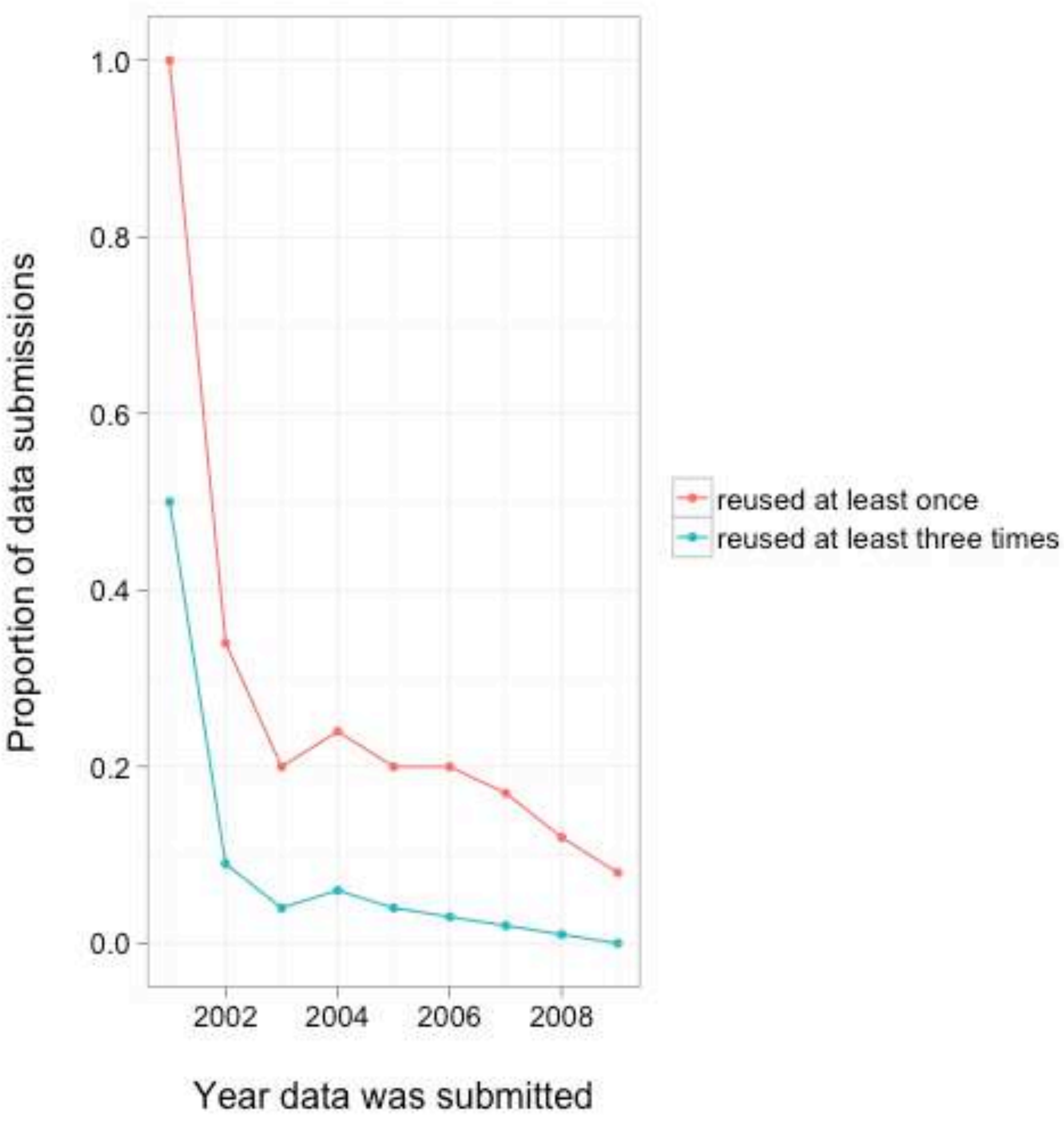

Figure 8: Proportion of data reused by third-party papers vs year of data submission. Lower bound, because only considers reuse by papers in PubMed Central, and only when reuse is attributed through direct mention of a GEO accession number

Distribution of the age of reused data

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

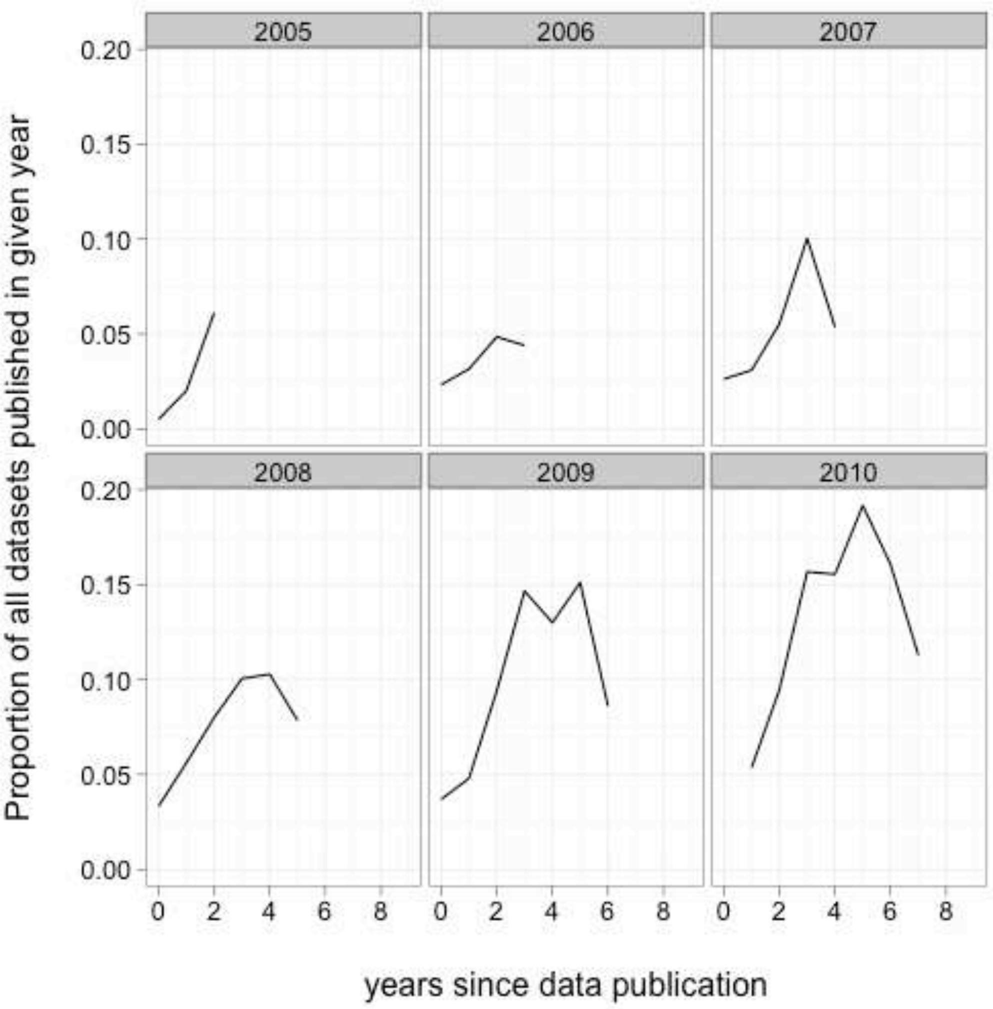

Figure 9: Proportion of data submissions that contributed to data reuse papers, by year of reuse paper publication and dataset submission. Each panel includes a cohort of data reuse papers published in a given year. The lines indicate the proportion of datasets that were mentioned, in aggregate, by the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

Discussion

The open data citation boost

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

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).

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.

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).

Challenges collecting citation data

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.

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

Patterns of data reuse

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

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

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

Reasons for the data citation boost

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

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

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

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

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

Future work

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

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

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

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

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

Conclusions

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. 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.

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

Acknowledgements

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

References

Publication references are available in a publicly-available Mendeley group to facilitate exploration.

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

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

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