# Who shares? Who doesn't? Bibliometric factors associated with open archiving of biomedical datasets

**Heather A Piwowar**
National Evolutionary Synthesis Center
2024 W. Main Street, Suite A200
Durham, NC 27705
hpiwowar@nescent.org

## ABSTRACT

Many initiatives encourage investigators to share their raw research datasets in pursuit of increased research quality and efficiency. Despite these investments of time and money, we do not yet understand the impact of these initiatives. In this study, I use bibliometric methods to understand the prevalence and patterns with which investigators publicly share their raw gene expression microarray datasets after study publication.

Automated methods were used to identify 11,603 published studies that created gene expression microarray data. At least 25% of these studies have datasets in one of the two predominant public databases for microarray data, increasing from 5% in 2001 to 35% in 2009. Fifteen factors that described authorship, funding, institution, publication, and domain environments were derived from 124 article attributes. Most factors associated with the prevalence of data sharing ($p<0.01$). In particular, publishing in a journal with a relatively strong data sharing policy, receiving funding from many NIH grants, publishing in an open access journal, and having prior experience sharing gene expression data were associated with the highest data sharing rates. In contrast, increased first author age and experience, having no experience reusing data, and studying cancer and human subjects were associated with the lowest data sharing rates.

In second-order factor analysis, previously sharing gene expression microarray data was most positively associated with high data sharing rates, whereas publishing a study on cancer or human subjects was strongly associated with a negative probability of data sharing.

*ASIST 2010,* October 22–27, 2010, Pittsburgh, PA, USA.



I hope these methods and results will contribute to a deeper understanding of data sharing behavior and eventually more effective data sharing initiatives.

### Keywords

data sharing, data archiving, bioinformatics, bibliometrics, science policy, human information behavior

## INTRODUCTION

Reuse of raw research datasets has the potential to increase research efficiency and quality. Eager to realize these benefits, funders, publishers, societies, and individual research groups have developed tools, resources, and policies to encourage investigators to make their data publicly available. Despite these investments of time and money, we do not yet understand the impact of these initiatives.

Dimensions of data sharing attitudes and action have been investigated by several surveys, such as Blumenthal (2006) and Hedstrom (2006). Studying research data-sharing behavior through bibliometric analyses may provide additional insights. I have collected and analyzed a large set of observed data sharing actions and associated experiment, investigator, journal, funding, and institutional variables. Common factors behind the attributes were explored, and the association between these factors and data sharing prevalence were explored through multivariate analysis.

## METHOD

The set of studies that *created* gene expression microarray datasets was identified by querying the title, abstract, and full-text of PubMed, PubMed Central, Highwire Press, Scirus, and Google Scholar with portal-specific variants of a previously-derived full-text query. Previous evaluation suggested the query could identify articles that created microarray data with a precision of 90% and a recall of 56%, compared to manual curation.

To determine whether these studies had an associated dataset archived in a public centralized repository, I queried the NCBI's Gene Expression Omnibus and EBI's ArrayExpress databases with article PubMed identifiers. In

a previous evaluation, this approach located 77% of all publicly archived datasets (Piwowar, 2010).

I collected 124 attributes of each study from a wide variety of bibliometric and web-based sources, including MEDLINE, NIH award history, SCImago Institutions Rankings, ISI Journal Citation Reports, journal Instruction to Author policies, Authori-ity authorship clusters (Torvik & Smalheiser, 2009), and a gender-guessing web service.

First and second order factors were extracted from these attributes and the statistical association between the factor scores and data archiving behaviour was determined through logistic regression.

## RESULTS

Queries for identifying microarray data-producing articles returned PubMed identifiers for 11,603 studies. PubMed identifiers were found in GEO or ArrayExpress primary citation fields for 2,901 of the 11,603 articles in our dataset, suggesting that at least 25% of the studies have data in these databases. The archiving rate increased across time, as seen in Figure 1.

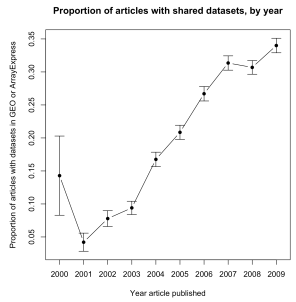


**Figure 1: Archiving rate over time**

In univariate analyses, several factors were correlated with frequency of data sharing. Publishing in a journal with a relatively strong data sharing policy, having funding from many NIH grants, publishing in an open access journal, and having prior experience sharing gene expression microarray data were associated with the highest data sharing rates. In contrast, increased first author age and experience, having no experience reusing data, and studying cancer and human subjects were associated with the lowest data sharing rates.

All five second-order factors were associated with data sharing in multivariate logistic regression, $p<0.001$. Previously sharing gene expression microarray data was most positively associated with high data sharing rates, whereas publishing a study on cancer or human subjects was strongly associated with a negative probability of data sharing.

## DISCUSSION

Although the current results should be considered preliminary, it is disheartening to discover that datasets of human and cancer studies have such low rates of data archiving, particularly because gene expression data can be shared without breaching patient privacy. It is intriguing that publishing in an open access journal, previously sharing gene expression data, and previously reusing gene expression data were associated with high levels of data sharing.

Analyzing data sharing through bibliometric and data-mining attributes has several advantages: we can look at a very large set of studies and attributes, results are not biased by survey response self-selection or reporting bias, and the analysis can be repeated over time with little additional effort.

However, this approach does suffer its own limitations. Filters for identifying microarray creation studies do not have perfect precision, so I may have included some non-data-creation studies in the analysis. Because studies that do not create data will not have data deposits, their inclusion alters the composition of what I consider to be studies that create but do not share data. The method for detecting data deposits overlooks data deposits that are missing PubMed identifiers in GEO and ArrayExpress. Missing data may have obscured important information. Finally, this study did not find deposits that had been submitted to GEO as a series, unless they had been assembled into a DataSet, a curation step for which GEO admits a current backlog.

Due to these limitations, care should be taken in interpreting the estimated levels of absolute data sharing and the data-sharing status of any particular study listed in the raw data. Nonetheless, the aggregate data supports relative trends worthy of additional investigation.

## AVAILABILITY

In the spirit of the topic, raw data and code for this study are openly available online:
http://openwetware.org/wiki/User:Heather_A_Piwowar


## ACKNOWLEDGMENTS

This work was done while HAP was a grad student in the Department of Biomedical Informatics at the University of Pittsburgh. Thanks to advisor Wendy Chapman for support.


## REFERENCES


Blumenthal D, Campbell EG, Gokhale M, Yucel R, Clarridge B, et al. (2006) Data withholding in genetics and the other life sciences: prevalences and predictors. *Acad Med.* 81: 137-145.

Hedstrom M. (2006) Producing Archive-Ready Datasets: Compliance, Incentives, and Motivation. *IASSIST conference*.

Piwowar HA, Chapman WW. (2010) Recall and bias of retrieving gene expression microarray datasets through PubMed identifiers. *J Biomed Discov Collab*. Mar 28;5:7-20.

Torvik VI and Smalheiser NR. (2009) Author Name Disambiguation in MEDLINE. *ACM Trans Knowl Discov Data*. Jul 1; 3(3).