

A review of journal policies for sharing research data

Heather A. Piwowar; Wendy W. Chapman

Department of Biomedical Informatics, University of Pittsburgh
200 Meyran Avenue, Pittsburgh PA, USA
e-mail: hpiwowar@gmail.com; wec6@pitt.edu

Abstract

Background: Sharing data is a tenet of science, yet commonplace in only a few subdisciplines. Recognizing that a data sharing culture is unlikely to be achieved without policy guidance, some funders and journals have begun to request and require that investigators share their primary datasets with other researchers. The purpose of this study is to understand the current state of data sharing policies within journals, the features of journals that are associated with the strength of their data sharing policies, and whether the strength of data sharing policies impact the observed prevalence of data sharing. **Methods:** We investigated these relationships with respect to gene expression microarray data in the journals that most often publish studies about this type of data. We measured data sharing prevalence as the proportion of papers with submission links from NCBI's Gene Expression Omnibus (GEO) database. We conducted univariate and linear multivariate regressions to understand the relationship between the strength of data sharing policy and journal impact factor, journal subdiscipline, journal publisher (academic societies vs. commercial), and publishing model (open vs. closed access). **Results:** Of the 70 journal policies, 53 made some mention of sharing publication-related data within their Instruction to Author statements. Of the 40 policies with a data sharing policy applicable to gene expression microarrays, we classified 17 as weak and 23 as strong (strong policies required an accession number from database submission prior to publication). Existence of a data sharing policy was associated with the type of journal publisher: 46% of commercial journals had data sharing policy, compared to 82% of journals published by an academic society. All five of the open-access journals had a data sharing policy. Policy strength was associated with impact factor: the journals with no data sharing policy, a weak policy, and a strong policy had respective median impact factors of 3.6, 4.9, and 6.2. Policy strength was positively associated with measured data sharing submission into the GEO database: the journals with no data sharing policy, a weak policy, and a strong policy had median data sharing prevalence of 8%, 20%, and 25%, respectively. **Conclusion:** This review and analysis begins to quantify the relationship between journal policies and data sharing outcomes. We hope it contributes to assessing the incentives and initiatives designed to facilitate widespread, responsible, effective data sharing.

Keywords: data sharing; editorial policies; instructions for authors; bibliometrics; gene expression microarrays

1. Background

Widespread adoption of the Internet now allows research results to be shared more readily than ever before. This is true not only for published research reports, but also for the raw research data points that underlie the reports. Investigators who collect and analyze data can submit their datasets to online databases, post them on websites, and include them as electronic supplemental information – thereby making the data easy to examine and reuse by other researchers.

Reusing research data has many benefits for the scientific community. New research hypotheses can be tested more quickly and inexpensively when duplicate data collection is reduced. Data can be aggregated to study otherwise-intractable issues, and a more diverse set of scientists can become involved when analysis is opened beyond those who collected the original data. Ethically, it has long been considered a tenet of scientific behavior to share results[1], thereby allowing close examination of research conclusions and facilitating others to build directly on previous work. The ethical position is even stronger when the research has been funded by public money[2], or the data are donated by patients and so should be used to advance science by the greatest extent permitted by the donors[3].

Unfortunately, these advantages only indirectly benefit the stakeholders who bear most of the costs for sharing their datasets: the primary data-producing investigators. Data sharing is often time consuming, confusing, scary,

and potentially damaging to future research plans. Consequently, sharing data is commonplace in only a few subdisciplines.

Recognizing that a data sharing culture is unlikely to be achieved without policy guidance, some funders and journals have begun to request and require that investigators share their primary datasets with other researchers. Funders are motivated by the promise of resource efficiency and rapid progress. The motivation for journals to act as an advocate and gatekeeper for data sharing is less straightforward. Journals seek to publish “well-written, properly formatted research that meets community standards” and in so doing have assumed monitoring tasks to “remind researchers of community expectations and enforce some behaviors seen as advantageous to the progress of science.”[4] This role has been encouraged by many letters[5, 6], white-papers[7, 8], and editorials in high-profile journals[9].

Journal policies are usually expressed within “instruction for authors” statements. A study by McCain in 1995[4] explored the statements of 850 journals, looking for mandates for the dissemination of data (and the sharing of biological materials). She found that 132 (16%) natural science and technology journals had a policy regarding sharing of some type of research-related information. While McCain covered a wide breadth and depth of journals (especially given that her review predated electronic access to instruction for author statements), she did not attempt to associate the policies with journal attributes, nor did she measure the actual data sharing behavior of authors and correlate the prevalence with journal policy strength. We believe looking at these issues could help us better understand the causes and effects of journal data sharing policies.

The purpose of this study is to understand the current state of data sharing policies within journals, to identify which characteristics of journals are associated with the strength of their data sharing policies, and to measure whether the strength of data sharing policies impacts the observed prevalence of data sharing.

2. Methodology

Our study involved three steps. First, we identified a set of journals for examination. For each journal, based on a manual review of the instruction to author statement, we classified the strength of its policy for data sharing as none, weak, or strong. Second, we studied the relationship between the strength of a journal’s data sharing policy and selected journal attributes. Third, for each journal, we measured how many of its recently published articles have submitted datasets to a centralized database. We used these estimates to study the relationship between data sharing prevalence and the strength of the journal’s data sharing policy. Each of these steps is described below in more detail.

2.1 Collecting the journal’s policies on sharing data

To avoid unnecessary complexity, we chose to investigate data sharing policies for a single type of data: biological gene expression microarrays. These “chips” allow investigators to measure the relative level of RNA expression across tens of thousands (exponentially more each year, as the technology improves) of different genes for each cell line in their study. For example, a clinical trial might involve extracting a small piece of breast cancer tumor from each of 100 patients who responded to a given chemotherapy treatment and from another 100 patients who did not. Cells from each patient’s tumor would be hybridized to a microarray chip, then the investigators would compare the relative levels of RNA expression across all the patients to identify a set of genes with expression levels that with chemotherapy response. This high-throughput dataset would include at least a million data points. The dataset is expensive and time-consuming to collect, but very valuable not only to the original investigators for their original purpose but also to other investigators who may wish to study different questions.

Microarray data provide a useful environment for exploring data sharing policies and behaviors, for several reasons. Despite being valuable for reuse, microarray data are often but not yet universally shared. The best-practice guidelines for sharing microarray data are fairly mature, including standards for formatting and minimum-inclusion reporting developed by the active Microarray and Gene Expression Data (MGED) Society. A few centralized databases have emerged as best-practice repositories: the Gene Expression Omnibus (GEO)[10] and ArrayExpress[11]. Several high-profile letters have called for strong data sharing policies[5, 6]. Finally, the National Center for Biotechnology Information's Entrez website (<http://www.ncbi.nlm.nih.gov/>) makes it easy to identify journal articles that have submitted datasets to GEO, allowing us to study the association between journal policies and observed data sharing practice.

We identified journals with more than 15 articles published on "gene expression profiling" in 2006, using Thomson's Journal Citation Reports. We extracted the journal impact factors, subdiscipline categories, and

publishing organizations. We looked up each journal in The Directory of Open Access Journals to determine which are based on an open-access publishing model.

We used Google to locate the Instructions for Author policies for each of the journals. We manually downloaded and reviewed each policy for all mentions of data sharing.

2.2 Classifying the relative strength of the data sharing policies

We classified each of the policies into one of three categories: no mention of sharing microarray data, a relatively weak data sharing policy, or a strong policy. We defined a weak policy as one that is unenforceable, echoing McCain's terminology.[4] This included policies that merely suggest or request that microarray data be shared, as well as policies that require sharing but fail to require evidence that data has been shared. Strong policies, in contrast, require microarray data to be shared and insist upon a database accession number as a condition of publication.

We conducted univariate and linear multivariate regressions to understand the relationship between the strength of data sharing policy and journal impact factor, journal subdiscipline, journal publisher (academic societies vs. commercial), and publishing model (open vs. closed access).

2.3 Measuring the frequency with which authors share their data

To make a preliminary estimate of data sharing prevalence, we began by querying PubMed for journal articles published in 2006 or 2007 that were likely to have generated gene expression microarray data. These articles form the denominator of our prevalence estimate, so ideally only studies that produced raw data – articles with potentially shareable data – would be included. Unfortunately, PubMed does not provide a straightforward way to accurately identify only studies that produced their own data; a PubMed query for articles about gene expression microarray data (“*Gene Expression Profiling*”[MeSH] AND “*Oligonucleotide Array Sequence Analysis*”[MeSH]) returns not only studies that produced their own data, but also studies that strictly reused previous datasets (and therefore don't have their own raw microarray data to share) and even articles about new tools for storing and analyzing gene expression microarray data. A more accurate retrieval of data-producing studies would require access to the article's full text, and was beyond the scope of this paper.

Nonetheless, if we assume that articles about data reuse and tools occur in journals independently of the journal's data sharing policy, we can use the rough PubMed query to provide a preliminary estimate of relative prevalence. It is crucial, however, that we interpret these estimates relative to one another and not compare them to a theoretical ideal of 100%. Since the denominator of our percentages is not “number of papers that produced microarray data and could have shared it” but rather “number of papers about microarrays,” even if all studies that produced data in fact shared it our estimates would still be less than 100%.

Using the NCBI's Entrez website, for each journal in our cohort, we counted the total number of articles returned by our PubMed query and the percentage of those articles that had links to the GEO data repository. We conducted univariate and linear multivariate regressions over the journal data-sharing prevalence percentages to understand if strength of data sharing policy was associated with observed data sharing prevalence, including covariates for journal impact factor, journal subdiscipline, publisher type, and publishing model.

3. Results

3.1 Journal's policies on sharing data

Seventy journals met the selection criteria, spanning a wide range of impact factors (0.9 to 30.0, median: 4.5). A minority are published by academic societies (22). Only 5 use an open-access publishing model. Thomson's Journal Citation Reports identified 27 subdisciplines covered by these journals. We retained the categories with more than five members: Biochemistry and Molecular Biology (19), Biotechnology and Applied Microbiology (11), Cell Biology (11), Genetics and Heredity (11), Oncology (19), and Plant Sciences (7). We also retained Multidisciplinary Sciences (n=4) because we were curious about the policies for high-profile journals such as *Nature* and *Science*.

Of the 70 journal policies, 30 (43%) had no policy applicable to microarrays. This included 17 journals that make no mention of sharing publication-related data within their Instruction to Author statements, and 13 journal policies that request or require the sharing of non-microarray types of data (usually DNA and protein sequences), but no statement covering data in general or microarray data in particular.

The remaining 40 journals had a policy applicable to microarrays. We classified 17 of the microarray-applicable policies as relatively weak and 23 as strong, as detailed in Table 1.

No Policy	Weak Policy	Strong (Enforceable) Policy
<i>Acta Biochimica Et Biophysica Sinica</i>	<i>Bioinformatics</i>	<i>Applied And Environmental Microbiology</i>
<i>Annals Of The New York Academy Of Sciences</i>	<i>BMC Bioinformatics</i>	<i>Blood</i>
<i>Biochemical And Biophysical Research Communications</i>	<i>BMC Cancer</i>	<i>Cancer Research</i>
<i>British Journal Of Cancer</i>	<i>BMC Genomics</i>	<i>Cell</i>
<i>Cancer Letters</i>	<i>Breast Cancer Research</i>	<i>Clinical Cancer Research</i>
<i>Carcinogenesis</i>	<i>FASEB Journal</i>	<i>Developmental Biology</i>
<i>Experimental Cell Research</i>	<i>Genome Biology</i>	<i>FEBS Letters</i>
<i>Frontiers In Bioscience</i>	<i>Genome Research</i>	<i>Gene Expression Patterns</i>
<i>Gene</i>	<i>International Journal Of Cancer</i>	<i>Infection And Immunity</i>
<i>Genes Chromosomes & Cancer</i>	<i>Molecular Endocrinology</i>	<i>Journal Of Bacteriology</i>
<i>Genomics</i>	<i>Physiological Genomics</i>	<i>Journal Of Biological Chemistry</i>
<i>Human Molecular Genetics</i>	<i>Plant Journal</i>	<i>Journal Of Experimental Botany</i>
<i>IEEE-ACM Transactions On Computational Biology And Bioinformatics</i>	<i>Plant Physiology</i>	<i>Journal Of Immunology</i>
<i>International Journal Of Molecular Medicine</i>	<i>Proteomics</i>	<i>Journal Of Pathology</i>
<i>International Journal Of Oncology</i>	<i>Stem Cells</i>	<i>Journal Of Virology</i>
<i>Journal Of Clinical Oncology</i>	<i>Toxicological Sciences</i>	<i>Molecular Cancer Therapeutics</i>
<i>Journal Of Leukocyte Biology</i>	<i>Virology</i>	<i>Molecular And Cellular Biology</i>
<i>Journal Of Neurochemistry</i>		<i>Nature Biotechnology</i>
<i>Leukemia Research</i>		<i>Nature</i>
<i>Leukemia</i>		<i>Nucleic Acids Research</i>
<i>Mammalian Genome</i>		<i>Plant Cell</i>
<i>Microbes And Infection</i>		<i>Proceedings Of The National Academy Of Sciences Of The USA (PNAS)</i>
<i>Molecular Immunology</i>		<i>Science</i>
<i>Molecular Plant-Microbe Interactions</i>		
<i>Oncogene</i>		
<i>Oncology Reports</i>		
<i>Pharmacogenomics</i>		
<i>Plant Molecular Biology</i>		
<i>Planta</i>		

Table 1: Classification of journal data-sharing policies for gene expression microarray data

The policies varied widely across a number of dimensions. We explore several of these dimensions below, using excerpts from the policies.

3.1.1 *Statements of policy motivation*

Several journals introduce their policies with a motivation for sharing data. These statements explain the anticipated benefits to the scientific community, the intended service to readers, or the principles of the journal. Examples are given in Table 2.

Journal	Excerpt from Instructions to Authors: motivation for data sharing policy
<i>Stem Cells, Blood</i> (similar statement)	Stem Cells supports the efforts of the National Academy of Sciences (NAS) to encourage the open sharing of publication-related data. Stem Cells adheres to the beliefs that authors should include in their publications the data, algorithms, or other information that is central or integral to the publication, or make it freely and readily accessible; use public repositories for data whenever possible; and make patented material available under a license for research use.
<i>Bioinformatics</i>	Bioinformatics fully supports the recommendations of the National Academies regarding data sharing.
<i>Genome Research</i>	Genome Research encourages all data producers to make their data as freely accessible as possible prior to publication. Open data resources accompanied by fair use will serve to greatly enhance the scientific quality of work by the entire community and for society at large.
<i>Plant Cell</i>	The purpose of this policy is to ensure that conclusions are scientifically sound
<i>Physiological Genomics</i>	Work published in the APS Journals must necessarily be independently verifiable [...] Within a short time span, microarrays have become an important, commonly used tool in molecular genetics and physiology research. For microarray analysis of gene expression to have any long-term impact, it is crucial that the issue of reproducibility be adequately addressed.
<i>Proceedings Of The National Academy Of Sciences Of The USA</i>	To allow others to replicate and build on work published in PNAS, authors must make materials, data, and associated protocols available to readers
<i>Science</i>	After publication, all data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of Science.
<i>Journal Of Biological Chemistry</i>	... will substantially enhance an author's ability to communicate important research information and will also greatly benefit readers.

Table 2: Selected excerpts to illustrate the variety of data-sharing policy motivations

In addition, 22 policies included general-purpose sharing statements, thereby implying their support for the principle of data sharing. An example from *Bioinformatics*:

All data on which the conclusions given in the publication are based must be publicly available.

From *BMC Bioinformatics*:

Submission of a manuscript to BMC Bioinformatics implies that readily reproducible materials described in the manuscript, including all relevant raw data, will be freely available to any scientist wishing to use them for non-commercial purposes.

3.1.2 Datatype-specific policies

The journals with general data-sharing policies almost always supplement this with additional instructions for certain datatypes. In fact, many policies only have policies for certain datatypes and not for data sharing in general.

The policies for depositing nucleotide sequences are usually more strict than policies for other datatypes, including gene expression microarray data. The *FASEB Journal*, in contrast, explicitly treats all datatypes the same:

The FASEB Journal also does not distinguish between microarray data and other sorts of data (proteomics, sequence data, organic syntheses, crystal structures, etc.) All methods must be publicly available and described. Anything published in The FASEB Journal must have all data available not only for review but to every reader, electronic or print.

3.1.3 *Sharing requested or required*

Most journals with a policy for sharing microarray data state it as a requirement, using phrases like *must*, *required*, and *as a condition of publication*. A few policies (n=4) are less strict, stating their policies as requests through the words *should*, *recommend*, and *request*.

3.1.4 *Data location*

Most policies state that microarray data must be made available in a public database. A few are less specific, stating that sharing via public webpages or supplementary journal information is sufficient, or the policy leaves location unspecified. Some policies are more specific, insisting that the database be of a certain standard. *Plant Cell*, for example, specifies a permanent public database. *Plant Physiology* expands on this theme:

Links to web sites other than a permanent public repository are not an acceptable alternative because they are not permanent archives.

Two databases, GEO and ArrayExpress, are the predominant centralized storage locations for microarray datasets. Many of the policies suggest that data be deposited into one of these two locations, and a few policies limit the choice to one of these centralized options.

3.1.5 *Data format*

None of the policies explicitly specified a data format. By recommending or requiring submission to one of the permanent public databases, the journals implicitly stipulate the standard formats used within those databases.

3.1.6 *Data completeness*

The Microarray and Gene Expression Data (MGED) Society has developed guidelines for the Minimum Information About a Microarray Experiment (MIAME) that is “needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment.”[12] Because the experimental conditions for collecting microarray data can be very complex, these MIAME guidelines are very helpful for both data sharers and data reusers. *Physiological Genomics* includes rationale for adopting the MIAME guidelines within their instruction for authors statement:

Within a short time span, microarrays have become an important, commonly used tool in molecular genetics and physiology research. For microarray analysis of gene expression to have any long-term impact, it is crucial that the issue of reproducibility be adequately addressed. In addition, since microarray analytic standards are certain to change, it is crucial that authors identify the nature of the experimental conditions prevalent at the time of their research. If today’s research is to be relevant tomorrow, the core elements that are immune to obsolescence must be made clear. The APS Journals are adopting the MIAME standards to ensure that what is cutting edge today is not obsolete few years later.

More than 30 of the data-sharing policies recommend that data be compliant with the MIAME guidelines. As an example of one of the strictest policies, *Gene Expression Patterns* requires adherence to the MIAME standards and even asks for a completed MIAME checklist to be submitted with the manuscript:

Authors submitting manuscripts relying on microarray or similar screens must supply the data as Supplementary data [...] at the time of submission, along with the completed MIAME checklist. The data must be MIAME-compliant and supplied in a form that is widely accessible.

3.1.7 *Timeliness of public availability*

A few policies specify that microarray data must be available to the public upon publication. None of the policies explicitly allow data to be withheld until a date after publication.

3.1.8 *Consequences for not sharing data*

Several policies stipulate consequences for authors who fail to comply with journal conditions, as listed in Table 3. No weak policies included consequences, even though weak policies would benefit most since their requirements are the least enforceable prior to publication.

Journal	Excerpt from Instructions to Authors: consequences for NOT sharing data
<i>Applied And Environmental Microbiology,</i> <i>Infection And Immunity,</i> <i>Journal Of Bacteriology,</i> <i>Journal Of Virology,</i> <i>Molecular And Cellular Biology</i>	Failure to comply with the policies described in these Instructions may result in a letter of reprimand, a suspension of publishing privileges in ASM journals, and/or notification of the authors’ institutions.

<i>Nucleic Acids Research</i>	The Editors are prepared to deny further publication rights in the Journal to authors unwilling to abide by these principles.
<i>Mammalian Genome</i>	Failure to comply with this policy may result in exclusion from publication in <i>Mammalian Genome</i> .
<i>Nature</i> , <i>Nature Biotechnology</i>	After publication, readers who encounter a persistent refusal by the authors to comply with these guidelines should contact the chief editor of the <i>Nature</i> journal concerned, with "materials complaint" and publication reference of the article as part of the subject line. In cases where editors are unable to resolve a complaint, the journal reserves the right to refer the correspondence to the author's funding institution and/or to publish a statement of formal correction, linked to the publication, that readers have been unable to obtain necessary materials or reagents to replicate the findings.

Table 3: Selected excerpts of consequences for noncompliance with data-sharing journal policies

Although only tangentially related to dataset sharing, it is interesting to note the tough stance that some journals are willing to take when authors refuse to share their biological reagents after publication. From *Blood*:

Although the Editors appreciate that many of the reagents mentioned in Blood are proprietary or unique, neither condition is considered adequate grounds for deviation from this policy. ... if a reasonable request is turned down and not submitted to the Editor-in-Chief, the corresponding author will be held accountable. The consequence for noncompliance is simple: the corresponding author will not publish in Blood for the following 3 years.

From *PNAS*:

Authors must make Unique Materials (e.g., cloned DNAs; antibodies; bacterial, animal, or plant cells; viruses; and computer programs) promptly available on request by qualified researchers for their own use. Failure to comply will preclude future publication in the journal... Contact pnas@nas.edu if you have difficulty obtaining materials.

3.1.9 Exceptions to data sharing policies

At least one journal, *Genome Research*, explicitly disallows any exceptions to their principle of public data sharing. In contrast, a few other journals state or imply that they are willing to be flexible in some circumstances. Relevant excerpts are included in Table 4.

Journal	Excerpt from Instructions to Authors: forbidding exceptions to data sharing policies
<i>Genome Research</i>	Genome Research will NOT consider manuscripts where data used in the paper is not freely available on either a publicly held Web site or, in the absence of such a Web site, on the Genome Research Web site. There are NO exceptions.
Journal	Excerpt from Instructions to Authors: permitting exceptions to data sharing policies
<i>Proceedings Of The National Academy Of Sciences Of The USA</i>	Authors must disclose upon submission of the manuscript any restrictions on the availability of materials or information.
<i>Developmental Biology, Gene Expression Patterns</i>	The editors understand that on occasion authors may not feel it appropriate to deposit the entire data set at the time of publication of this paper. We are therefore willing to consider exceptions to this requirement in response to a request from the authors, which must be made at the time of initial submission or as part of an informal pre-submission enquiry
<i>Science</i>	We recognize that discipline-specific conventions or special circumstances may occasionally apply, and we will consider these in negotiating compliance with requests. Any concerns about your ability to meet Science's requirements must be disclosed and discussed with an editor.

Table 4: Selected excerpts to illustrate forbidden and permitted exceptions from data-sharing policies

3.2 The relative strength of the data sharing policies

Based on univariate analysis, data sharing policy strength was associated with impact factor. As seen in Figure 1, the journals with no data sharing policy, a weak policy, and a strong policy had respective median impact factors of 3.6, 4.9, and 6.2.

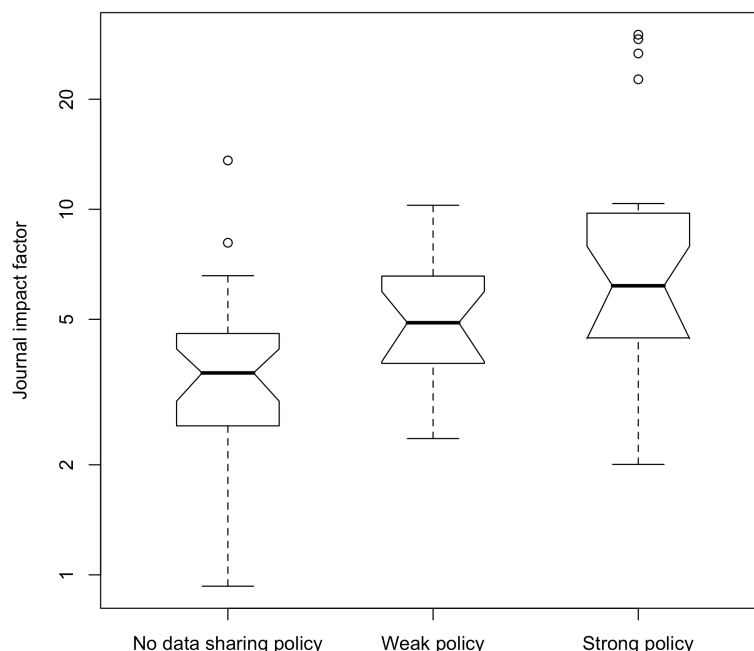


Figure 1: A boxplot of the impact factors for each journal, grouped by the strength of the journal's data-sharing policy. For each group, the heavy line indicates the median, the box encompasses the interquartile range (IQR, 25th to 75th percentiles), the whiskers extend to datapoints within 1.5xIQR from the box, and the notches approximate the 95% confidence interval of the median

Data sharing policy was also associated with journal publisher: 46% of commercial publishers had a data sharing policy, compared to 82% of journals published by an academic society. All five of the open-access journals had a policy.

In multivariate analysis, we found that the following variables were positively associated with the existence of a microarray data sharing policy: impact factor, open access, and academic society publishing. In contrast, the subdisciplines of Biochemistry&Molecular Biology and Oncology were negatively associated with the existence of a microarray data sharing policy. Details including all the covariates are provided in Table 5.

Journal Attribute	Estimate	p-value
Impact Factor, natural log	0.34	<0.001 ***
Open Access	0.63	0.002 **
Published by Association	0.23	0.046 *
Biochemistry & Molecular Biology	-0.28	0.031 *
Biotechnology & Applied Microbiology	0.04	0.784
Plant Sciences	-0.08	0.636
Oncology	-0.37	0.004 **
Cell Biology	0.10	0.485
Genetics & Heredity	-0.11	0.456
Multidisciplinary Sciences	-0.29	0.207

Table 5: Results of linear multivariate regression over the existence of a journal's data-sharing policy

3.3 The frequency with which authors share their data

Journals with the strongest data sharing policies had the highest proportion of papers with shared datasets. As seen in Figure 2, the journals with no data sharing policy, a weak policy, and a strong policy had a median data sharing prevalence of 8%, 20%, and 25% respectively. As mentioned in the Methodology section, these proportions should be interpreted relative to each other rather than to a theoretical maximum of 100%.

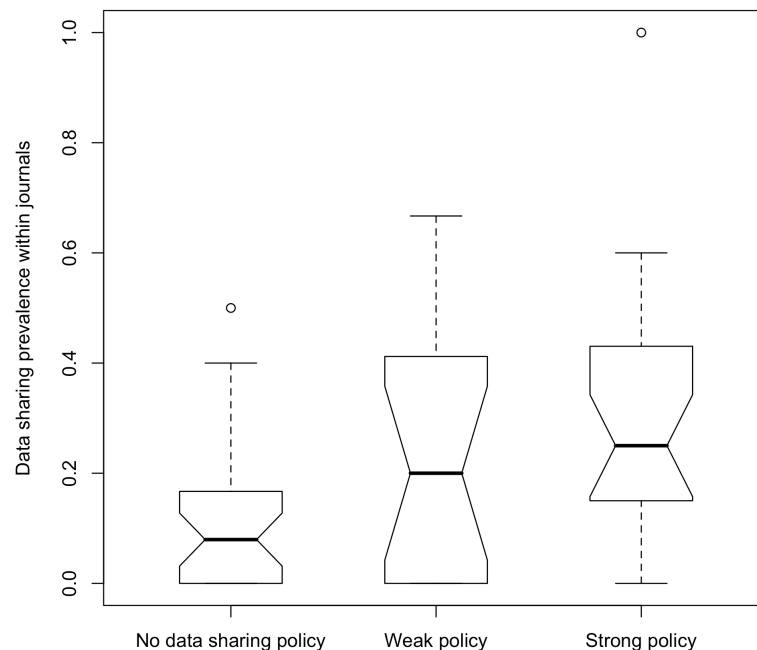


Figure 2: A boxplot of the relative data-sharing prevalence for each journal, grouped by the strength of the journal's data-sharing policy. For each group, the heavy line indicates the median, the box encompasses the interquartile range (IQR, 25th to 75th percentiles), the whiskers extend to datapoints within 1.5xIQR from the box, and the notches approximate the 95% confidence interval of the median

Based on multivariate analysis, we found that articles were more likely to have submitted primary data to GEO when they were published in journals with a data sharing policy, published by an academic society, or in the subdisciplines of Genetics&Heredity or Multidisciplinary Sciences. Details are given in Table 6.

Journal Attribute	Estimate	p-value
Has a Data Sharing Policy	0.11	0.037 *
Impact Factor, natural log	0.06	0.118
Open Access	-0.07	0.386
Published by Association	0.15	0.002 **
Biochemistry & Molecular Biology	0.01	0.850
Biotechnology & Applied Microbiology	-0.01	0.866
Plant Sciences	0.08	0.232
Oncology	0.02	0.737
Cell Biology	0.04	0.475
Genetics & Heredity	0.27	<0.001 ***
Multidisciplinary Sciences	0.28	0.004 **

Table 6: Results of linear multivariate regression over the prevalence with which the articles in a journal submit their microarray data to a centralized database

4. Discussion

We found wide variation amongst journal policies on data sharing, even for a data type with well-defined reporting standards and centralized repositories. Journals with a high impact factor, an open access publishing model, and a non-commercial publisher were most likely to have a data-sharing policy. This could be expected, as journals with a high impact factor are able to stipulate conditions to ensure research is of the highest quality without eroding their appeal, open-access journals are often particularly advocates for all aspects of open scholarship, and journals published by academic societies have previously been found to endorse data sharing more readily than commercial journals.[4] Surprisingly, our study did not identify any subdisciplines with an unusually-high number of data sharing policies. In contrast, we found that Oncology journals and Biochemistry&Molecular Biology journals were relatively unlikely to have a data sharing policy. The Oncology result is consistent with our observation that medical journals have been slower to embrace new publishing paradigms and open scholarship principles than journals within biology and bioinformatics. This is unfortunate, since cancer microarray data holds particular promise and is often especially expensive and time-consuming to collect. It is also unnecessary, since microarray data can be (and is) shared without compromising patient privacy.

We found that the existence of a data sharing policy was associated with an increase in data sharing behavior. A non-commercial publisher and the subdisciplines of Genetics&Heredity and Multidisciplinary Sciences were also significantly associated with a relatively high frequency of dataset submissions into the GEO database, as a percentage of all published gene expression papers. Studies of Genetics&Heredity often reuse data, so perhaps authors in that field are well acquainted with the value of sharing data. Interestingly, the two subdisciplines that were negatively associated with the existence of a data sharing policy were not less likely than usual to share their data when other factors are held constant. We were surprised that impact factor was not strongly associated with data sharing prevalence in multivariate analysis, because we suspect that well-funded and high-profile studies are under more pressure to share their data. In the future, we'd like to include variables about funding in these analyses.

A large number of journals had a policy for microarray data but not data in general. This probably reflects the success of MGED's efforts in actively encouraging and supporting microarray data exchange. As such, the results we have found are illuminating but may not be representative for other datatypes with a less mature infrastructure. A study by Brown[13] in 2000 used several methods to investigate the adoption and usage of Genbank, one of the most mature and successful biological databases. She tracked changes in instruction to author statements across 23 journals over 20 years, and noted that the data sharing policies for sequences have become stronger over time. As she explains, the authors who published in the *Journal of Biological Chemistry* were urged to deposit sequence data into Genbank in 1984, told they "should" deposit data in 1985, and were required to submit data as a condition of publication by 1991. It would be interesting to study whether, as the microarray field continues to mature, the journals we consider to have weak data sharing policies will evolve stronger policies with time.

Journals ought to give careful consideration to changing their policies[14]. Although there may be direct benefits to journals when authors must share their raw research data (reducing fraud, encouraging more careful research), data sharing mandates are controversial.[15] It is possible new mandates may cause authors to shop for an alternative publishing venue to avoid hassle. To measure the acceptance of a policy change, the editorial team at *Physiological Genomics* surveyed their authors and reviewers two years after instituting a data sharing requirement. They found that the vast majority of authors (92%) believed depositing microarray data was of significant value to the scientific community, and "67% of those who responded said they did not find the deposit of microarray data into GEO to be an obstacle to submission or review of articles".[16] Database tools have evolved since that survey, and submitting data continues to get easier.

Nonetheless, there are many personal difficulties for those who undertake to share their data, resulting in a variety of reasons why investigators may choose to withhold it. First, sharing data is often time-consuming: the data have to be formatted, documented, and uploaded. Second, releasing data can induce fear. There is a possibility that the original conclusions may be challenged by a re-analysis, whether due to possible errors in the original study, a misunderstanding or misinterpretation of the data, or simply more refined analysis methods. Future data miners might discover additional relationships in the data, some of which could disrupt the planned research agenda of the original investigators. Investigators may fear they will be deluged with requests for assistance, or need to spend time reviewing and possibly rebutting future re-analyses. They might feel that sharing data decreases their own competitive advantage, whether future publishing opportunities, information

trade-in-kind offers with other labs, or potentially profit-making intellectual property. Finally, it can be complicated to release data. If not well-managed, data can become disorganized and lost. Some informed consent agreements may not obviously cover subsequent uses of data. De-identification can be complex. Study sponsors, particularly from industry, may not agree to release raw detailed information, or data sources may be copyrighted such that the data subsets can not be freely shared.

Given all of these hurdles, it is natural that authors may need extra encouragement to share their data. We suggest that journal editors take a few simple steps to increase adherence to data sharing policies and thus bring about a more open scholarship. First, journals that already mandate data sharing should require the inclusion of an accession number (or web address for datatypes without databases) upon submission, since “prepublication compliance is much easier to monitor and enforce than postpublication compliance”[4] Second, journals should instruct their editors and reviewers to confirm that accession numbers are included in the manuscripts, as some journals do for their clinical trial reporting policies[17]. Third, journals should require that authors complete a MIAME checklist to increase the likelihood that shared data is complete and well-annotated, following the example of *Gene Expression Patterns*. To take this step further, journals could contract with a service like the one offered by ArrayExpress[18] to verify that submitted datasets meet a threshold of annotation quality. Fourth, journals need to implement their consequences: don’t publish papers that don’t uphold the policies.

Finally, during this cultural transition, we recommend that journals support measures that recognize and reward investigators who share data.[19] For example, journals could educate authors and reviewers on responsible data reuse and acknowledgement practices, either as part of instructions to authors statements or in editorials (see *Nature* journals [20, 21, 22]) Acknowledging data sources in a machine-readable way (through references, urls, and accession numbers) will allow the benefits of data reuse to be automatically linked back to the original data producers through citation counts[23] or other usage metrics, and thus provide a positive motivation for sharing data. Innovative attempts to provide microattribution or a data reuse registry may offer additional opportunities for journals to support these goals.[21, 22, 24]

Our study has several important limitations: we explored journal policies for only one type of data, our measured data sharing behavior predated the policy downloads, and the policy classifications were performed by only one investigator. Our method of measuring data sharing behavior captures many but not all articles that shared data; we plan to use natural language processing techniques to find a wider variety of data sharing instances in the future[25]. Similarly, a full-text query to identify articles that produce primary, shareable data – perhaps using laboratory terms like *purify* and *hybridize* – could improve our preliminary estimates of data sharing prevalence. Finally, we note that the reported associations do not imply causation: we have not demonstrated that changing a journal’s data sharing policy will change the behavior of authors.

Nonetheless, we believe this review and analysis is an important step in understanding the relationship between journal policies and data sharing outcomes. Policies are implemented with the hopes of affecting change. It is often said, “You cannot manage what you do not measure.” We need to understand the motivation and impact of our various incentives and initiatives if we hope to unleash the benefits of widespread data sharing.

5. Acknowledgements

HP is supported by NLM training grant 5T15-LM007059-19 and WC is funded through NLM grant 1R01LM009427-01.

Raw data and statistical analysis code from this study are available at <http://www.dbmi.pitt.edu/piwowar/>.

6. References

- (1) MERTON R: The sociology of science: Theoretical and empirical investigations. 1973
- (2) GASS A: Open Access As Public Policy. *PLoS Biology* 2(10):e353, 2004
- (3) VICKERS A: Whose data set is it anyway? Sharing raw data from randomized trials. *Trials* 7:15, 2006
- (4) MCCAIN K: Mandating Sharing: Journal Policies in the Natural Sciences. *Science Communication* 16(4):403-431, 1995
- (5) BALL CA et al.: Standards for microarray data. *Science (New York, NY)* 298(5593)2002
- (6) BALL CA et al.: Submission of microarray data to public repositories. *PLoS Biol* 2(9)2004
- (7) CECH TR et al.: Sharing publication-related data and materials: responsibilities of authorship in the life sciences. *Plant physiology* 132(1):19-24, 2003
- (8) PANEL ON SCIENTIFIC RESPONSIBILITY AND THE CONDUCT OF RESEARCH: Responsible Science, Volume I: Ensuring the Integrity of the Research Process. 1992

- (9) Microarray standards at last. *Nature* 419(6905)2002
- (10) BARRETT T et al.: NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic Acids Res* 35(Database issue)2007
- (11) PARKINSON H et al.: ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* 35(Database issue)2007
- (12) BRAZMA A et al.: Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 29(4):365-371, 2001
- (13) BROWN C: The changing face of scientific discourse: Analysis of genomic and proteomic database usage and acceptance. *Journal of the American Society for Information Science and Technology* 54(10):926-938, 2003
- (14) Democratizing proteomics data. *Nat Biotech* 25(3):262-262, 2007
- (15) CAMPBELL P: Controversial Proposal on Public Access to Research Data Draws 10,000 Comments. *The Chronicle of Higher Education* A42, 1999
- (16) VENTURA B: Mandatory submission of microarray data to public repositories: how is it working? *Physiol Genomics* 20(2):153-156, 2005
- (17) HOPEWELL S et al.: Endorsement of the CONSORT Statement by high impact factor medical journals: a survey of journal editors and journal 'Instructions to Authors'. *Trials* 9:20, 2008
- (18) BRAZMA A, PARKINSON H: ArrayExpress service for reviewers/editors of DNA microarray papers. *Nature Biotechnology* 24(11):1321-1322, 2006
- (19) Got data? *Nat Neurosci* 10(8):931-931, 2007
- (20) SCHRIGER DL, ARORA S, ALTMAN DG: The content of medical journal Instructions for authors. *Ann Emerg Med* 48(6)2006
- (21) Human variome microattribution reviews. *Nat Genet* 40(1)2008
- (22) Compete, collaborate, compel. *Nat Genet* 39(8)2007
- (23) PIWOWAR HA, DAY RS, FRIDSMA DB: Sharing detailed research data is associated with increased citation rate. *PLoS ONE* 2(3)2007
- (24) PIWOWAR HA, CHAPMAN WW: Envisioning a Biomedical Data Reuse Registry. Blog post on March 24, 2008: <http://researchremix.wordpress.com/2008/03/24/envisioning-a-biomedical-data-reuse-registry/>
- (25) PIWOWAR HA, CHAPMAN WW: Identifying data sharing in the biomedical literature. AMIA Annual Symposium [submitted] 2008