FOUNDATIONAL STUDIES FOR MEASURING THE IMPACT, PREVALENCE, AND PATTERNS OF PUBLICLY SHARING BIOMEDICAL RESEARCH DATA

by

Heather Alyce Piwowar

Bachelor of Science in Electrical Engineering and Computer Science, MIT, 1995

Master of Engineering in Electrical Engineering and Computer Science, MIT, 1996
Master of Science in Biomedical Informatics, University of Pittsburgh, 2006

Submitted to the Graduate Faculty of

the School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2010

UNIVERSITY OF PITTSBURGH
SCHOOL OF MEDICINE

This dissertation was presented

by

Heather Alyce Piwowar

It was defended on
March 24, 2010
and approved by
Brian B. Butler, PhD, Associate Professor,
Katz Graduate School of Business, University of Pittsburgh

Ellen G. Detlefsen, PhD, Associate Professor,
School of Information Sciences, University of Pittsburgh

Gunther Eysenbach, MD, MPH, Associate Professor,
Department of Health Policy, Management and Evaluation, University of Toronto

Madhavi Ganapathiraju, PhD, Assistant Professor,
Department of Biomedical Informatics, University of Pittsburgh

Dissertation Advisor: Wendy W. Chapman, PhD, Assistant Professor,
Department of Biomedical Informatics, University of Pittsburgh

FOUNDATIONAL STUDIES FOR MEASURING THE IMPACT, PREVALENCE, AND PATTERNS OF PUBLICLY SHARING BIOMEDICAL RESEARCH DATA

Heather A. Piwowar, PhD
University of Pittsburgh, 2010

Many initiatives encourage research investigators to share their raw research datasets in hopes of increasing research efficiency and quality. Despite these investments of time and money, we do not have a firm grasp on the prevalence or patterns of data sharing and reuse. Previous survey methods for understanding data sharing patterns provide insight into investigator attitudes, but do not facilitate direct measurement of data sharing behaviour or its correlates. In this study, we evaluate and use bibliometric methods to understand the impact, prevalence, and patterns with which investigators publicly share their raw gene expression microarray datasets after study publication.

To begin, we analyzed the citation history of 85 clinical trials published between 1999 and 2003. Almost half of the trials had shared their microarray data publicly on the internet. Publicly available data was significantly (p=0.006) associated with a 69% increase in citations, independently of journal impact factor, date of publication, and author country of origin.

Digging deeper into data sharing patterns required methods for automatically identifying data creation and data sharing. We derived a full-text query to identify studies that generated gene expression microarray data. Issuing the query in PubMed Central®, Highwire Press, and Google Scholar found 56% of the data-creation studies in our gold standard, with 90% precision. Next, we established that searching ArrayExpress and the Gene Expression Omnibus databases for PubMed® article identifiers retrieved 77% of associated publicly-accessible datasets.

We used these methods to identify 11603 publications that created gene expression microarray data. Authors of at least 25% of these publications deposited their data in the predominant public databases. We collected a wide set of variables about these studies and derived 15 factors that describe their authorship, funding, institution, publication, and domain environments. In second-order analysis, authors with a history of sharing and reusing shared gene expression microarray data were most likely to share their data, and those studying human subjects and cancer were least likely to share.

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

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I am truly grateful I've had the opportunity to pursue my research passion in such a supportive environment for the last five years.

I thank my advisor Wendy Chapman. Wendy, you allowed me a great deal of independence and backed it up with unfailing encouragement, support, and feedback. I couldn't imagine a better advisor for my dissertation.

Thanks to Mike Gabrin and Sean McDonald for keeping us in Pittsburgh, introducing me to biomedicine, and continuing to inspire my efforts to make a difference in the real world. I am thankful for time with the late Sam Wieand: he will forever be a role model to me in biostatistics and beyond.

Thanks to Roger Day and Jim Lyons-Weiler for conducting challenging, though-provoking classes during my first semester at Pitt, thereby hooking me back in to academic life. Thanks to Doug Fridsma for early discussions, Greg Cooper for always providing an insightful comment on my work, and Brian Chapman for articulately framing many of my messy thoughts. I thank my committee for their contributions, Toni Porterfield for making paperwork hassles disappear, and fellow students for camaraderie. I'm also very grateful to the support and flexibility of DBMI in allowing me to bring a sleeping kiddo to colloquium for months on end, and come and go as life required. You are a friendly and supportive department, and I will miss you.

I thank the NLM for the biomedical informatics training fellowship (5T15-LM007059), and especially for their monetary recognition of previous work experience in computer science and IT. I wouldn't have initiated a research career had it not been for this support.

I'm grateful to Todd Vision for initiating contact that has led to my next career step: a postdoc position studying data sharing. Finishing was much easier because I had something fantastic to look forward to.

I offer a personal thanks to the giants who built tools and performed research that made my work possible, though you are too many to name. Special thanks to those who release their research and creative output openly: Flickr photos with Creative Commons licenses, open source code (including blog snippets!), and open access articles. I thank scientists who share their data. It is hard to do. Thank you.

I am grateful to everyone who organized workshops, conferences, and symposia where I presented preliminary and tangential work: the NLM training conference, AMIA, ISMB, ELPUB, JCDL, PSB, ASIS&T. These opportunities gave me experience, exposure, confidence, and valuable feedback. In particular I thank those who put extra effort into organizing doctoral consortiums, student awards, and special tracks in Open Science.

Thanks to the open science community itself. You are inspirational, affirming, helpful, and make me want to be my best self.

I send a shout-out to all of the caffeine and wifi-fueled “third spaces” and their friendly faces that facilitated my flextime life in Pittsburgh and Vancouver, and to all of my friends and relations who helped keep work in perspective and life fun.

Thanks to my Maple Ridge family. Mom, you are always interested, always make time, and demonstrate a can-do and must-do attitude. Dad, I enjoy coming to you for insightful advice, and relish your example of unabashed joy in single-minded focus. Robyn: your passion is matched only by your intellect, and I admire both more than I can say. Callum and Kris, you make a rich life look easy: I draw strength from your example.

My Scottdale family, you put a human face on the medical and teaching professions. You go after your dreams, and offer unwavering support and love to those around you. Thank you.

I save a place of honour for all of the caregivers in our lives. Grandparents! Also, the staff at UCDC and Escuelita, and particularly Niki, B, Christa, Katie, Rosi, Lorenza, Lisa: thank you so much for your warmth and care. Niki: even more thanks on top, because you helped our family navigate the early days, and made me feel good about being a brand new mom and a PhD student and both at the same time.

Finally, first, last, and always: John, for doing everything you did to make this happen.

I dedicate this work to two people:

Kira

without whom I’d never have started, and

John

without whom I’d never have finished.

1.0 INTRODUCTION

Many initiatives encourage research data sharing in hopes of increasing research efficiency and quality, but the effectiveness of these early initiatives is not well understood. Sharing and reusing scientific datasets have many potential benefits: in addition providing detail for original analyses, 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.

Eager to encourage the realization of such 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 rewards, prevalence or patterns of data sharing and reuse, the effectiveness of initiatives, or the costs, benefits, and impact of repurposing biomedical research data.

Studies examining current data sharing behavior would be useful in three ways. First, an estimate of the prevalence with which data is shared, either voluntarily or under mandate, would provide a valuable baseline for assessing future adoption and continued intervention. Second, analyses of current behavior will likely identify subfields (perhaps research areas with a particular disease or organism focus, or those in well funded research groups) with relatively high prevalence of data sharing; digging into these can illuminate valuable best practices. Third, the same analyses will likely reveal subareas in which researchers rarely share their research datasets. Future research could focus on these challenging areas, to understand their unique obstacles for data sharing and refine future initiatives accordingly. You can not manage what you do not measure: understanding the rewards, prevalence, and patterns of data sharing and withholding will facilitate effective refinement of data sharing initiatives to better address real-world needs.

1.1 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].

However, whereas the general research community benefits from shared data, much of the burden for sharing the data falls to the study investigator. A major cost is time: the data have to be formatted, documented, and released. Further, it is sometimes complicated to decide where to best publish data, since supplementary information and laboratory sites are transient [4-6]. Beyond a time investment, 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 [7], a misunderstanding or misinterpretation of the data [8], 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. Data sources may be copyrighted such that the data subsets cannot be freely shared.

Recognizing that these disincentives make the establishment of a voluntary data sharing culture unlikely without policy guidance, many initiatives actively encourage or require that investigators make their raw data available for other researchers. There is a well known adage inspired by William Thomson (Lord Kelvin) [9]: you cannot manage what you do not measure. For those with a goal of promoting responsible data sharing, it would be helpful to evaluate the effectiveness of requirements, recommendations, and tools. When data sharing is voluntary, insights could be gained by learning which datasets are shared, on what topics, by whom, and in what locations. When policies make data sharing mandatory, monitoring is useful to understand compliance and unexpected consequences.

Unfortunately, it is difficult to monitor data sharing because data can be shared in so many different ways. Previous assessments of data sharing have included manual curation [10-12] and investigator self-reporting [13]. These methods are only able to identify instances of data sharing and data withholding in a limited number of cases, and therefore are unable to support widespread inquiry into patterns of data sharing behavior. We hope this project supplements previous research to address these limitations.

1.1.1 The potential benefits of data sharing

Sharing information facilitates science. Reusing previously-collected data in new studies allows these valuable resources to contribute far beyond their original analysis [14]. 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.

Believing that that these benefits outweigh the costs of sharing research data, many initiatives actively encourage investigators to make their data available. Some journals require the submission of detailed biomedical data to publicly available databases as a condition of publication [15, 16]. Since 2003, the NIH has required a data sharing plan for all large funding grants and has more recently introduced stronger requirements for genome-wide association studies [17, 18]; other funders have similar policies. Several government whitepapers [14, 19] and high-profile editorials [19-25] call for responsible data sharing and reuse, large-scale collaborative science is providing the opportunity to share datasets within and outside of the original research projects [20, 21], and tools, standards, and databases are developed and maintained to facilitate data sharing and reuse.

1.1.2 Current data sharing practice: forces in support

As highlighted above, sharing research data has many potential benefits to society. Although sharing of data has always been an aspiration of the scientific enterprise, it has only been common in a few subdisciplines. Forces are now converging to make it an achievable and everyday practice.

Datasets are larger than they have ever been – and larger than any single team of scientists can analyze exhaustively. The ubiquitous sharing and reuse of DNA sequences in Genbank® has clearly demonstrated the power of openly shared data. Other high-throughput hypothesis-generating datasets, such as genome-wide association studies [17, 22], gene expression microarrays [23], proteomics mass spectra [24], and brain images [25] allow data to be repurposed to answer multiple research questions. Extensive datasets are also generated within the clinical setting, particularly through the adoption of electronic health records. Stakeholders have begun to develop recommendations and guidelines for the complex ethical, legal, and technical issues surrounding the reuse and sharing of health data beyond primary healthcare [26].

Research is increasingly performed within networks of multi-disciplinary teams. The NIH Roadmap [27] and other initiatives [21, 28-30] have recognized that significant scientific progress requires collaboration. Collaborations develop and adopt frameworks, standards, tools, and policies to share data among investigators. This work can facilitate sharing their data beyond the boundaries of the original research partners.

Today’s collaborative science on large datasets is performed within an extremely tight biomedical funding environment. Many funding agencies have instituted data-sharing policies [31], hoping to accelerate scientific progress while avoiding the cost of duplicative collection efforts. The NIH Data Sharing Policy, adopted in 2003, requires a data sharing plan for all research grants over $500K [17].   The NIH stipulates additional requirements for specific domains. For example, all funded genome-wide association studies (GWAS) are now expected to share their data in the centralized NCBI database, dbGaP [18, 22]. Complementing and extending these funding agency requirements, many biomedical journals require or recommend that data be shared as a condition of publication [15, 16, 19]. Some journals delineate the responsibilities in detail and include procedures for addressing data sharing noncompliance [16, 33].

Open, centralized databases such as Genbank, Uniprot, and the Gene Expression Omnibus have evolved into de facto homes for specific types of data [34]. Standards for minimum data inclusion and data formats have been developed for many types of datasets. The challenge of integrating datasets has spurred research progress on ontologies and semantic description methods. Projects such as NCBI’s Entrez database suite [35], the Semantic Web for Life Sciences [36], the National Center for

Biomedical Ontology's Bioportal framework [37], and caBIG [29, 38, 39] provide visions for the future of research when data is more universally available and interoperable.

Data sharing and integration are being actively pursued outside of biomedical research, in other scientific fields (physics, astronomy, environmental science) and also by the general public [40]. Several websites encourage uploading and visualizing all sorts of data: the "Tasty Data Goodies" at Swivel (http://www.swivel.com) and IBM's Many Eyes (http://www.many-eyes.com) are popular examples. Widespread adoption of Web 2.0 technologies, including blogging, tagging, wikis, and mashups, suggest that our next generation of scientists will expect and embrace a world of research remixes [40].

Finally, I note the complementary forces of open access and pre-print publications, open notebook science projects [41], open source code [42], Creative Commons copyright licenses (http://creativecommons.org/) for many kinds of original content (including data), and two recent public access policies. The NIH Public Access Policy requires all NIH-funded investigators to submit their peer-reviewed manuscripts to PubMed Central to ensure public access, as of April 2008 [43]. In February 2008, the faculty of Harvard University voted to make all faculty scholarly publications freely available in an online open-access repository [44], the first such resolution by a university in the United States. While these policies do not apply to data beyond that provided within the manuscripts, they clearly demonstrate a political will to support sharing research results "to help advance science and improve human health" (http://publicaccess.nih.gov) and "promote free and open access to significant, ongoing research" [44].

1.1.3 Current data sharing practice: forces in opposition

While many forces are converging to enhance our ability to share data, there are significant social, organizational, technical and legislative factors that may impede them.

Investigators may restrict access to data to maximize the professional and economic benefit that they accrue from data they generate, even though they also gain advantage by accessing data produced by others.

A review of genomic data sharing highlighted the complexity of stakeholder interests both for and against data sharing [45], beyond those of the original investigators. Study subjects may have personal interests in privacy and confidentiality that exceed their personal interests in contributing to new methods of detecting and treating disease. Academic health centers may view data sharing as a threat to intellectual property, with the potential to impede spin-offs and start-ups that bring revenue and act as incubators for future research. Industrial sponsorship may hinder plans for sharing data. Changes in the regulatory environment make the sharing of data more complex, and may necessitate more stringent oversight to ensure compliance and minimize risk. Finally, limitations imposed by specific technologies undermine the ability of a uniform approach to generalize across different data types and regulatory requirements.

It is often difficult to effectively incent and mandate data sharing. Mandates are often controversial [46-48] while requests and unenforced mandates are often ignored [49]. The effect of funder policies like the NIH Data Sharing Policy have not been systematically studied, but anecdotal evidence suggests that many researchers view funder policies as optional, since they data sharing plans are not considered as part of scientific evaluation and there are no penalties for noncompliance [50].

I believe that a critical element in balancing these opposing forces is a better understanding of current data sharing behavior, patterns, and predictors to be used for communicating and refining sharing best-practices.

1.2 PREVIOUS RESEARCH ON DATA SHARING BEHAVIOR

A few investigations into data sharing behavior and attitudes have initiated work in this area. Findings and outstanding challenges are highlighted below.

1.2.1 Measuring and modeling data sharing behavior

Most measurements of data sharing prevalence have manually searched for shared datasets across a subset of journals [10, 11, 49], or systematically contacted authors to ask for shared datasets [51]. These studies have found that data sharing levels are high (but less than 100%) in a few cases, but overall prevalence is low. For example, Ochsner et al. [10] found that despite the maturity of gene expression microarray data sharing infrastructure and multitude of funder and journal mandates, overall rates of sharing gene expression microarray data online is about 50%.

These analyses have not correlated their prevalence findings with other variables to detect patterns. Multivariate analyses have relied upon surveyed attitudes and intentions (described below), rather than measured characteristics.

1.2.2 Measuring and modeling data sharing attitudes and intentions

The largest body of knowledge about motivations and predictors for biomedical data sharing and withholding comes from Campbell and co-authors. They surveyed researchers, asking whether they have ever requested data and been denied, or themselves denied other researchers from access to data. Results indicated that participation in relationships with industry, mentors' discouragement of data sharing, negative past experience with data sharing, and male gender were associated with data withholding [13]. In another survey, among geneticists who said they intentionally withheld data related to their published work, 80% said it was too much effort to share the data, 64% said they withheld data to protect the ability of a junior team member to publish, and 53% withheld data to protect their own publishing opportunities [52].

Occasionally, the administrators of centralized data servers publish feedback surveys of their users. As an example, Ventura reports a survey of researchers who submitted and reviewed microarray studies in the Physiological Genomics journal after its mandatory data submission policy had been in place for two years. Almost all (92%) authors said that they believed depositing microarray data was of value to the scientific community and about half (55%) were aware of other researchers reusing data from the database [53].

In related research, the information science and management of information systems communities have developed several models of knowledge sharing. These models often use either case studies [54] or opinions and attitudes gathered through validated survey instruments ([44, 55-57], and many more). Studied domains include knowledge sharing within an organization, volunteering knowledge in open social networks, physician knowledge sharing in hospitals, participation in open source projects, academic contributions to institutional archives, and other related activities.

1.2.3 Identifying instances of data sharing

While surveys have provided insight into sharing and reuse behavior, other issues are best examined by studying the demonstrated behavior of scientists. Unfortunately, observed measurement of data behavior is difficult because of the complexity in identifying all episodes of data sharing and reuse. Although indications of sharing and reuse usually exist within a published research report, the descriptions are in unstructured free text and thus complex to extract.

Most studies of data sharing to date have used a manual review to identify shared datasets (e.g. [10, 11, 49]).

One automated approach for identifying data sharing behavior is to follow the "primary citation" field of database submission entries. Unfortunately, this is imperfect, since these references often missing when data is submitted prior to study publication. Populating the submission citation fields retrospectively requires intensive manual effort, as demonstrated by the recent Protein Data Bank remediation project [57, 58], and thus is not usually performed. No effective way exists to automatically retrieve and index data housed on personal or lab websites or journal supplementary information.

Related research has examined the degree to which data remains available after it has been shared. Multiple studies underscore the transience of supplementary information [5], website URLs [6], and corresponding author email addresses [44].

1.2.4 Evaluating the impact of data sharing policies

Despite many funder and journal policies requesting and requiring data sharing, the impact of these policies have only been measured in small and disparate studies. McCain manually categorized the journal “Instruction to Author” statements in 1995 [15]. A more recent manual review of gene sequence papers found that, despite requirements, up to 15% of articles did not submit their datasets to Genbank [11]. Analyses of reproducibility in the political science literature suggests that only actively enforced journal policies are effective [49].

Studying the impact of data sharing policies is difficult because policies are often confounded with other variables. If, for example, impact factor is positively correlated with a strong journal data sharing policy as well as a large research impact, it is difficult to distinguish the direction of causation. Evaluating data sharing policies would ideally involve a randomized controlled trial, but unfortunately this is impractical.

In related work, evaluations have been done to estimate the impact of reporting guidelines [59].

1.2.5 Estimating the costs and benefits of data sharing

Estimating the costs and benefits of data sharing would be challenging even with a comprehensive dataset of occurrences. A complete evaluation would require comparing projects that shared with other similar projects that did not, across a wide variety of variables including person-hours-till-completion, total project cost, received citations and their impact, the number and impact of future publications, promotion, success in future grant proposals, and general recognition and respect in the field.

Pienta [60] is currently investigating these questions with respect to social science research data and publications. Zimmerman [61] has studied the ways in which ecologists find and validate datasets to overcome the personal costs and risks of data reuse.

Examining variables for their benefits on research impact is a common theme within the field of bibliometrics. Research impact is usually approximated by citation metrics, despite their recognized limitations.

1.2.6 Related research fields

Evaluation of data sharing and reuse behavior is related to a number of other active research fields: code reusability in software engineering, motivation in open source projects, online knowledge sharing communities, and corporate knowledge sharing, tools for collaboration, evaluating research output, the sociological study of altruism, information retrieval, usage metrics, data standards, the semantic web, open access, and open notebook science.

1.3 RESEARCH DESIGN AND METHODS

The long-term goal of this research is to accelerate research progress by increasing effective data reuse through informed improvement of data sharing and reuse tools and policies. The objective of this research project is to examine the feasibility of evaluating data sharing behavior based on examination of the biomedical literature.

This research addressed the following specific aims:

1.3.1 Aim 1: Does sharing have benefit for those who share?

I investigated the association between sharing raw microarray data and subsequent citation rate of published studies. I used datasets generated by a small, relatively homogeneous set of cancer gene expression microarray clinical trials. Multivariate analysis was used to statistically controlling for potential confounders. The results of Aim 1 provided motivation for Aim 2 and preliminary work for Aim 3.

1.3.2 Aim 2: Can sharing and withholding be systematically measured?

Because the manual methods used to conduct Aim 1 did not scale to larger analyses, I investigated automatic methods for measuring data sharing and withholding behavior. First, articles that generated gene expression microarray data were identified using NLP on full-text research. Second, to assess whether the authors of these data-generating studies shared or withheld their data, I investigated using database submission citation links as evidence of data sharing. The results of Aim 2 were used to generate a dataset for Aim 3.

1.3.3 Aim 3: How often is data shared? What predicts sharing? How can we model sharing behavior?

First, I applied the classification systems described in Aim 2 to a wide spectrum of the biomedical literature to identify articles that generated gene expression microarray data and, subsequently, which of the articles that generated data also shared it. Then, for each of the articles, I collected and analyzed features related to the authors, their institutional and funding environment, the study itself, and the publishing mechanism. I used univariate and multivariate statistics to investigate which of these features are associated with dataset sharing. Finally, I used exploratory factor analysis to derive a model that could be used to explain data sharing decisions based on my measured variables.

1.4 RELATED RESEARCH APPLICATIONS OF METHODS

1.4.1 Citation analysis for adoption and impact of open science

Citation analysis has been used to assess several aspects of the adoption and impact of open science, particularly literature open access. Eysenbach [62] found that authors who chose to make their articles open access in the Proceedings of the National Academy of Sciences received more citations within the first year after publication, Wren [63] correlated journal impact factor with the adoption rate of author-shared reprints, and many others. Other research have used citations to see how scientists use each other's work [64] and the relative impact of various study designs [65].

Many authors study factors that underlie citation rate; these highlight important factors to include as potential confounders whenever a detailed citation analysis of a new variable [66, 67]. Ongoing research attempts to identify the best way to represent various issues such as author ambiguity [68], author productivity [69, 70], institutional environment [71], journal impact factor [72-76] and clean and comprehensive citation data [77].

Finally, several researchers have proposed methods for citations of data to make studying the issue of reuse easier in the long run, such as [43] and [78], and examined the extent to which citations are an accurate proxy for peer ratings of quality [79].

1.4.2 Natural language processing of the biomedical literature

Natural language processing of the biomedical literature is traditionally organized into information retrieval, entity recognition, information extraction, hypothesis generation, and heterogeneous integration [80]. Most work has been on abstracts, because they are free, easy to obtain, and in a standardized format from MEDLINE®. Unfortunately, a great deal of information resides only in article full text. The TREC Genomics 2006/2007 tasks opened up a selection of free text for Information Retrieval research, and the Open Access subset at PubMed Central is another homogeneous, free, easy subset to obtain. Consequently, more research is beginning to focus on full-text [81].

Most research has focused on the needs of biologists or curators [82], but starting to be some investigations into automated techniques to help find articles for review based on the text [91-94], identification of the relationships between citing and cited articles [83, 84], and analysis of the methods section to enumerate the diversity of wet lab method use [85].

Techniques vary depending on the task, but stemming, synonyms, and n-grams are a mainstay [86]. Query expansion to include all query aspects have also been shown to help [87]. The availability of full text articles in PMC, Google Scholar, and other portals is spurring new approaches [88].

Finally, NLP techniques applied to clinical text might be of informative. For example, Melton et al. [89] also faces the issue of identifying records based on snippets of full text, though in their case it is adverse reactions in clinical discharge summaries.

1.4.3 Regression and factor analysis for deriving and evaluating models of sharing behavior

Most models of sharing behavior are based on established surveys, and thus evaluate their models using confirmatory analysis [101-105]. However, a few research projects instead use linear regression, such as [13, 56, 90-92]. Siemsen et al. [93] compare a regression model to that derived from constraining factor analysis. Finally, several studies involve exploratory factor analysis [71, 94, 95].

1.5 OUTLINE OF THE DISSERTATION

This chapter has provided an introduction to the dissertation and its topic. Each aim is described separately as a self-contained research report including an introduction, methods, results, and discussion. Aim 1 is covered in Chapter 2, Aim 2 in Chapters 3 and 4, and Aim 3 in Chapter 5. An overall discussion of contributions, implications, and future work is provided in the final chapter.

2.0 AIM 1: SHARING DETAILED RESEARCH DATA IS ASSOCIATED WITH INCREASED CITATION RATE

Background

Sharing research data provides benefit to the general scientific community, but the benefit is less obvious for the investigator who makes his or her data available.

Principal Findings

We examined the citation history of 85 cancer microarray clinical trial publications with respect to the availability of their data. The 48% of trials with publicly available microarray data received 85% of the aggregate citations. Publicly available data was significantly (p = 0.006) associated with a 69% increase in citations, independently of journal impact factor, date of publication, and author country of origin using linear regression.

Significance

This correlation between publicly available data and increased literature impact may further motivate investigators to share their detailed research data.

2.1 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 [14]. 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.

Believing that that these benefits outweigh the costs of sharing research data, many initiatives actively encourage investigators to make their data available. Some journals, including the PLoS family, require the submission of detailed biomedical data to publicly available databases as a condition of publication [15, 96, 97]. Since 2003, the NIH has required a data sharing plan for all large funding grants. The growing open-access publishing movement will perhaps increase peer pressure to share data.

However, while the general research community benefits from shared data, much of the burden for sharing the data falls to the study investigator. Are there benefits for the investigators themselves?

A currency of value to many investigators is the number of times their publications are cited. Although limited as a proxy for the scientific contribution of a paper [98], citation counts are often used in research funding and promotion decisions and have even been assigned a salary-increase dollar value [99]. Boosting citation rate is thus is a potentially important motivator for publication authors.

In this study, we explored the relationship between the citation rate of a publication and whether its data was made publicly available. Using cancer microarray clinical trials, we addressed the following questions: Do trials which share their microarray data receive more citations? Is this true even within lower profile trials? What other data-sharing variables are associated with an increased citation rate? While this study is not able to investigate causation, quantifying associations is a valuable first step in understanding these relationships. Clinical microarray data provides a useful environment for the investigation: despite being valuable for reuse and extremely costly to collect, is not yet universally shared.

2.2 MATERIALS AND METHODS

2.2.1 Identification and Eligibility of Relevant Studies

We compared the citation impact of clinical trials which made their cancer microarray data publicly available to the citation impact of trials which did not. A systematic review by Ntzani and Ioannidis [100] identified clinical trials published between January 1999 and April 2003 which investigated correlations between microarray gene expression and human cancer outcomes and correlates. We adopted this set of 85 trials as the cohort of interest.

2.2.2 Data Extraction

We assessed whether each of these trials made its microarray data publicly available by examining a variety of publication and internet resources. Specifically, we looked for mention of Supplementary Information within the trial publication, searched the Stanford Microarray Database (SMD) [101], Gene Expression Omnibus (GEO) [102], ArrayExpress [103], CIBEX [104], and the NCI GeneExpression Data Portal (GEDP)(gedp.nci.nih.gov), investigated whether a data link was provided within Oncomine [105], and consulted the bibliography of data re-analyses. Microarray data release was not required by any journals within the timeframe of these trial publications. Some studies may make their data available upon individual request, but this adds a burden to the data user and so was not considered “publicly available” for the purposes of this study.

We attempted to determine the date data was made available through notations in the published paper itself and records within the WayBackMachine internet archive (www.archive.org/web/web.php). Inclusion in the WayBackMachine archive for a given date proves a resource was available, however, because archiving is not comprehensive, absence from the archive does not itself demonstrate a resource did not exist on that date.

The citation history for each trial was collected through the Thomson Scientific Institute for Scientific Information (ISI) Science Citation Index at the Web of Science Database (www.isinet.com). Only citations with a document type of ‘Article’ were considered, thus excluding citations by reviews, editorials, and other non-primary research papers.

For each trial, we also extracted the impact factor of the publishing journal (ISI Journal Citation Reports 2004), the date of publication, and the address of the authors from the ISI Web of Science. Trial size, clinical endpoint, and microarray platform were extracted from the Ntzani and Ioannidis review [100].

2.2.3 Analysis

The main analyses addressed the number of citations each trial received between January 2004 and December 2005. Because the pattern of citations rates is complex–changing not only with duration since publication but also with maturation of the general microarray field–a confirmatory analysis was performed using the number of citations each publication received within the first 24 months of its publication.

Although citation patterns covering a broad scope of literature types are left-skewed [106], we verified that citation rates within our relatively homogeneous cohort were roughly log-normal and thus used parametric statistics.

Multivariate linear regression was used to evaluate the association between the public availability of a trial's microarray data and number of citations (after log transformation) it received. The impact factor of the journal which published each trial, the date of publication, and the country of authors are known to correlate to citation rate [107], so these factors were included as covariates. Impact factor was log-transformed, date of publication was measured as months since January 1999, and author country was coded as 1 if any investigator has a US address and 0 otherwise.

Since seminal papers–often those published early in the history a field or in very high-impact journals–receive an unusually high number of citations, we performed a subset analysis to determine whether our results held when considering only those trials which were published after 2000 and in lower-impact (<25) journals.

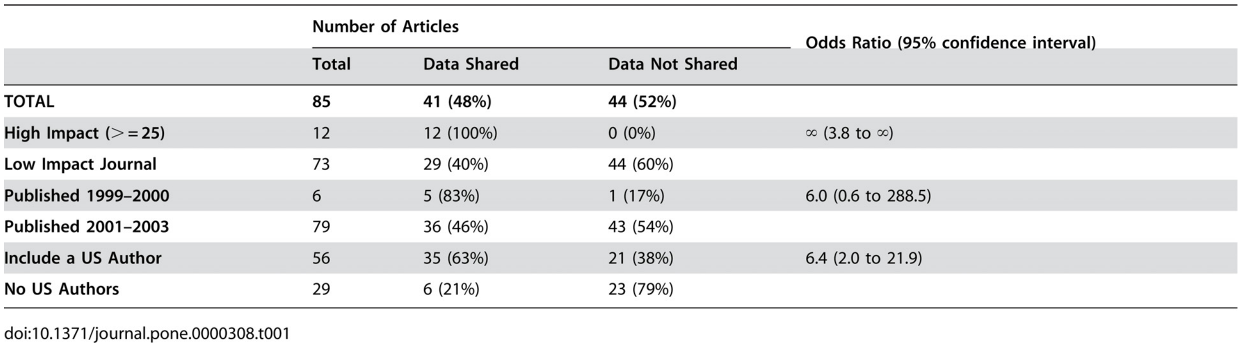
Finally, as exploratory analysis within the subset of all trials with publicly available microarray data, we looked at the linear regression relationships between additional covariates and citation count. Covariates included trial size, clinical endpoint, microarray platform, inclusion in various public databases, release of raw data, mention of supplementary information, and reference within the Oncomine [105] repository.

Statistical analysis was performed using the stats package in R version 2.1 [108]. P-values are two-tailed.

2.3 RESULTS

We studied the citations of 85 cancer microarray clinical trials published between January 1999 and April 2003, as identified in a systematic review by Ntzani and Ioannidis [100] and listed in Supplementary Text S1. We found 41 of the 85 clinical trials (48%) made their microarray data publicly available on the internet. Most data sets were located on lab websites (28), with a few found on publisher websites (4), or within public databases (6 in the Stanford Microarray Database (SMD) [101], 6 in Gene Expression Omnibus (GEO) [102], 2 in ArrayExpress [103], 2 in the NCI GeneExpression Data Portal (GEDP) (gedp.nci.nih.gov); some datasets in more than one location). The internet locations of the datasets are listed in Supplementary Text S2. The majority of datasets were made available concurrently with the trial publication, as illustrated within the WayBackMachine internet archives (www.archive.org/web/web.php) for 25 of the datasets and mention of supplementary data within the trial publication itself for 10 of the remaining 16 datasets. As seen in Table 1, trials published in high impact journals, prior to 2001, or with US authors were more likely to share their data.

Table 1: Characteristics of eligible publications



	Number of Articles			Odds Ratio (95% confidence interval)
	Total	Data Shared	Data Not Shared	
TOTAL	**85**	**41 (48%)**	**44 (52%)**	
High Impact (>=25)	12	12 (100%)	0 (0%)	∞ (3.8 to ∞)
Low Impact Journal	73	29 (40%)	44 (60%)	
Published 1999–2000	6	5 (83%)	1 (17%)	6.0 (0.6 to 288.5)
Published 2001–2003	79	36 (46%)	43 (54%)	
Include a US Author	56	35 (63%)	21 (38%)	6.4 (2.0 to 21.9)
No US Authors	29	6 (21%)	23 (79%)	

doi:10.1371/journal.pone.0000308.t001

The cohort of 85 trials was cited an aggregate of 6239 times in 2004–2005 by 3133 distinct articles (median of 1.0 cohort citation per article, range 1–23). The 48% of trials which shared their data received a total of 5334 citations (85% of aggregate), distributed as shown in Figure 1.

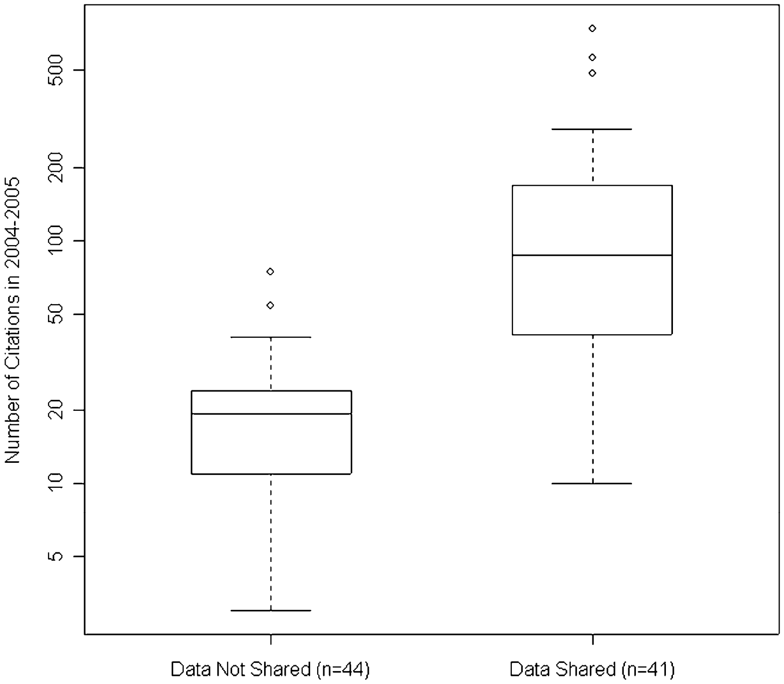
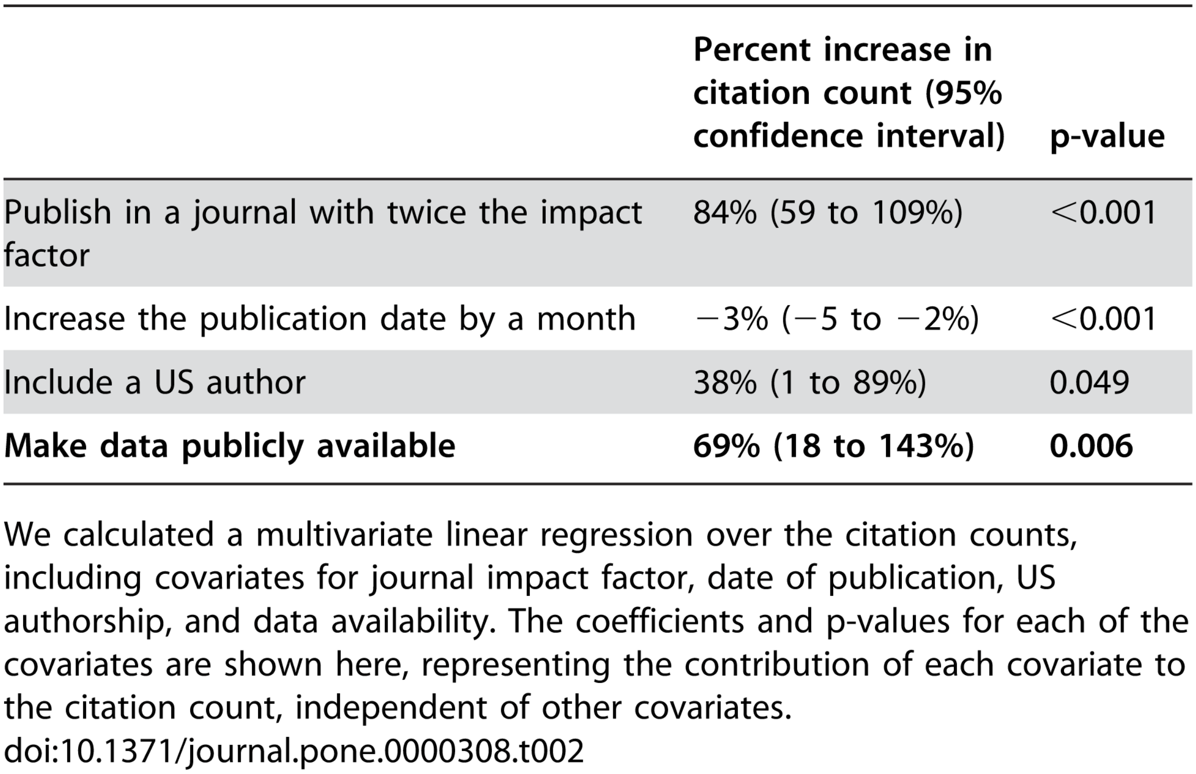


Figure 1: Distribution of 2004-2005 citation counts of 85 publications

Whether a trial's dataset was made publicly available was significantly associated with the log of its 2004–2005 citation rate (69% increase in citation count; 95% confidence interval: 18 to 143%, p=0.006), independent of journal impact factor, date of publication, and US authorship. Detailed results of this multivariate linear regression are given in Table 2. A similar result was found when we regressed on the number of citations each trial received during the 24 months after its publication (45% increase in citation count; 95% confidence interval: 1 to 109%, p = 0.050).

Table 2: Multivariate regression on citation count of 85 publications



	Percent increase in citation count (95% confidence interval)	**p-value**
Publish in a journal with twice the impact factor	84% (59 to 109%)	<0.001
Increase the publication date by a month	−3% (−5 to −2%)	<0.001
Include a US author	38% (1 to 89%)	0.049
Make data publicly available	**69% (18 to 143%)**	**0.006**

We calculated a multivariate linear regression over the citation counts, including covariates for journal impact factor, date of publication, US authorship, and data availability. The coefficients and p-values for each of the covariates are shown here, representing the contribution of each covariate to the citation count, independent of other covariates.
doi:10.1371/journal.pone.0000308.t002

To confirm that these findings were not dependent on a few extremely high-profile papers, we repeated our analysis on a subset of the cohort. We define papers published after the year 2000 in journals with an impact factor less than 25 as lower-profile publications. Of the 70 trials in this subset, only 27 (39%) made their data available, although they received 1875 of 2761 (68%) aggregate citations. The distribution of the citations by data availability in this subset is shown in Figure 2. The association between data sharing and citation rate remained significant in this lower-profile subset, independent of other covariates within a multivariate linear regression (71% increase in citation count; 95% confidence interval: 19 to 146%, p = 0.005).

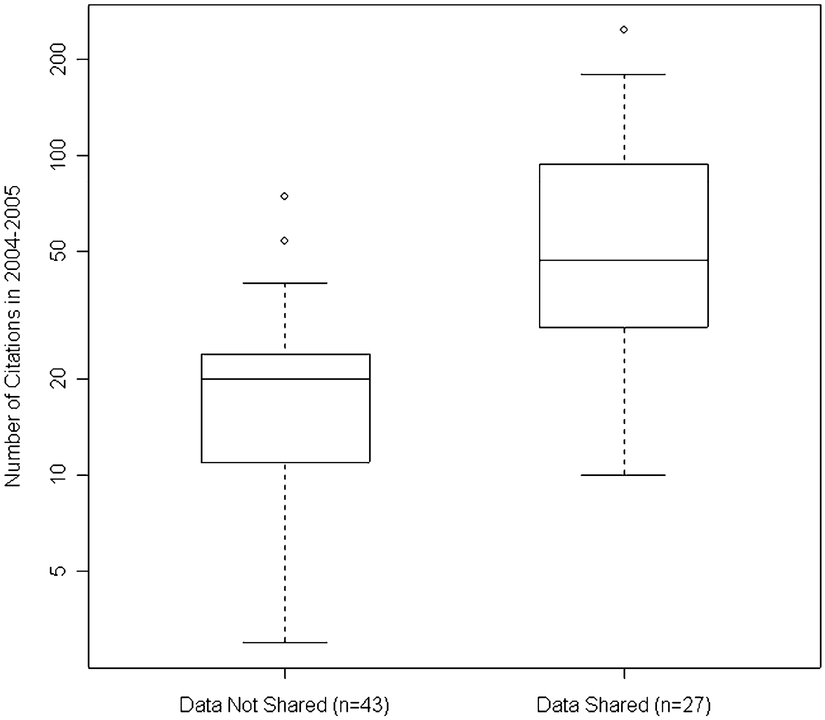
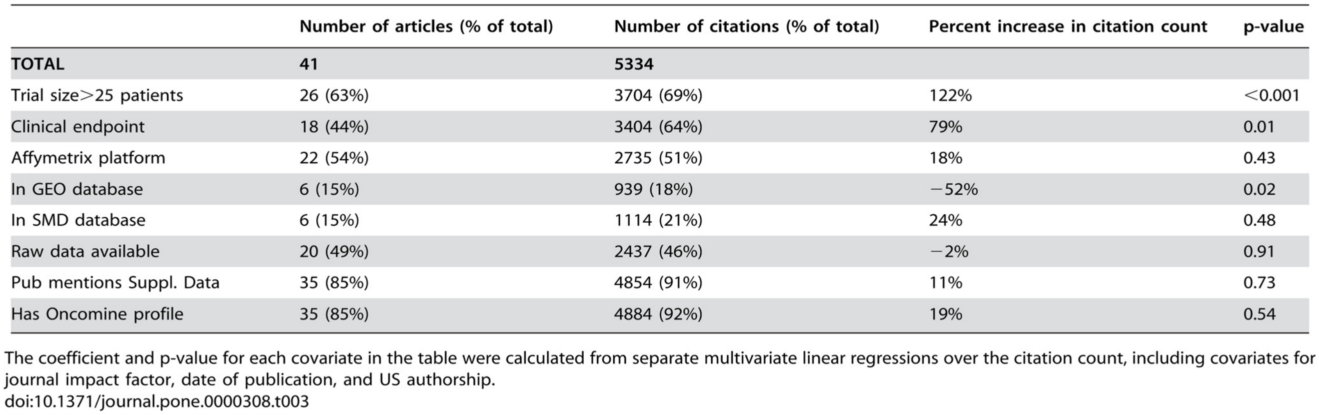


Figure 2: Distribution of 2004-2005 citation counts of 70 lower-profile publications

Lastly, we performed exploratory analysis on citation rate within the subset of trials which shared their microarray data; results are given in Table 3. The number of patients in a trial and a clinical endpoint correlated with increased citation rate. Assuming shared data is actually re-analyzed, one might expect an increase in citations for those trials which generated data on a standard platform (Affymetrix), or released it in a central location or format (SMD, GEO, GEDP) [109]. However, the choice of platform was insignificant and only those trials located in SMD showed a weak trend of increased citations. In fact, the 6 trials with data in GEO (in addition to other locations for 4 of the 6) actually showed an inverse relationship to citation rate, though we hesitate to read much into this due to the small number of trials in this set. The few trials in this cohort which, in addition to gene expression fold-change or other preprocessed information, shared their raw probe data or actual microarray images did not receive additional citations. Finally, although finding diverse microarray datasets online is non-trivial, an additional increase in citations was not noted for trials which mentioned their Supplementary Material within their paper, nor for those trials with datasets identified by a centralized, established data mining website. In summary, only trial design features such as size and clinical endpoint showed a significant association with citation rate; covariates relating to the data collection and how the data was made available only showed very weak trends. Perhaps with a larger and more balanced sample of trials with shared data these trends would be more clear.

Table 3: Exploratory regression on citation count for 41 publications with shared data



	Number of articles (% of total)	Number of citations (% of total)	Percent increase in citation count	p-value
TOTAL	**41**	**5334**		
Trial size>25 patients	26 (63%)	3704 (69%)	122%	<0.001
Clinical endpoint	18 (44%)	3404 (64%)	79%	0.01
Affymetrix platform	22 (54%)	2735 (51%)	18%	0.43
In GEO database	6 (15%)	939 (18%)	−52%	0.02
In SMD database	6 (15%)	1114 (21%)	24%	0.48
Raw data available	20 (49%)	2437 (46%)	−2%	0.91
Pub mentions Suppl. Data	35 (85%)	4854 (91%)	11%	0.73
Has Oncomine profile	35 (85%)	4884 (92%)	19%	0.54

The coefficient and p-value for each covariate in the table were calculated from separate multivariate linear regressions over the citation count, including covariates for journal impact factor, date of publication, and US authorship.
doi:10.1371/journal.pone.0000308.t003

2.4 DISCUSSION

We found that cancer clinical trials which share their microarray data were cited about 70% more frequently than clinical trials which do not. This result held even for lower-profile publications and thus is relevant to authors of all trials.

A parallel can be drawn between making study data publicly available and publishing a paper itself in an open-access journal. The association with an increased citation rate is similar [110]. While altruism no doubt plays a part in the motivation of authors in both cases, studies have found that an additional reason authors choose to publish in open-access journals is that they believe their articles will be cited more frequently [62, 111], endorsing the relevance of our result as a potential motivator.

We note an important limitation of this study: the demonstrated association does not imply causation. Receiving many citations and sharing data may stem from a common cause rather than being directly causally related. For example, a large, high-quality, clinically important trial would naturally receive many citations due to its medical relevance; meanwhile, its investigators may be more inclined to share its data than they would be for a smaller trial-perhaps due greater resources or confidence in the results.

Nonetheless, if we speculate for a moment that some or all of the association is indeed causal, we can hypothesize several mechanisms by which making data available may increase citations. The simplest mechanism is due to increased exposure: listing the dataset in databases and on websites will increase the number of people who encounter the publication. These people may then subsequently cite it for any of the usual reasons one cites a paper, such as paying homage, providing background reading, or noting corroborating or disputing claims ([112] provides a summary of research into citation behavior). More interestingly, evidence suggests that shared microarray data is indeed often reanalyzed [53], so at least some of the additional citations are certainly in this context. Finally, these re-analyses may spur enthusiasm and synergy around a specific research question, indirectly focusing publications and increasing the citation rate of all participants. These hypotheses are not tested in this study: additional research is needed to study the context of these citations and the degree, variety, and impact of any data re-use. Further, it would be interesting to assess the impact of reuse on the community, quantifying whether it does in fact lead to collaboration, a reduction in resource use, and scientific advances.

Since it is generally agreed that sharing data is of value to the scientific community [19, 53, 113-116], it is disappointing that less than half of the trials we looked at made their data publicly available. It is possible that attitudes may have changed in the years since these trials were published, however even recent evidence (in a field tangential to microarray trials) demonstrates a lack of willingness and ability to share data: an analysis in 2005 by Kyzas et al. [117] found that primary investigators for 17 of 63 studies on TP53 status in head and neck squamous cell carcinoma did not respond to a request for additional information, while 5 investigators replied they were unable to retrieve raw data.

Indeed, there are many personal difficulties for those who undertake to share their data [14]. A major cost is time: the data have to be formatted, documented, and released. Unfortunately this investment is often larger than one might guess: in the realm of microarray and particularly clinical information, it is nontrivial to decide what data to release, how to de-identify it, how to format it, and how to document it. Further, it is sometimes complicated to decide where to best publish data, since supplementary information and laboratory sites are transient [4, 5]. Beyond a time investment, 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 [118], a misunderstanding or misinterpretation of the data [8], 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. Data sources may be copyrighted such that the data subsets can not be freely shared, though it is always worth asking.

Although several of these difficulties are challenging to overcome, many are being addressed by a variety of initiatives, thereby decreasing the barriers to data sharing. For example, within the area of microarray clinical trials, several public microarray databases (SMD [119], GEO [102], ArrayExpress [103], CIBEX [104], GEDP(gedp.nci.nih.gov)) offer an obvious, centralized, free, and permanent data storage solution. Standards have been developed to specify minimal required data elements (MIAME [120] for microarray data, REMARK [121] for prognostic study details), consistent data encoding (MAGE-ML [122] for microarray data), and semantic models (BRIDG (www.bridgproject.org) for study protocol details). Software exists to help de-identify some types of patient records (De-ID [123]). The NIH and other agencies allow funds for data archiving and sharing. Finally, large initiatives (NCI's caBIG [39]) are underway to build tools and communities to enable and advance sharing data.

Research consumes considerable resources from the public trust. As data sharing gets easier and benefits are demonstrated for the individual investigator, hopefully authors will become more apt to share their study data and thus maximize its usefulness to society.

3.0 AIM 2A: USING OPEN ACCESS LITERATURE TO GUIDE FULL-TEXT QUERY FORMULATION

Background

Much scientific knowledge is contained in the details of the full-text biomedical literature. Most research in automated retrieval presupposes that the target literature can be downloaded and preprocessed prior to query. Unfortunately, this is not a practical or maintainable option for most users due to licensing restrictions, website terms of use, and sheer volume. Scientific article full-text is increasingly queryable through portals such as PubMed Central, Highwire Press, Scirus, and Google Scholar. However, because these portals only support very basic Boolean queries and full text is so expressive, formulating an effective query is a difficult task for users. We propose improving the formulation of full-text queries by using the open access literature as a proxy for the literature to be searched. We evaluated the feasibility of this approach by building a high-precision query for identifying studies that perform gene expression microarray experiments.

Methodology and Results

We built decision rules from unigram and bigram features of the open access literature. Minor syntax modifications were needed to translate the decision rules into the query languages of PubMed Central, Highwire Press, and Google Scholar. We mapped all retrieval results to PubMed identifiers and considered our query results as the union of retrieved articles across all portals. Compared to our reference standard, the derived full-text query found 56% (95% confidence interval, 52% to 61%) of intended studies, and 90% (86% to 93%) of studies identified by the full-text search met the reference standard criteria. Due to this relatively high precision, the derived query was better suited to the intended application than alternative baseline MeSH® queries.

Significance

Using open access literature to develop queries for full-text portals is an open, flexible, and effective method for retrieval of biomedical literature articles based on article full-text. We hope our approach will raise awareness of the constraints and opportunities in mainstream full-text information retrieval and provide a useful tool for today's researchers.

3.1 BACKGROUND

Much scientific information is available only in the full body of a scientific article. Full-text biomedical articles contain unique and valuable information not encapsulated in titles, abstracts, or indexing terms. Literature-based hypothesis generation, systematic reviews, and day-to-day literature surveys often require retrieving documents based on information in full-text only.

Progress has been made in accurately retrieving documents and passages based on their full-text content. Research efforts, relying on advanced machine-learning techniques and features such as parts of speech, stemmed words, n-grams, semantic tags, and weighted tokens, have focused on situations in which complete full-text corpora are available for preprocessing. Unfortunately, most users do not have an extensive, local, full-text library. Establishing and maintaining a machine-readable archive involves complex issues of permissions, licenses, storage, and formats. Consequently, applying cutting-edge full-text information retrieval and extraction research is not feasible for mainstream scientists.

Several portals offer a simple alternative: PubMed Central, Highwire Press, Scirus, and Google Scholar provide full-text query interfaces to an increasingly large subset of the biomedical literature. Users can search for full-text keywords and phrases without maintaining a local archive; in fact, they need not have subscription nor access privileges for the articles they are querying. Portals return a list of articles that match the query (often with a matching snippet). Users can manually review this list and download articles subject to individual licensing agreements.

It is difficult, however, to formulate an effective query for these portals: Full-text has so much lexical variation that query terms are often too broad or too narrow. This standard information retrieval problem has been extensively researched for queries based on titles, abstracts, and indexing terms. Much less research has been done on query expansion and refinement for full-text. Today's full-text portals offer very basic Boolean query interfaces only, with little support for synonyms, stemming, n-grams, or "nearby" operations.

We suggest that open access literature can help users build better queries for use within full-text portals. An increasingly large proportion of the biomedical literature is now published in open access journals such as the BMC family, PLoS family, Nucleic Acids Research, and the Journal of Medical Internet Research [124]. Papers published in these journals can be freely downloaded, redistributed, and preprocessed by anyone for any purpose. Furthermore, the NCBI provides a daily zipped archive of biomedical articles published by most open access publishers in a standard format, making it easy to establish and maintain a local archive of this content. If a proposed seed query has sufficient coverage, we believe that the open access literature could provide valuable information to expand and focus the query when it is applied to the general literature though established full-text portals.

We propose a method to facilitate the retrieval of biomedical literature through full-text queries run in publicly accessible interfaces. In this initial implementation, users provided a list of true positive and true negative PubMed identifiers within the open access literature. Standard text mining techniques were used to generate a query that accurately retrieved the documents based on the provided examples. We chose text-mining techniques that resulted in query syntax that was compatible with full-text portal interfaces, such as Boolean combinations, n-grams, wildcards, stemming, and stop words. The returned query was ready to be run through the simple interfaces of existing, publicly available full-text search engines. Full-text document hits could then be manually reviewed and downloaded by the user, subject to article subscription restrictions.

To evaluate the feasibility of this query-development approach, we applied it to the task of identifying studies that use a specific biological wet-laboratory method: running gene expression microarray experiments.

3.2 METHOD

3.2.1 Query development corpus

To assemble articles on the general topic of interest, we used the title and abstract filter proposed by Ocshner et al. [10]. We limited our results to those in the open access literature by running the following PubMed query:

"open access" [filter] *AND*
(microarray [tiab] OR microarrays [tiab] OR genome-wide [tiab]
OR "expression profile" [tiab] OR "expression profiles" [tiab]
OR "transcription profile" [tiab] OR "transcription profiling" [tiab])

We translated the returned PubMed identifiers to PubMed Central (PMC) identifiers, then to locations on the PubMed Central server. We downloaded the full text for the first 4000 files from PubMed Central and extracted the component containing the raw text in xml format.

To automatically classify our development corpus, we used raw dataset sharing into NCBI's Gene Expression Omnibus(GEO) database [125] as a proxy for running gene expression microarray experiments. This approach will incorrectly classify many gene-expression data articles, because either the authors did not share their gene expression data (about 50% [10]) or they did share but did not have a link to their gene expression study in GEO (about 35% [126]). Nonetheless, we expected the number of false negative instances to be small compared to the number of true negatives and thus sufficiently accurate for training. We implemented this filter by querying PubMed

Central with the development-corpus identifiers and the filter *AND "pmc_gds" [filter],* using the NCBI's EUtils web service. We considered articles returned by this filter to be positive examples, or gene expression microarray sharing/creation articles, and articles not returned in this subset to be negative examples.

3.2.2 Query development features

We assembled unigram and bigram features of the article full-text. Specifically, we removed all xml and split on spaces and all punctuation except hyphens. We excluded any unigram or bigram that included a word less than 3 characters long, more than 30 characters long, or that did not include at least one alphabetic character. We excluded unigrams and bigrams that included PubMed (and PubMed Central) stop words [127]. Due to the nature of our specific-use case for the query, we also excluded a manually derived list of bioinformatics data words, such as "geo", "omnibus", "accession number", "Agilent," and journal and formatting words, such as "bmc", "plos", "dtd", and "x000b0."

We eliminated unigrams and bigrams that did not have at least 20% precision, 20% recall, and a 35% f-measure on the entire training set.

3.2.3 Query development algorithm

Preliminary investigations using established rule-generation algorithms (JRip, Ridor, and others) in Weka returned queries with high f-measure but relatively low precision. Attempts to alter parameters to achieve high precision and acceptable recall were not successful, even with cost-weighted learning. Therefore, we decided to use a simple technique to build our own binary rules: assemble features with the highest recall joined with AND, assemble features with the highest precision joined by OR, and then AND the two assemblies together. This is illustrated in Figure 3.

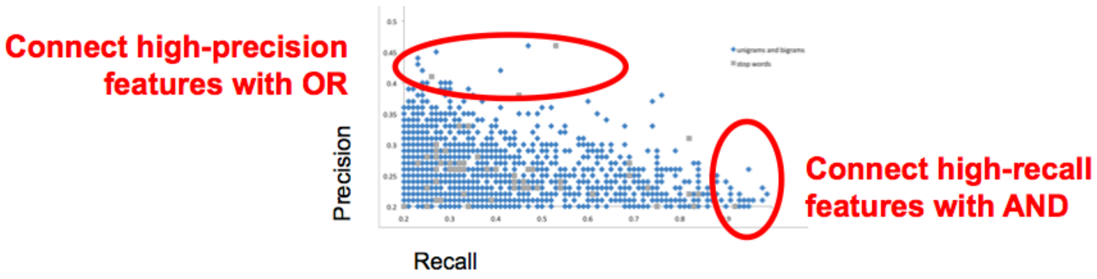


Figure 3: Method for building boolean queries from text features

We determined NOT phrases through a manual error analysis of the false positives in the development set.

3.2.4 Query syntax

The search syntax supported by established full-text portals is usually not well documented. We read available help files and experimented to determine capabilities, limitations, and syntax. We then translated the derived rules into the slightly different syntaxes of each of the query engines: PubMed Central, Highwire Press, Scirus, and Google Scholar.

3.2.5 Query evaluation corpus

We evaluated the performance of our derived query against the reference standard established by Ochsner et al. [10]. Although many of the reference articles have full-text freely available in PubMed Central, none are open access and thus none were in the development set.

Because the emphasis of Ochsner et al. was precision rather than recall, their analysis failed to identify a number of true positives. We searched for these misclassifications automatically by identifying whether any of the articles that were considered non-data-generating actually had linked database submissions in GEO: an indication that they did in fact generate data. We also manually examined all classification errors.

3.2.6 Query execution

We ran our query for all journals that included their complete content in PubMed Central first, then Highwire Press, and finally Google Scholar. This order allowed us to maximize the degree to which the query execution could be automated, as per the terms of use of the websites. We ran the queries in each location for articles published in 2007.

We used the EUtils library to automatically execute the query and obtain the results from PubMed Central. For the other query engines, we manually executed the query and manually saved the resulting html files on our computer. We parsed these html files with python scripts to extract the citations and submitted the citation lists to the PubMed Citation Matcher to obtain PubMed identifier (PMID) lists.

3.2.7 Query evaluation statistics

We calculated the precision and recall of the developed filters and compared this performance to that of the two most obvious baseline Medical Subject Heading (MeSH) filters:

"Gene Expression Profiling" AND "Oligonucleotide Array Sequence Analysis"
"Gene Expression Profiling" OR "Oligonucleotide Array Sequence Analysis"

We also used Fisher's exact test to verify that the filter was indeed adding value. For our use case, an eventual study of data sharing prevalence, we hoped to achieve recall of at least 50% and precision of at least 90%.

3.3 RESULTS

3.3.1 Queries

We applied our query-formulation approach to the task of identifying studies that performed gene expression microarray experiments. Using the open access literature as a development corpus and links to a gene expression microarray database as a proxy endpoint, we derived the full-text queries shown in Table 4.

Table 4: Derived microarray data creation queries for full-text portals

Portal	Query
PubMed Central	("gene expression" [text] AND "microarray" [text] AND "cell" [text] AND "rna" [text]) AND ("rneasy" [text] OR "trizol" [text] OR "real-time pcr" [text]) NOT ("tissue microarray*" [text] OR "cpg island*" [text])
HighWire Press	Anywhere in Text, ANY: ("gene expression" AND microarray AND cell AND rna) AND (rneasy OR trizol OR "real-time pcr") NOT (“tissue microarray*” OR “cpg island*”)
Google Scholar	+"gene expression” +microarray +cell +rna +(rneasy OR trizol OR "real time pcr") -"cpg island*" -"tissue microarray*"
Scirus	Anywhere in Text, ALL: ("gene expression" AND microarray AND cell AND rna) (rneasy OR trizol OR "real-time pcr") ANDNOT ("cpg island*" OR "tissue microarray*")

3.3.2 Evaluation portal coverage

Our evaluation corpus spanned 20 journals. We preferred to execute queries in PubMed Central when possible, since it allows automated query and results processing: As seen in Table 5, three of the 20 journals have deposited all of their content in PubMed Central. HighWire Press is also easy to use, though it does require manual querying and saving of results. Eight of the non-PubMed Central journals made their articles queryable by HighWire Press. The remaining journals listed their content in Scirus. Unfortunately, we were unable to reliably query full-text through Scirus, so we queried the remaining journals through Google Scholar for this study.

Table 5: Full-text portal coverage of reference journals, by portal preference

Portal	Journal
PubMed Central	Am J Pathol
	EMBO J
	PNAS
Highwire Press	Blood
	Cancer Res.
	Endocrinology
	FASEB J
	J. Biol. Chem.
	J. Endocrinol.
	J. Immunol.
	Mol. Cell. Biol.
	Mol. Endocrinol.
Scirus/Google Scholar	Cell
	Molecular Cell
	Nature
	Nature Cell Biology
	Nature Genetics
	Nature Medicine
	Nature Methods
	Science

3.3.3 Query performance

Ochsner et al. [10] identified 768 articles generally related to gene expression microarray data. Through a manual review, they determined that 391 of the articles documented the execution of a gene expression microarray experiment for a true positive rate of 51%. Our query replicated these results with a precision of 83%, recall of 62%, and f-measure of 69%.

Since the emphasis of the Ochsner review was precision rather than recall, we found that they were missing quite a few true positives. We searched for these misclassifications automatically by identifying whether any of the articles that were considered non-data-generating actually had linked database submissions in GEO: an indication that they did in fact generate data. Forty-four articles were reclassified based on this analysis. Our queries found seven of these reclassified articles and missed 37, resulting in a precision of 86% and recall of 57%.

We then manually examined all 41 remaining errors to see if any were due to erroneous manual classification. Based on our manual examination, we reclassified 28 articles as true positives, a true positive rate of 60%. Our query retrieved 12 of these and missed 18. Using this gold standard, the queries achieved a precision of 90% (95% confidence intervals: 86% to 93%), recall of 56% (52% to 61%), and f-measure of 69%. This performance was much improved over chance (p<0.001). We used the performance against this final gold standard for the remaining analyses.

To investigate if the queries would be effective in each of the full text portals, we examined the performance by portal, as shown in Table 6.

Table 6: Query accuracy by portal source

	N	precision	recall	f-measure
PubMed Central	149	96%	50%	65%
Highwire Press	498	91%	61%	73%
Google Scholar	121	67%	30%	42%
Weighted average	**768**	**90%**	**56%**	**69%**

The performance of all of these portals was improved over chance (p < 0.001), indicating that even the relatively poor performance of Google Scholar was adding value.

Finally, we compare the results of the derived query to two naïve queries based on Medical Subject Heading (MeSH) terms. As seen in Table 7, the derived query had better precision than either of the MeSH queries at an acceptable recall for our intended task.

Table 7: Query accuracy compared to baseline MeSH queries

	N	precision	recall	f-measure
"gene expression profiling" [mesh] OR "Oligonucleotide Array Sequence Analysis" [mesh]	768	81%	66%	73%
"gene expression profiling" [mesh] AND "Oligonucleotide Array Sequence Analysis" [mesh]	768	88%	24%	38%
Derived query	**768**	**90%**	**56%**	**69%**

3.4 DISCUSSION

We described a mechanism for formulating effective queries for use in publicly available, established full-text search portals, using the open access literature as training material. As a proof of concept, we applied this approach to a task that requires searching the full text of research articles: identifying studies that ran gene expression microarray experiments. The query we derived achieved 90% precision and 56% recall, making it a better fit for our intended application than lower-precision baseline MeSH queries. Although the evaluation demonstrates the usefulness of this approach in only one situation, we believe the method for deriving full-text queries could have widespread potential.

Effectively querying full-text is difficult: Synonyms, variant spellings, acronyms, and inexperience make it difficult to form effective queries [128]. Although difficult, searching full-text is often the only way to identify methods [85], detect harm [129], extract detailed data, or identify all of the biomedical concepts or genes explored in the study [130, 131]. There is also evidence that searching full-text is more effective than searching meta-data or abstracts for identifying articles of overall relevance [132, 133].

Domain-specific biomedical NLP and data integration systems, such as Textpresso [134], Pharmspresso [135], BioText [81], and BioLit [136], illustrate the potential value of accessing, exploring, and analyzing full-text, though none of these tools is designed to facilitate searching across domain-independent open-access and closed-access biomedical literature. Other systems have been built to take a preassembled corpus of positive and negative examples to build a filter query for execution in PubMed [137, 138], but to our knowledge, none suggest an easily accessed open-source training set nor result in a full-text query for use in domain-independent, publicly accessible online portals.

Existing full-text search portals, such as Google Scholar, Scirus, Highwire Press, and PubMed Central, differ in their features and performance [154, 155], though we believe their full-text searching capabilities have not yet been compared. We found differences in retrieval performance, but because our dataset was relatively small, it was not clear if any differences between portals were due to the portal or the subset of journals we searched.

While portals provide a source of articles, many prohibit systematic downloads [139]. Furthermore, it is unclear whether standard licensing agreements and fair use allow text mining, “a question on which informed people continue to disagree [157, 158]. Luckily, open access articles are available for download and all kinds of reuse. Evidence suggests that these articles have similar textual characteristics to traditional journal articles [140], and so we used them as a proxy for all articles.

Our method offers several advantages over alternatives: It is easy to maintain, it is free and open to query both open- and subscription-based content, and the user can be in direct control of recall/precision balance by setting recall and precision thresholds. It does have several limitations, however. This technique can only identify articles with full-text available for query in full-text portals, although we estimate that this is a sizeable amount of the total literature when results from PubMed Central, Highwire Press, Scirus, and Google Scholar are aggregated. A related limitation is that the distribution of articles in full-text portals could influence the distribution of retrieved articles. Articles published within the last year are unlikely to be retrieved, since many journals take full advantage of the NIH Public Access embargo period [141]. Furthermore, while a few journals have made their entire back archives digitally queryable, we suspect that recall of articles more than 10-years old would be relatively poor.

We also recognize that since this technique uses open access articles as a proxy for all articles, our queries would be most refined in areas that are well represented in open access articles. To the extent that there are topics poorly covered by open access articles, this technique could have difficulty deriving keywords to find them.

The system could be expanded in many ways. Its input could instead involve a seed query and a list of "true positive" passages. Other publicly available resources could also be consulted, including the UMLS®, WordNet, MEDLINE fields, and MeSH terms. Active learning might allow for further refinement. The system could run parts of speech analysis or domain-specific named entity recognition on the open access training set, if that helped to identify valuable features. It could extract features only from a certain subsection of manuscripts, if there were reason to believe that all relevant information would be in the Methods section, for example. The system could be enhanced to use bootstrapping to identify phrase variants [88]. Since some portals have some wildcard capabilities, we would like to experiment with learning regular expressions [142], though there is some evidence that this may not help [143]. Finally, more sophisticated natural language processing algorithms would become easier if this method were implemented within a system like LingPipe [143].

To better understand the relative strengths and weaknesses of this approach, it would be informative to compare its performance to other systems and algorithms on a standard task, such as the TREC Genomics corpus [86, 133], or a query that has been developed just on abstracts [144].

While our system will undoubtedly underperform compared with those at the cutting edge of research, we believe it will raise awareness of the constraints in mainstream full-text information retrieval and provide a useful tool for today's researchers.

4.0 AIM 2B: RECALL AND BIAS OF RETRIEVING GENE EXPRESSION MICROARRAY DATASETS THROUGH PUBMED IDENTIFIERS

Background

The ability to locate publicly available gene expression microarray datasets effectively and efficiently facilitates the reuse of these potentially valuable resources. Centralized biomedical databases allow users to query dataset metadata descriptions, but these annotations are often too sparse and diverse to allow complex and accurate queries. In this study we examined the ability of PubMed article identifiers to locate publicly available gene expression microarray datasets, and investigated whether the retrieved datasets were representative of publicly available datasets found through statements of data sharing in the associated research articles.

Results

In a recent article, Ochsner and colleagues identified 397 studies that had generated gene expression microarray data. Their search of the full text of each publication for statements of data sharing revealed 203 publicly available datasets, including 179 in the Gene Expression Omnibus (GEO) or ArrayExpress databases. Our scripted search of GEO and ArrayExpress for PubMed identifiers of the same 397 studies returned 160 datasets, including six not found by the original search for data sharing statements. As a proportion of datasets found by either method, the search for data sharing statements identified 91.4% of the 209 publicly available datasets, compared to 76.6% found by our search for PubMed identifiers. Searching GEO or ArrayExpress alone retrieved 63.2% and 46.9% of all available datasets, respectively. Studies retrieved through PubMed identifiers were representative of all datasets in terms of research theme, technology, size, and impact, though the recall was highest for datasets published by the highest-impact journals.

Conclusions

Searching database entries using PubMed identifiers can identify the majority of publicly available datasets. We urge authors of all datasets to complete the citation fields for their dataset submissions once publication details are known, thereby ensuring their work has maximum visibility and can contribute to subsequent studies.

4.1 BACKGROUND

The number of publicly available biomedical research datasets, such as those based on gene expression microarray experiments, continues to increase. The ability to access and process these large datasets enables other scientists to perform their own data driven studies, reduces duplicate data collection, allows the study of issues that require combining multiple datasets, and facilitates the training of future scientists through the analysis of real experimental data.

To realize these potential benefits, it is necessary that datasets can easily be found when needed. Biomedical databases typically include structured data fields indicating number of data samples, experimental platform and organism and tissue-type or disease of study. The experimental design, controls, and interventions involved are usually described in free-text fields. Unfortunately, the content of these descriptions is often sparse and diverse [145]. As a result, although basic queries of the structured fields can be effective, more complex queries may require pre-processing steps [146] and lack the accuracy required for some applications [147, 148].

Many publicly available datasets are associated with rich annotation outside the database: the published article describing the primary generation and analysis of the data. Centralized biomedical databases often include a “primary citation” field to link to the original published article or articles. This unambiguous link permits a user to query the article metadata, indexing terms, abstracts, or even the full text of the article, and then receive links to datasets relevant to the query.

The usefulness of Medical Subject Heading (MeSH) indexing terms for annotating gene expression datasets has been described by Butte and colleagues [147, 149, 150]. For example, they found that 53% of gene expression microarray datasets in the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database were linked to articles with disease related MeSH terms [147], that control/intervention gene expression data are publicly available for diseases contributing to 30% of all disease-related mortality in the United States }[149], and that approximately 10% of microarray experiments in GEO have MeSH terms related to pharmacological substances [150]. We expect that the use of MEDLINE annotations for dataset retrieval will increase, particularly as combining text and data analysis becomes more common [80, 136, 148, 151-154].

To identify the links between articles and their accompanying datasets, ideally a scientist could simply query PubMed, PubMed Central, or a specialized value added interface (e.g. MedMiner [155], BioText [81], or others [156]) and receive links to related datasets. This is possible within the Entrez network of databases. By appending "AND pubmed_gds [filter]" to any PubMed query, the set of returned articles is limited to those identified as a primary citation in a **G**ene Expression Omnibus GEO **D**ata**S**et record. While viewing PubMed results, selecting "GEO Datasets" in the Database dropdown list under "Find related data" in the right-hand column will retrieve the associated datasets. The data can then be explored or downloaded. In many cases, this primary citation query process can be automated. The Entrez databases can be queried through a web service eUtilities interface (http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html). Other databases offer similar web services or application programming interfaces.

As with any information retrieval strategy, retrieving datasets through their citation field identifiers has limitations. Not all publicly available datasets are submitted to centralized databases, and many are hosted on publisher or laboratory websites. Dataset citation fields are often empty because datasets are frequently submitted to databases before the research article has been published and assigned a PubMed ID. If we use a retrieval strategy based on article metadata, how many datasets are we missing? Are the datasets that are found a representative sample? If not, what are the biases?

To address these questions, in this study we have compared searching for publicly available datasets through statements of data sharing in published articles as reported by Ochsner et al. [10] to searching through queries of centralized databases with article PubMed identifiers. We have focused on gene expression microarray data, which is expensive to collect, is often shared, has well established data-sharing standards, and is valuable for reuse. The National Center for Biotechnology Information (NCBI) Gene Expression Omnibus [125] (GEO) and the European Bioinformatics Institute (EBI) ArrayExpress [157] databases have emerged as the dominant centralized repositories for sharing gene expression microarray data. Both include fields for primary article citations as PubMed IDs and support querying of those links.

4.2 METHODS

4.2.1 Reference standard

Ochsner and colleagues [10] manually curated gene expression microarray studies published in 20 journals during 2007. They began with a PubMed filter to identify studies related to gene expression microarray data, reviewed the gene expression articles to identify the subset of studies that generated primary gene expression datasets, and finally searched the full text of the published research articles for statements that the datasets were publicly available either in centralized databases, as supplementary information, or on public websites.

4.2.2 Database search for PubMed identifiers

We attempted to replicate the results of Ochsner et al. with a scripted query of gene expression databases. We began with their list of PubMed identifiers for articles identified as generating primary gene expression datasets. We then ran scripts to query the “article submission citation” field of the GEO and ArrayExpress databases with this list of PubMed IDs, and tabulated the datasets thereby retrieved.

We issued scripted queries for GEO and ArrayExpress through their web programmatic interfaces. For example, to query GEO for PubMed IDs **17510434** and **17603471**, we wrote programmatically retrieved the following page:

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=(17510434%5Buid%5D+OR+17603471%5Buid%5D)+AND+pubmed_gds%5Bfilter%5D

and then extracted the <IDList> from the resulting XML. To search ArrayExpress, we issued a query for each PubMed ID:

*http://www.ebi.ac.uk/microarray-as/ae/xml/experiments?keywords=**17510434***
*http://www.ebi.ac.uk/microarray-as/ae/xml/experiments?keywords=**17603471***

and confirmed the returned pages listed the PubMed ID in the bibliography field. We performed these queries with custom Python scripts.

4.2.3 Data extraction

For each of the datasets found in centralized databases, we collected the PubMed ID(s), the number of samples in the dataset, the gene expression platform, and the species. We considered the variable for dataset size to be “missing” for datasets shared outside centralized databases because the number of dataset samples was rarely explicitly and consistently stated on journal or laboratory websites.

For each PubMed identifier we collected the name of the journal that published the article, its 2007 Thomson ISI Journal Impact Factor, whether the article was indexed with the MeSH keyword that identifies cultured cells, and whether the article was found by the PubMed “cancer” filter (cancer was the most frequent disease classification for microarray data identified by Butte [147]). We collected PubMed Central citation statistics using the Entrez EUtils web service.

We determined whether each journal published articles within one specific discipline or had a multidisciplinary scope. We also recorded whether the journal requires authors to include a gene expression microarray database submission accession number in their articles as a condition of publication, following our earlier analysis of journal requirements [16].

If identical datasets were found in more than one location, we made note of this and collected data for the most complete location. Data collection was performed in May 2009 by manual download and with customized scripts (Python 2.5.2 and the EUtils python library [158].

4.2.4 Statistical analysis

We calculated the proportion of datasets that were retrievable by the Ochsner search and PubMed identifier queries, using the union of datasets found by either method as a denominator. We estimated the odds that defined subsets of gene expression microarray datasets (those investigating cancer, performed on an Affymetrix platform, involving humans, or involving cultured cells) would be retrieved by querying a database for their PubMed identifiers, relative to the odds they would be found by the Ochsner search but overlooked by the scripted query for PubMed identifiers. Fisher's exact test was used to determine whether the odds were significantly different than 1.0, with 95% confidence intervals. Histograms and Wilcoxon Rank Sum tests were used to determine whether the distributions of journal impact factors, number of citations, and number of data samples were significantly different between datasets found or overlooked by the PubMed identifier query. Statistics were calculated using the sciplot [159], Hmisc, and Design [160] libraries in R version 2.7.0 [108].

4.3 RESULTS

A previous article by Ochsner et al. [10] identified 397 published studies that generated gene expression microarray data. Their examination of data sharing statements revealed that 186 (47%) of these studies had made their datasets publicly available. Fourteen studies had more than one associated dataset (13 studies had two associated datasets, one study had five). The combined 203 datasets were found in a variety of locations: 147 (72%) in the Gene Expression Omnibus (GEO) database, 32 (16%) in the ArrayExpress database, 12 (6%) hosted on journal websites, and 12 (6%) on laboratory websites and smaller online data repositories. Combined, GEO and ArrayExpress housed 179 (88%) of the datasets found by the Ochsner search.

In order to determine the effectiveness of retrieving microarray datasets through an automated search, we attempted to locate these publicly available datasets using scripted queries of centralized microarray databases. We queried the GEO and ArrayExpress databases with the PubMed identifiers of the 397 data producing studies. Our scripted queries returned 160 datasets in total: 132 datasets in GEO and 98 datasets in ArrayExpress, including 70 datasets in both databases (ArrayExpress imports selected GEO submissions).

We compared the retrieval results of the two search strategies: Ochsner‘s search for data sharing statements within the full text of the published studies and our query of centralized databases for PubMed identifiers. As shown in Table 8, the query of databases using PubMed identifiers returned 6 datasets that were overlooked by Ochsner's search. Data submission dates suggested that one of these six was submitted after publication of the Ochsner paper. Ochsner's search found 31 datasets in GEO and ArrayExpress that were not found by the PubMed identifier search strategy: 18 of these database entries listed no article citation, 10 listed a different citation by the same research group, two listed incomplete citations lacking a PubMed ID, and one dataset entry included citations to papers by what appears to be a different group of authors.

Table 8: Comparison of dataset retrieval by two retrieval strategies

a) a search of article full-text for statements of data sharing, and b) a scripted query of centralized microarray databases for PubMed identifiers.

	b) Number of datasets **Found** by querying databases for PubMed IDs	Number of datasets **Not Found** by querying databases for PubMed IDs	Total
a) Number of datasets **Found** by searching full-text for statements of data sharing	154	49 (31 in GEO and ArrayExpress + 18 elsewhere)	203
Number of datasets **Not Found** by searching full-text for statements of data sharing	6	An unknown number of data-producing studies have publicly available data not found by either search method	at least 6
Total	160	at least 49	at least 209

The union of retrieval results from both search strategies yielded 209 datasets. We defined this union as the set "all publicly available datasets" for subsequent analysis. As illustrated in Figure 4, 91% of the 209 publicly available datasets were identified by the Ochsner search, compared to 77% found by queries of GEO and ArrayExpress for PubMed identifiers. PubMed identifier queries of either GEO or ArrayExpress alone retrieved 63% and 47% of all available datasets, respectively.

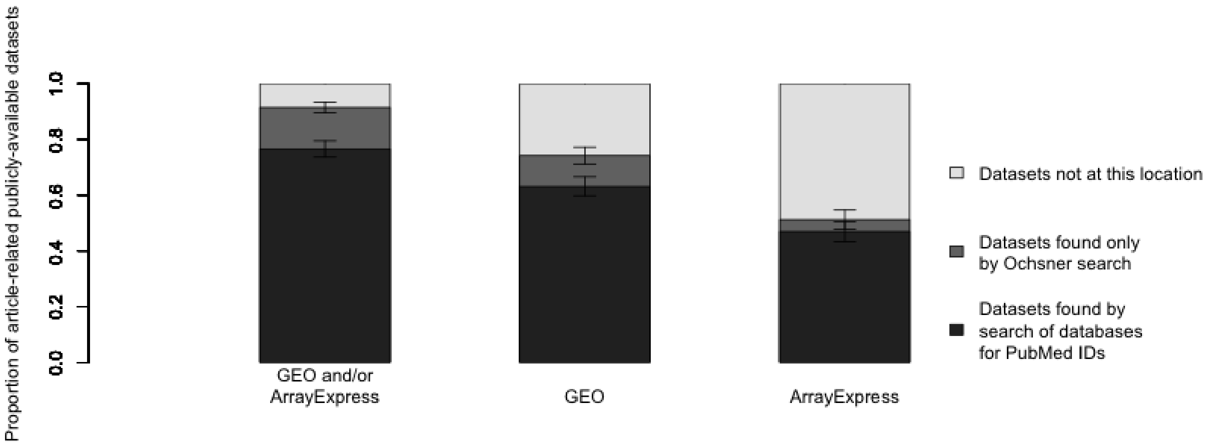


Figure 4: Datasets found or missed by PubMed ID queries, by database (bars indicate 95% confidence intervals of proportions)

Next, we looked at univariate patterns to determine whether the datasets retrieved through our search differed from those found only by the Ochsner search. The odds that a dataset was about cancer, performed on an Affymetrix platform, involved humans, or involved cultured cells were not significantly different whether the dataset was retrievable through our search method or not (p>0.3). The recall for datasets from disciplinary journals was similar to the recall from multidisciplinary journals (p>0.1). In ANOVA analysis, the distribution of species was not significantly different between the two search strategies (p>0.9).

Datasets found through PubMed identifiers were more likely to be associated with articles in higher impact journals than datasets overlooked by this retrieval method (p=0.01). Our PubMed identifier search found 92% of datasets from articles published in journals with impact factors greater than 20, 88% of those with impact factors between 10 and 20, and 73% of those with impact factors between three and 10. Journal data sharing policy and journal scope were strongly associated with journal impact factor (p<0.001), but stratifying our dataset by these features only slightly reduced the association between impact factor and recall (minimum p-value for stratified analysis was 0.06).

There was no association between the number of citations received by a study or the study sample size and whether or not the dataset was found by our PubMed identifier query. Histograms of the impact factors (Figure 5a), citations (Figure 5b), and dataset sample size (Figure 5c) found and overlooked by our query illustrate these patterns.

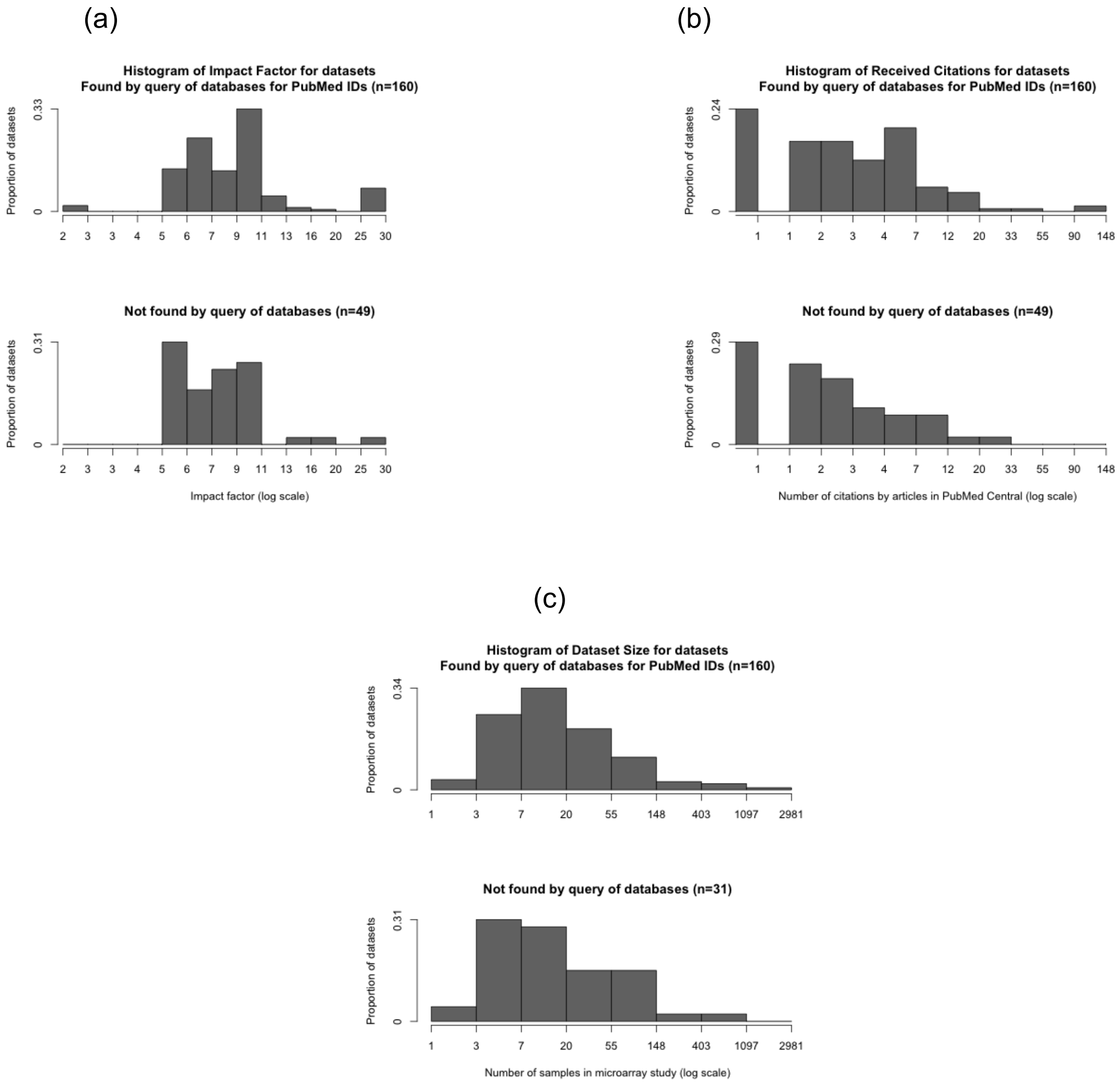


Figure 5: Datasets found or missed by PubMed ID queries, by impact and size

The ability to retrieve online datasets through PubMed identifiers differed across the twenty journals in our sample, as illustrated in Figure 6, although this difference was not statistically significant in an ANOVA test (p=0.9).

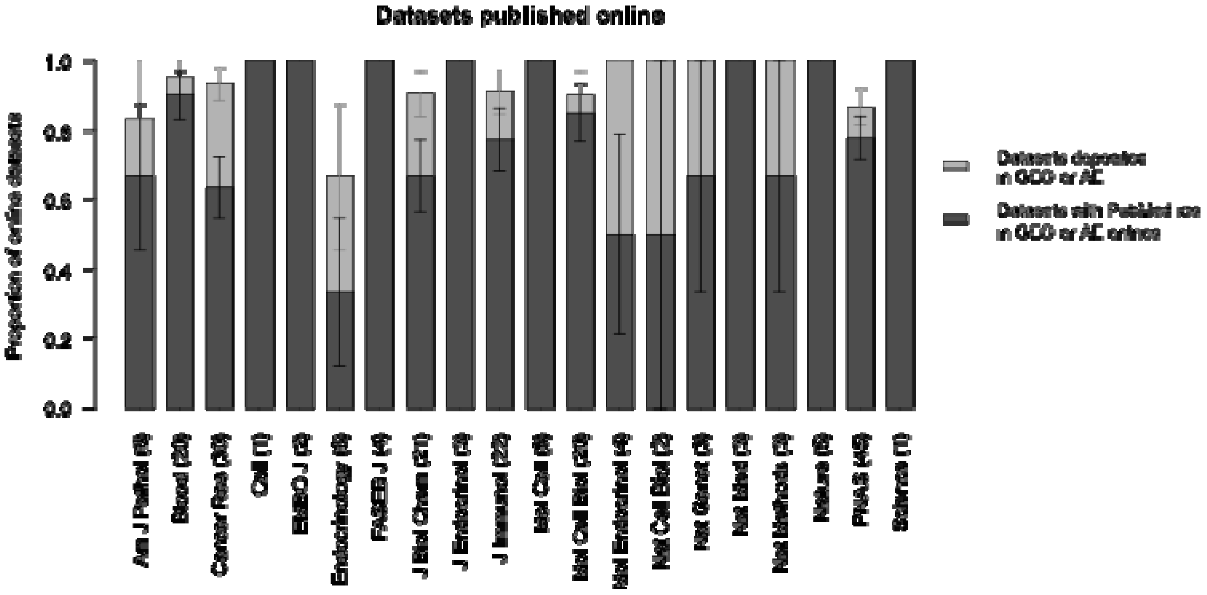


Figure 6: Datasets found or missed by PubMed ID queries, by journal (bars indicate 95% confidence intervals of proportions)

In Figure 6, light grey bars represent the proportion of online datasets available in the Gene Expression Omnibus or ArrayExpress databases. Dark grey bars represent the proportion of online datasets that include their publication PubMed identifier in the GEO or ArrayExpress entry, and thus can be found by our retrieval method. The number of online datasets in our sample follows the journal title, in parentheses.

Finally, we found some evidence that journal policy may be associated with whether a dataset is deposited into a database, complete with PubMed identifier citation. Our scripted queries found 78% of known publicly available datasets for articles published in journals that require a GEO or ArrayExpress submission accession number as a condition of publication. This is a higher retrieval rate than we found for publicly available datasets in journals without such a policy (65%), but the difference was not statistically significant (p=0.19).

4.4 DISCUSSION

In this study we found that scripted queries of centralized microarray databases using PubMed identifiers retrieved 76.6% of all publicly available datasets associated with the publications. The spectrum of datasets was similar to that found by a reference search [10] in terms of array platform, cell source, subject of study, sample size, and study impact.

Dataset retrieval through PubMed identifiers achieved the highest recall when applied to studies from the highest-impact journals. Additional research is needed to understand the reasons behind this finding since it is not fully explained by journal policy or scope, and may have to do with the implementation details of journal policy requirements. The importance of the retrieval bias depends on the intended use of the query results. For example, while there is likely no problem using the query to retrieve datasets for a combination analysis, caution is required when using the results for policy evaluation because query results are not fully representative of all online datasets,

Our evaluation has several limitations. The evaluation dataset was not chosen randomly and does not contain a representative distribution of journals: in particular, our evaluation subset lacked any journal with an impact factor below 2.5. Also, our reference standard classifications may contain errors, if there exist studies with publicly available data that were identified by neither the Ochsner search nor our PubMed identifier query.

We found that the number of gene expression microarray dataset entries with citation links could be increased by about 25% if all datasets now published on the internet were uploaded to centralized databases, and all primary article citation fields were fully completed. This is consistent with the findings of manual update efforts on the PDB database [57, 161]. We believe encouraging authors and enabling curators to document all link between datasets and research articles is effort well spent. In addition to use in retrieval, a clear relationship between a dataset and its research article allows synergistic documentation, integration for text mining and data mining, and facilitates rewards for publicly sharing data [162, 163].

This study considers the issue of retrieving datasets that are currently available on the internet. As noted by Ochsner et al., data from half of the published gene expression microarray studies does not appear to be publicly shared online [10]. Addressing incentives and policies for increasing the proportion of publicly available datasets is outside the scope of the current study but represents a crucial issue for unleashing the potential of research resources.

4.5 CONCLUSIONS

Efficient and accurate dataset retrieval can improve the efficiency of scientific progress, to the extent that it permits detailed review, facilitates integration, and reduces duplicate data collection. Our study suggests that querying gene expression microarray databases for PubMed identifiers is a feasible approach for identifying the majority of publication-related publicly available datasets, particularly when results from GEO and ArrayExpress are combined. The retrieved datasets are representative of all related publicly available datasets. We urge the authors of all datasets to complete the citation fields for their dataset submissions once publication details are known, thereby ensuring their work can have maximum visibility and fully contribute to future scientific studies.

5.0 AIM 3: WHO SHARES? WHO DOESN'T? FACTORS ASSOCIATED WITH SHARING GENE EXPRESSION MICROARRAY DATA

Many initiatives encourage research investigators to share their raw research datasets in hopes of increasing research efficiency and quality. Despite these investments of time and money, we do not have a firm grasp on the prevalence or patterns of data sharing and reuse; the effectiveness of initiatives; or the costs, benefits, and impact of repurposing biomedical research data. Previous survey methods for understanding data sharing patterns provide insight into investigator attitudes, but do not facilitate direct measurement of data sharing behaviour or its correlates. In this study, we use bibliometric methods to understand the prevalence and patterns with which investigators publicly share their raw gene expression microarray datasets after study publication.

We used automated methods to identify 11,603 publications that created gene expression microarray data and estimated that the authors of at least 25% of these publications deposited their data in the predominant public databases. We collected a wide set of variables about these studies and derived 15 factors that describe authorship, funding, institution, publication, and domain environments. Most factors were found to be statistically associated with the prevalence of data sharing. In particular, 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 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 analysis, previously sharing gene expression 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.

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

5.1 INTRODUCTION

Sharing and reusing primary research datasets has the potential to increase research efficiency and quality. Raw data can be used to explore related or new hypotheses, particularly when combined with other available datasets. Real data is indispensable for developing and validating 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 population resources by avoiding duplicate data collection.

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. For example, some journals require the submission of detailed biomedical datasets to publicly available databases as a condition of publication [15, 16]. Many funders require data sharing plans as a condition of funding: Since 2003, the National Institutes of Health (NIH) in the USA has required a data sharing plan for all large funding grants [17] and has more recently introduced stronger requirements for genome-wide association studies [164]. Several government whitepapers [14, 19] and high-profile editorials [165, 166] call for responsible data sharing and reuse. Large-scale collaborative science is increasing the need to share datasets [20, 167], and many guidelines, tools, standards, and databases are being developed and maintained to facilitate data sharing and reuse [120, 125].

Despite these investments of time and money, we do not yet understand the impact of these initiatives. There is a well-known adage: You cannot manage what you do not measure. For those with a goal of promoting responsible data sharing, it would be helpful to evaluate the effectiveness of requirements, recommendations, and tools. When data sharing is voluntary, insights could be gained by learning which datasets are shared, on what topics, by whom, and in what locations. When policies make data sharing mandatory, monitoring is useful to understand compliance and unexpected consequences.

Dimensions of data sharing action and intension have been investigated by a variety of studies. Manual annotations and systematic data requests have been used to estimate the frequency of data sharing within biomedicine [10, 11, 51, 117], though few attempts were made to determine patterns of sharing and withholding within these samples. Blumenthal [13], Campbell [52], Hedstrom [168], and others have used survey results to correlate self-reported instances of data sharing and withholding with self-reported attributes like industry involvement, perceived competitiveness, career productivity, and anticipated data sharing costs. Others have used surveys and interviews to analyze opinions about the effectiveness of mandates [53] and the value of various incentives [168-171]. A few inventories list the data-sharing policies of funders [172, 173] and journals [15, 174], and some work has been done to correlate policy strength with outcome [16, 175]. Surveys and case studies have been used to develop models of information behavior in related domains, including knowledge sharing within an organization [191, 192], physician knowledge sharing in hospitals [176], participation in open source projects [177], academic contributions to institutional archives [56, 178], the choice to publish in open access journals [179], sharing social science datasets [168], and participation in large-scale biomedical research collaborations [54].

Although these studies provide valuable insights and their methods facilitate investigation into an author’s intentions and opinions, they have several limitations. First, associations between an investigator’s intention to share data do not directly translate to an association with actually sharing data [180]. Second, associations that rely on self-reported data sharing and withholding likely suffer from underreporting and confounding, since people admit withholding data much less frequently than they report having experienced the data withholding of others [13].

We suggest a supplemental approach for investigating research data-sharing behavior. We have collected and analyzed a large set of observed data sharing actions and associated study, investigator, journal, funding, and institutional variables. In this report we explore common factors behind these attributes and look at the association between these factors and data sharing prevalence.

We chose to study data sharing for one particular type of data: biological gene expression microarray intensity values. Microarray studies provide a useful environment for exploring data sharing policies and behaviors. Despite being a rich resource valuable for reuse [181], microarray data are often, but not yet, universally shared. Best-practice guidelines for sharing microarray data are fairly mature [120, 182]. Two centralized databases have emerged as best-practice repositories: the Gene Expression Omnibus (GEO) [125] and ArrayExpress [157]. Finally, high-profile letters have called for strong journal data-sharing policies [34], resulting in unusually strong data sharing requirements in some journals [183].

5.2 METHODS

We identified a set of studies in which the investigators had generated gene expression microarray datasets, and then we identified the subset that had made their datasets publicly available on the internet. We analyzed attributes related to the investigators, journals, funding, institutions, and topic of the studies to determine which factors were associated with an increased frequency of data sharing.

5.2.1 Studies for analysis

The set of “gene expression microarray creation” articles was identified by searching the full-text of PubMed Central, Highwire Press, Scirus, and Google Scholar with portal-specific variants of the following query:

**("gene expression" AND "microarray" AND "cell" AND "rna")
AND ("rneasy" OR "trizol" OR "real-time pcr")
NOT ("tissue microarray*" OR "cpg island*")**

We found PubMed identifiers for the retrieved articles whenever possible and considered the union of these PubMed identifiers to be our studies for analysis. As discussed in Chapter 3, we previously evaluated the accuracy of this approach and found that it identified articles that created microarray data with a precision of 90% (95% confidence interval, 86% to 93%) and a recall of 56% (52% to 61%), compared to manual identification of articles that created microarray data.

Because Google Scholar only allows viewing of 1000 results per query, we were not able to identify all of its hits. We tried to identify as many as possible by iteratively appending a variety of attributes to the end of the query, including various publisher names, journal title words, and years of publication, thereby retrieving distinct subsets of the results 1000 hits at a time.

5.2.2 Study attributes

Our dependant variable was whether the gene expression microarray research articles had an associated dataset in a public centralized repository. As we showed in Chapter 4, we found that querying the NCBI's Gene Expression Omnibus and EBI's ArrayExpress with article PubMed identifiers located a representative 77% of all publicly available datasets associated with the published articles.

We implemented this same approach on the study articles; we queried GEO by submitting our PubMed identifiers to PubMed, then filtering them using the "pubmed_gds [filter]" query. We queried ArrayExpress by searching for each PubMed identifier in an offline copy of their public database. Those articles with an associated dataset in one of these two centralized repositories were considered to have "shared their data" for our endpoint, and those without such a link were considered not to have shared their data.

For every study article, we collected 124 attributes that were used as independent variables, as listed in the Appendix. The independent variables were collected automatically from a wide variety of sources. Basic bibliometric metadata was extracted from the MEDLINE record, including journal, year of publication, number of authors, Medical Subject Heading (MeSH) terms, number of citations from PubMed Central, inclusion in PubMed subsets for cancer, whether the journal is published with an open-access model and if it had data-submission links from Genbank, PDB, and SwissProt. The corresponding address was parsed for institution and country, following the methods of Yu et al.[184].

Institutions were cross-referenced to the SCImago Institutions Rankings 2009 World Report(http://www.scimagoir.com/) to estimate the relative degree of research output and impact of the institutions. The gender of the first and last authors were estimated using the Baby Name Guesser website at http://www.gpeters.com/names/baby-names.php. ISI Journal Impact Factors and associated metrics were extracted from the 2008 ISI Journal Citation Reports.

NIH grant details were extracted by cross-referencing grant numbers in the MEDLINE record with the NIH award information at http://report.nih.gov/award/state/state.cfm. From this information, we tabulated the amount of total funding received for each of the fiscal years from 2003 to 2008. We also estimated the date of renewal by identifying the most recent year in which a grant number was prefixed by a “1” or “2” —indication that the grant is “new” or “renewed,” respectively.

We quantified the content of journal data-sharing policies based on the “Instruction for Authors” for the most commonly occurring journals. We attempted to estimate if the paper itself reused publicly available gene expression microarray data by looking for its inclusion in the list that GEO keeps of reuse at http://www.ncbi.nlm.nih.gov/projects/geo/info/ucitations.html.

A list of prior publications in MEDLINE was extracted from Author-ity clusters, 2009 edition [185], for the first and last author of each article in our study. To limit the impact of extremely large “lumped” clusters that erroneously contain the publications of more than one actual author, we excluded prior publication lists for first or last authors in the largest 2% of clusters and instead considered this data to be missing. For all papers in an author's publication history with PubMed identifiers numerically less than the PubMed identifier of the paper in question, we queried for whether any of these prior publications had been published in an open source journal, were included in our "gene expression microarray creation" subset themselves, or had reused gene expression data. We recorded the date of the earliest publication by the author and the number of citations to date that their earlier papers received in PubMed Central.

Data collection scripts were coded in Python version 2.5.2 (many libraries, including EUtils, BeautifulSoup, pyparsing and nltk [186]) and SQLite version 3.4.

5.2.3 Statistical methods

Statistical analysis was performed in R version 2.10.1 [108]. P-values were two-tailed. Data was visually explored using Mondrian version 1.1 [187] and the Hmisc package [188]. We applied a square-root transformation to variables representing count data to improve their normality prior to calculating correlations.

To calculate variable correlations, we used the hector function in the polycor library. This computes polyserial correlations between pairs of numeric and ordinal variables and polychoric correlations between two ordinal variables. We modified it to calculate Pearson correlations between numeric variables using the rcorr function in the Hmisc library. We used a pairwise-complete approach to missing data and used the nearcor function in the sfsmisc library to make the correlation matrix positive definite. A correlation heatmap was produced using the gplots library.

We used the nFactors library to calculate and display the scree plot for our correlations.

Since our correlation matrix was not well-behaved enough for maximum-likelihood factor analysis, first-order exploratory factor analysis was performed with the fa function in the psych library, using the minimum residual (minres) solution and a promaxoblique rotation. Second-order factor analysis also used the minres solution but a varimax rotation, since we wanted these factors to be orthogonal. We computed the loadings on the original variables for the second-order factors using the method described by Gorsuch[189].

To compute the factor scores for the original dataset, we first had to impute the missing values. We did this using Gibbs sampling with two iterations through the mice library.

Using this complete dataset, we computed scores for each of our datapoints onto all of the first and second-order factors using Bartlett's algorithm as extracted from the factanal function. We submitted these factor scores to a logistic regression using the lrm function in the rms package. Continuous variables were modeled as cubic splines with 4 knots using the rcs function from the rms package, and all two-way interactions were explored.

Finally, we performed hierarchical supervised clustering on the datapoints to learn which factors were most predictive and then estimated the data sharing prevalence in a contingency table of these two clusters split at their medians.

5.3 RESULTS

Our queries for identifying microarray data-producing articles returned PubMed identifiers for 11,603 studies.

MEDLINE fields were still "in process" for 512 records, resulting in missing data for our MeSH-derived variables (Human, Mice, effectiveness, etc.). Impact factors were found for all but 1,001 articles. Journal policy variables were missing for 4,107 articles. The institution ranking attributes were missing for 6,185. We cross-referenced NIH grant details for 3,064 studies (some grant numbers could not be parsed, because they were incomplete or strangely formatted). We were able to determine the gender of the first and last authors, based on the forenames in the MEDLINE record, for all but 2,841 first authors and 2,790 last authors. All but 1,765 first authors and 797 last authors were found to have a publication history in the 2009 Author-ity clusters. A summary of the variables can be found in the Appendix and their correlations in Figure 7.

PubMed identifiers were found in GEO or ArrayExpress primary citation fields for 2,901 of the 11,603 articles in our dataset, indicating that 25% (95% confidence intervals: 24% to 26%) of the studies deposited their data in GEO or ArrayExpress and completed the “citation” fields with the primary article PubMed identifier. This is our estimate for the prevalence of gene expression microarray data deposited into the two predominant, centralized, publicly accessible databases. This data-sharing rate increased with each subsequent article publication year, as seen in Figure 8. The data sharing rate also varied across journals. Figure 9 shows the data sharing rate across the 50 journals with the most studies in our dataset.

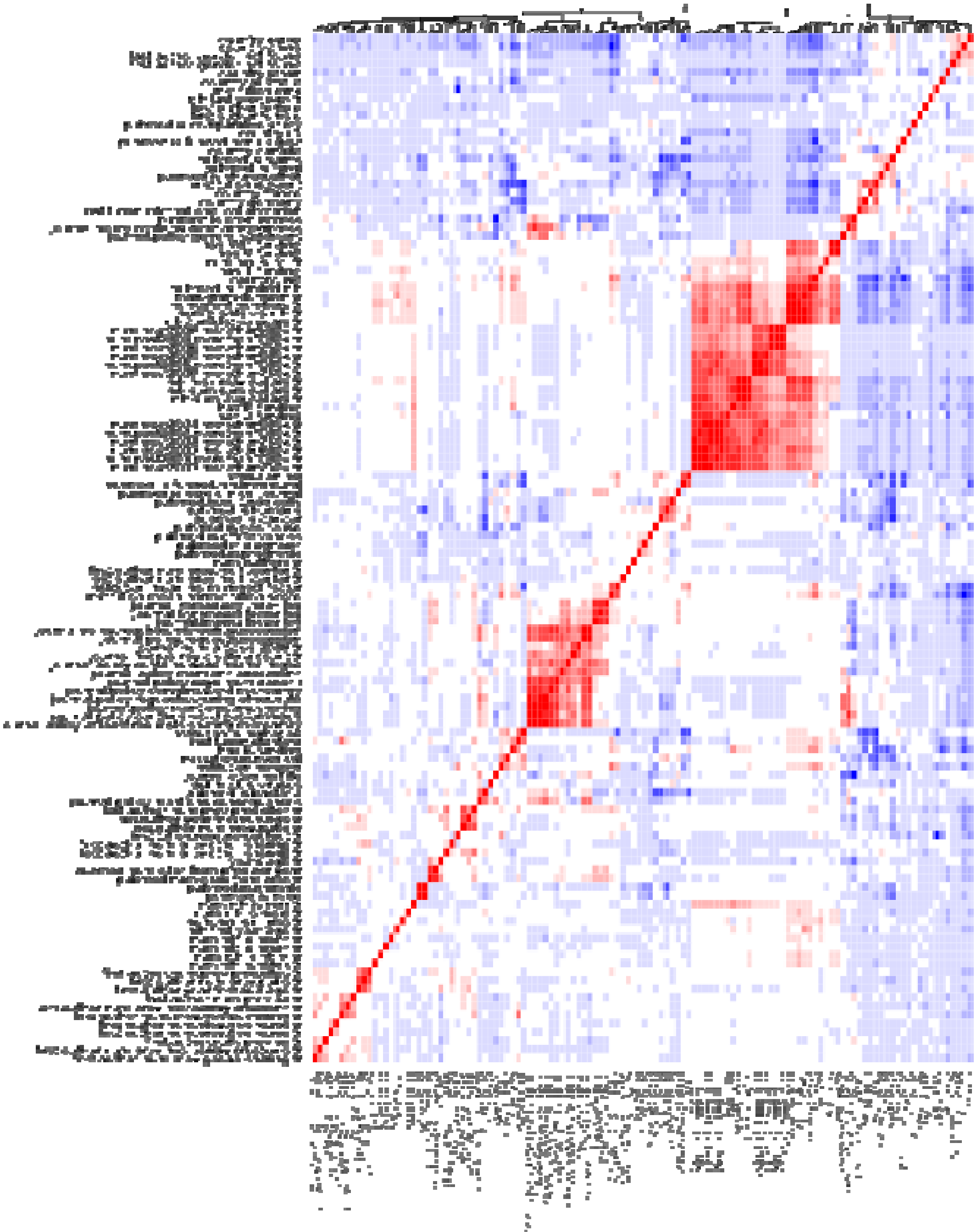


Figure 7: Covariance matrix of independent variables.

Positive correlations are red and negative correlations are blue.

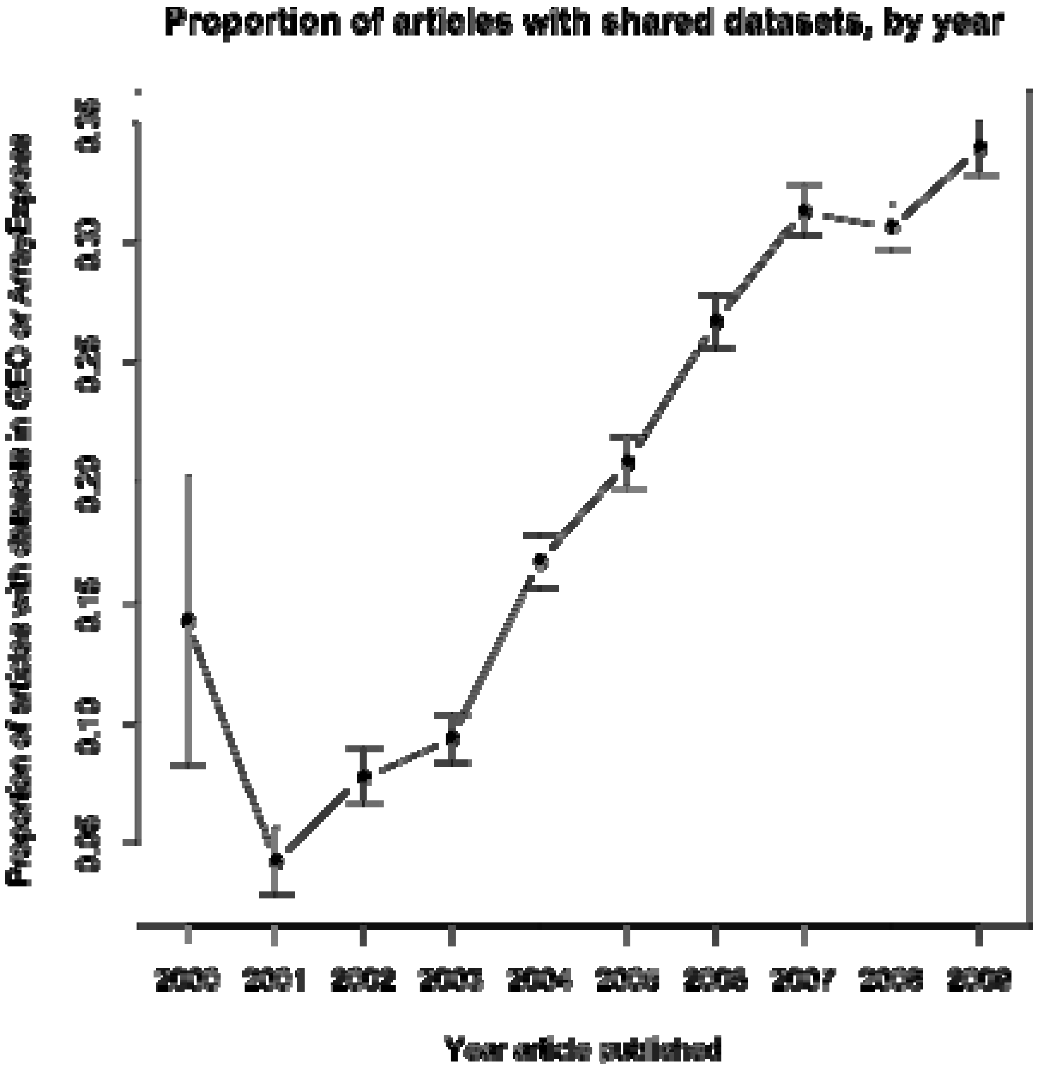


Figure 8: Proportion of articles with shared datasets, by year (error bars are 95% confidence intervals of the proportions)

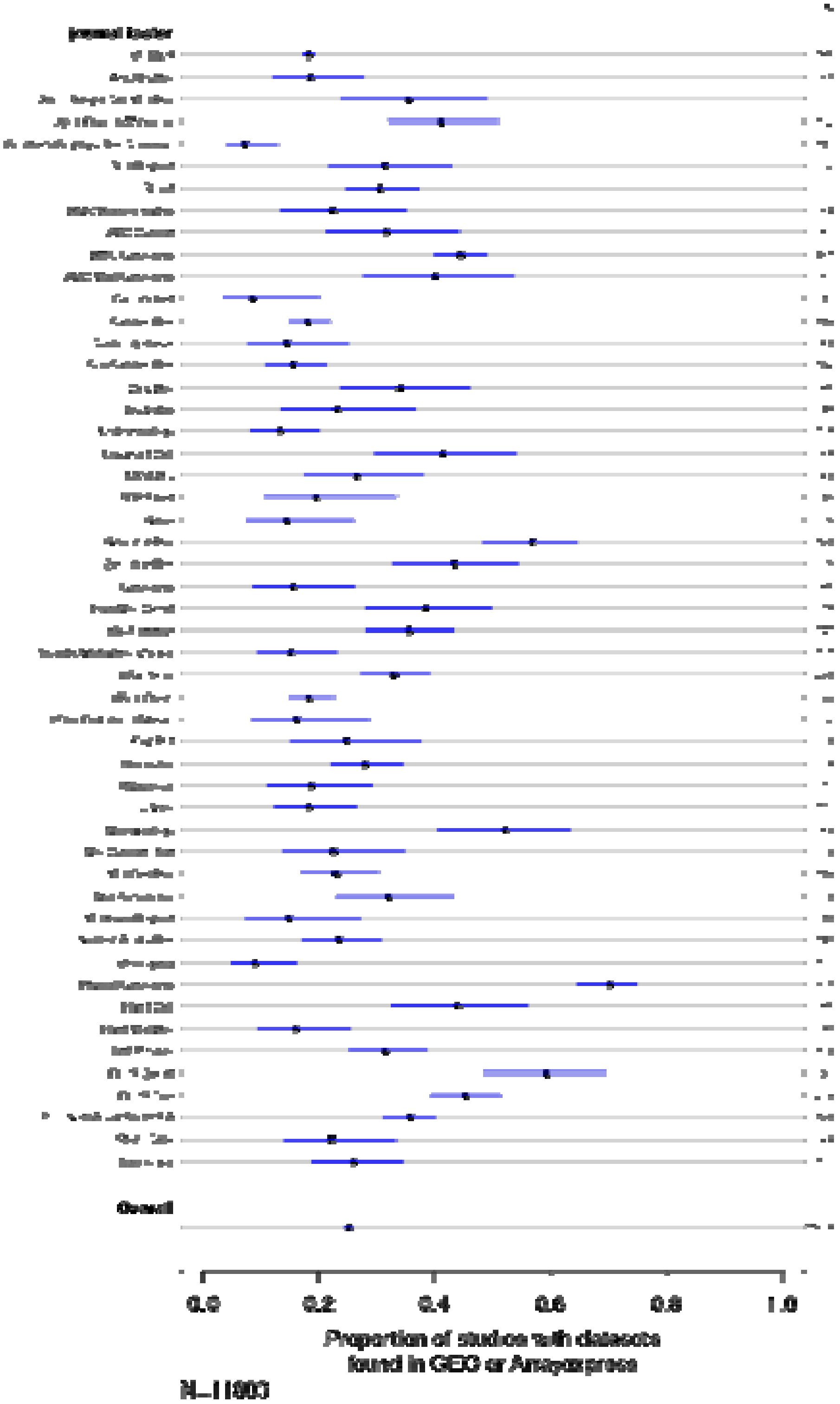


Figure 9: Proportion of articles with shared datasets, by journal (error bars are 95% confidence intervals of the proportions)

Many of our other attributes were also associated with the prevalence of data sharing in univariate analysis. Illustrations of these relationships are given in the Appendix.

5.3.1 First-order factors

We tried to use a scree plot to determine the number of factors for our first-order analysis. Since the scree plot did not have a clear drop-off, we experimented with a range of factor counts near the optimal coordinates index (as calculated by nScree in the nFactors R-project library) and finalized on 15 factors. Our correlation matrix was not sufficiently well-behaved for maximum-likelihood factor analysis, so we used a minimum residual (minres) solution. We chose to rotate our factors with the promax oblique algorithm, because we expected our first-order factors to have significant correlations with one another. The rotated first-order factors are given in Table 9 with loadings larger than 0.4 or less than -0.4. We named the factors based on the variables they load most heavily, using abbreviations for publishing in an Open Access journal (OA) and previously depositing data in the Gene Expression Omnibus (GEO) or ArrayExpress (AE) databases.

Table 9: First-order factor loadings

Large NIH grant

0.97	num.post2005.morethan1000k.tr
0.96	num.post2005.morethan750k.tr
0.92	num.post2004.morethan750k.tr
0.91	num.post2004.morethan1000k.tr
0.91	num.post2005.morethan500k.tr
0.89	num.post2006.morethan1000k.tr
0.89	num.post2006.morethan750k.tr
0.86	num.post2004.morethan500k.tr
0.85	num.post2006.morethan500k.tr
0.84	num.post2003.morethan750k.tr
0.84	num.post2003.morethan1000k.tr
0.80	num.post2003.morethan500k.tr
0.74	has.U.funding
0.71	has.P.funding
0.58	nih.sum.avg.dollars.tr
0.56	nih.sum.sum.dollars.tr
0.44	nih.max.max.dollars.tr

Has journal policy

1.00	journal.policy.contains..geo.omnibus
0.95	journal.policy.at.least.requests.sharing.array
0.95	journal.policy.mentions.any.sharing
0.93	journal.policy.contains.word.microarray
0.91	journal.policy.requests.sharing.other.data
0.85	journal.policy.says.must.deposit
0.83	journal.policy.contains.word.arrayexpress
0.72	journal.policy.requires.microarray.accession
0.71	journal.policy.requests.accession
0.58	journal.policy.contains.word.miame.mged
0.48	journal.microarray.creating.count.tr
0.45	journal.policy.mentions.consequences
0.42	journal.policy.general.statement

NOT institution NCI or intramural

0.59	pubmed.is.funded.non.us.govt
0.55	institution.is.higher.ed
-0.89	institution.nci
-0.86	pubmed.is.funded.nih.intramural
-0.42	country.usa

Count of R01 & other NIH grants

1.15	has.R01.funding
1.14	has.R.funding
0.89	num.grants.via.nih.tr
0.86	nih.cumulative.years.tr
0.82	num.grant.numbers.tr
0.80	max.grant.duration.tr
0.66	pubmed.is.funded.nih
0.50	nih.max.max.dollars.tr
0.45	num.nih.is.nigms.tr
0.44	country.usa
0.42	has.T.funding
0.41	num.nih.is.niaid.tr

Journal impact

0.88	journal.5yr.impact.factor.log
0.88	journal.impact.factor.log
0.85	journal.immediacy.index.log
0.70	journal.policy.mentions.exceptions
0.54	journal.num.articles.2008.tr
0.51	journal.policy.contains.word.miame.mged
-0.61	journal.policy.contains.word.arrayexpress
-0.48	pubmed.is.open.access

Last author num prev pubs & first year pub

0.84	last.author.num.prev.pubs.tr
0.74	last.author.year.first.pub.ago.tr
0.73	last.author.num.prev.pmc.cites.tr
0.68	last.author.num.prev.other.sharing.tr
0.48	country.japan
0.44	last.author.num.prev.microarray.creations.tr

Journal policy consequences & long half-life

0.78	journal.policy.mentions.consequences
0.73	journal.cited.halflife
0.60	pubmed.is.bacteria
0.42	journal.policy.requires.microarray.accession
-0.54	pubmed.is.open.access
-0.45	journal.policy.general.statement

Institution high citations & collaboration

0.76	institution.mean.norm.citation.score
0.72	institution.international.collaboration
0.64	institution.mean.norm.impact.factor
0.41	country.germany
-0.67	country.china
-0.61	country.korea
-0.56	last.author.gender.not.found
-0.43	country.japan

continued…

Table 9 (continued)

NO geo reuse & YES high institution output	
0.66	institution.research.output.tr
0.58	institution.harvard
0.46	has.K.funding
0.42	institution.stanford
-0.79	pubmed.is.geo.reuse
-0.62	country.australia
-0.46	institution.rank

NOT animals or mice	
0.51	pubmed.is.humans
0.43	pubmed.is.diagnosis
0.40	pubmed.is.effectiveness
-0.93	pubmed.is.animals
-0.86	pubmed.is.mice

Humans & cancer	
0.84	pubmed.is.humans
0.75	pubmed.is.cancer
0.67	pubmed.is.cultured.cells
0.52	institution.is.medical
0.47	pubmed.is.core.clinical.journal
-0.68	pubmed.is.plants
-0.49	pubmed.is.fungi

Institution is government & NOT higher ed	
0.92	institution.is.govnt
0.70	country.germany
0.65	country.france
0.46	institution.international.collaboration
-0.78	institution.is.higher.ed
-0.56	country.canada
-0.51	institution.stanford
-0.42	institution.is.medical

NO K funding or P funding	
0.56	has.R01.funding
0.49	has.R.funding
0.41	num.post2006.morethan500k.tr
0.41	num.post2006.morethan750k.tr
0.40	num.post2006.morethan1000k.tr
-0.65	has.K.funding
-0.63	has.P.funding

Authors prev GEOAE sharing & OA & arry creation	
0.83	last.author.num.prev.geoae.sharing.tr
0.74	last.author.num.prev.microarray.creations.tr
0.73	last.author.num.prev.oa.tr
0.60	first.author.num.prev.geoae.sharing.tr
0.47	first.author.num.prev.oa.tr
0.46	first.author.num.prev.microarray.creations.tr
0.40	institution.stanford
-0.44	years.ago.tr

First author num prev pubs & first year pub	
0.83	first.author.num.prev.pubs.tr
0.77	first.author.year.first.pub.ago.tr
0.73	first.author.num.prev.pmc.cites.tr
0.52	first.author.num.prev.other.sharing.tr

After imputing missing values, we calculated scores for each of the 15 factors for each of our 11,603 datapoints.   In univariate analysis, several of the factors demonstrated a correlation with frequency of data sharing, as seen in Figure 10. Several factors seemed to have a linear relationship with data sharing across their whole range. For example, whereas the data sharing rate was relatively low for studies that had the lowest score on the factor “Authors prev GEOAE sharing & OA & microarray creation” (in Figure 10, the first line under the heading “Authors prev GEOA sharing…”), the data sharing rate was higher for studies that had scores within the 25th to 50th percentile of all the studies in our sample, higher still for studies with “Authors prev GEO sharing…” factor scores in the third quartile, and studies that had a very high score on the factor, above the 75th percentile, had a relatively high rate of data sharing. A trend in the opposite direction can be seen for the factor “Humans & cancer”: the higher a study scored on that factor, the less likely they were to have shared their data.

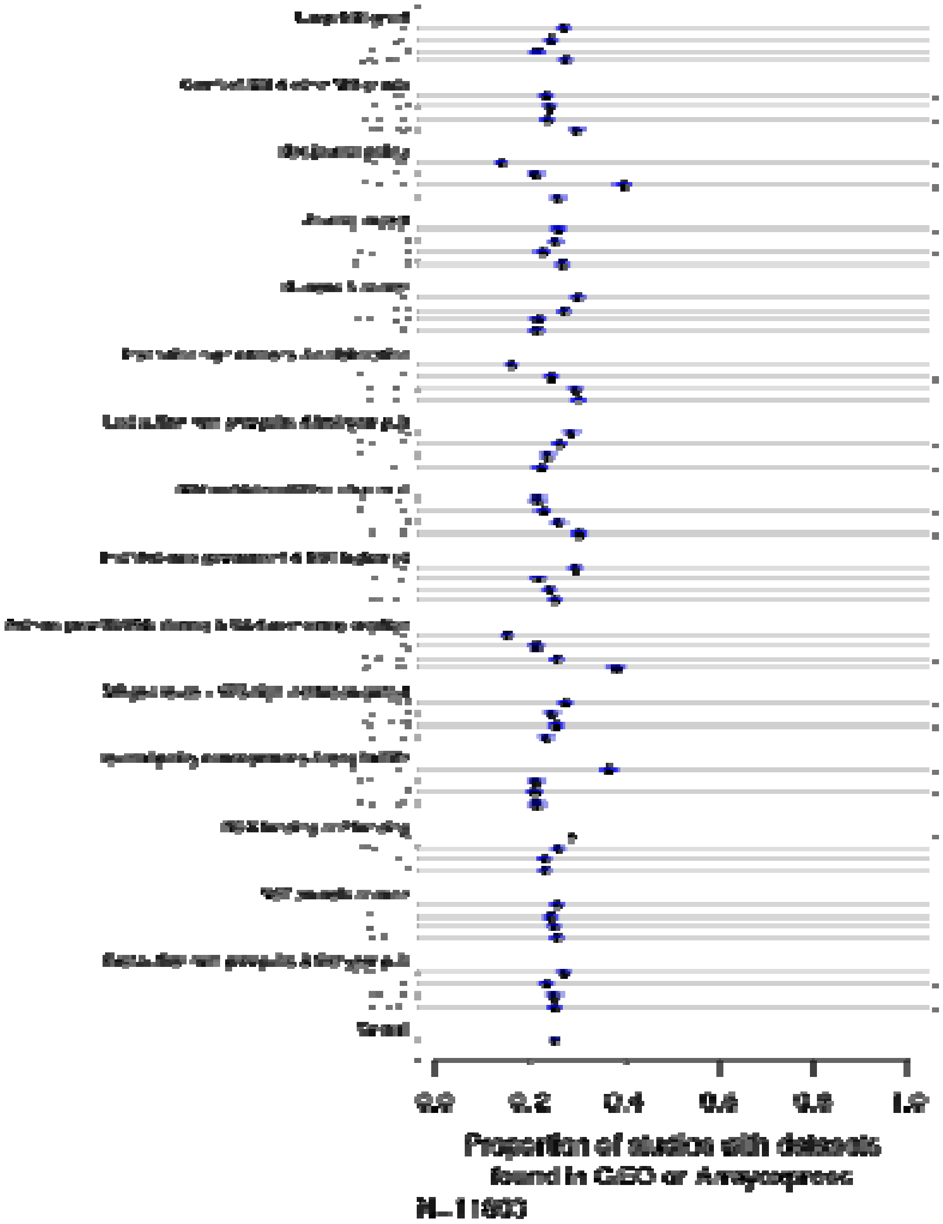


Figure 10: Association between shared data and first-order factors

Percentage of studies with shared data is shown for each quartile for each factor.

Univariate analysis.

Most of these factors were significantly associated with data-sharing behavior in a multivariate logistic regression: p=0.18 for "Large NIH grant", p<0.05 for "No GEO reuse & YES high institution output" and "No K funding or P funding", and p<0.005 for the other first-order factors. The increase in odds of data sharing is illustrated in Figure 11, as each factor in the model is moved from its 25th percentile value to its 75th percentile value.

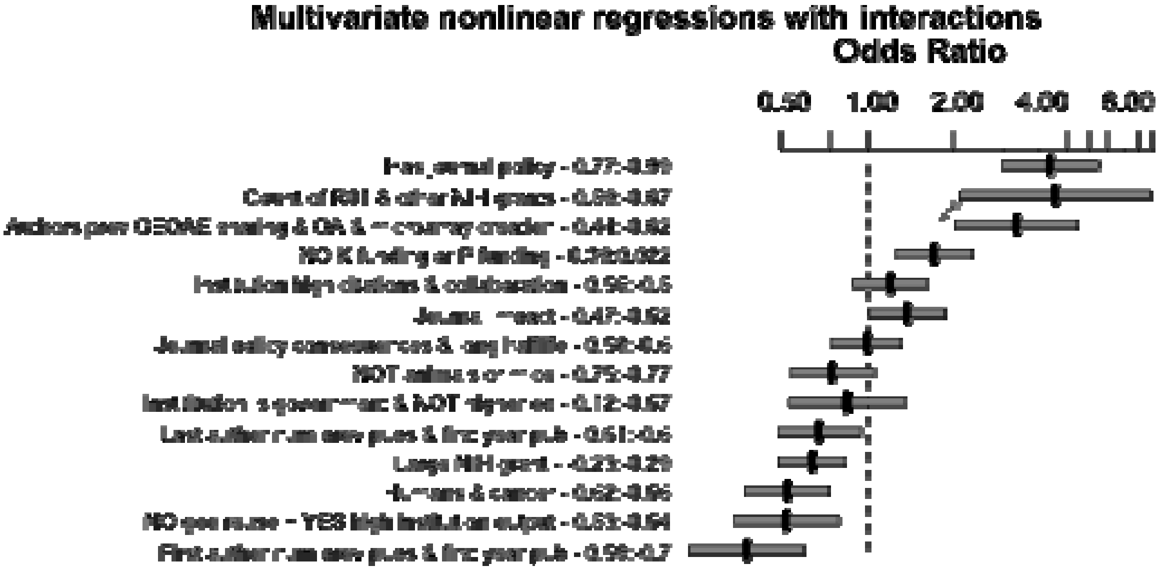


Figure 11: Odds ratios of data sharing for first-order factor, multivariate model
Odd ratios are calculated as factor scores are each varied from
their 25th percentile value to their 75th percentile value.
Horizontal lines show the 95% confidence intervals of the odds ratios.

5.3.2 Second-order factors

The heavy correlations between these factors suggest that second-order factors may be illuminating. Scree plot analysis of the correlations between the first-order factors suggested that we explore a solution containing five second-order factors. We calculated the factors using a “varimax” rotation to find orthogonal factors. The loadings on the first-order factors are given in Table 10.

Table 10: Second-order factor loadings, by first-order factors

Amount of NIH funding		
	0.88	Count of R01 & other NIH grants
	0.49	Large NIH grant
	-0.55	NO K funding or P funding
Cancer & humans		
	0.83	Humans & cancer
OA journal & previous GEO-AE sharing		
	0.59	Authors prev GEOAE sharing & OA & microarray creation
	0.43	Institution high citations & collaboration
	0.31	First author num prev pubs & first year pub
	-0.36	Last author num prev pubs & first year pub
Journal impact factor and policy		
	0.57	Journal impact
	0.51	Last author num prev pubs & first year pub
Higher Ed in USA		
	0.40	NO geo reuse + YES high institution output
	-0.44	Institution is government & NOT higher ed

Since interactions make these second-order variables slightly difficult to interpret, we followed the method explained by Gorsuch [189] to calculate the loadings of the second-order variables directly on the original variables. The results are listed in Table 11. We named the second-order factors based on the loadings on the original variables.

Table 11: Second-order factor loadings, by original variables

Amount of NIH funding

0.87	nih.cumulative.years.tr
0.85	num.grants.via.nih.tr
0.84	max.grant.duration.tr
0.82	num.grant.numbers.tr
0.80	pubmed.is.funded.nih
0.79	nih.max.max.dollars.tr
0.70	nih.sum.avg.dollars.tr
0.70	nih.sum.sum.dollars.tr
0.59	has.R.funding
0.59	num.post2003.morethan500k.tr
0.58	country.usa
0.58	has.U.funding
0.57	has.R01.funding
0.55	num.post2003.morethan750k.tr
0.53	has.T.funding
0.53	num.post2003.morethan1000k.tr
0.49	num.post2004.morethan500k.tr
0.45	num.post2004.morethan750k.tr
0.44	has.P.funding
0.43	num.post2004.morethan1000k.tr
0.43	num.nih.is.nci.tr
0.35	num.post2005.morethan500k.tr
0.32	num.nih.is.nigms.tr
0.31	num.post2005.morethan750k.tr

Cancer & humans

0.60	pubmed.is.cancer
0.59	pubmed.is.humans
0.52	pubmed.is.cultured.cells
0.43	pubmed.is.core.clinical.journal
0.39	institution.is.medical
-0.58	pubmed.is.plants
-0.50	pubmed.is.fungi
-0.37	pubmed.is.shared.other
-0.30	pubmed.is.bacteria

OA journal & previous GEO-AE sharing

0.40	first.author.num.prev.geoae.sharing.tr
0.37	pubmed.is.open.access
0.37	first.author.num.prev.oa.tr
0.35	last.author.num.prev.geoae.sharing.tr
0.32	pubmed.is.effectiveness
0.32	last.author.num.prev.oa.tr
0.31	pubmed.is.geo.reuse
-0.38	country.japan

Journal impact factor and policy

0.48	journal.impact.factor.log
0.47	jour.policy.requires.microarray.accession
0.46	jour.policy.mentions.exceptions
0.46	pubmed.num.cites.from.pmc.tr
0.45	journal.5yr.impact.factor.log
0.45	jour.policy.contains.word.miame.mged
0.42	last.author.num.prev.pmc.cites.tr
0.41	jour.policy.requests.accession
0.40	journal.immediacy.index.log
0.40	journal.num.articles.2008.tr
0.39	years.ago.tr
0.36	jour.policy.says.must.deposit
0.35	pubmed.num.cites.from.pmc.per.year
0.33	institution.mean.norm.citation.score
0.32	last.author.year.first.pub.ago.tr
0.31	country.usa
0.31	last.author.num.prev.pubs.tr
0.31	jour.policy.contains.word.microarray
-0.31	pubmed.is.open.access

Higher Ed in USA

0.36	institution.stanford
0.36	institution.is.higher.ed
0.35	country.usa
0.35	has.R.funding
0.33	has.R01.funding
0.30	institution.harvard
-0.37	institution.is.govnt

We then calculated factor scores for each of these second-order factors using the original attributes of our 11,603 datapoints. In univariate analysis, scores on several of the five factors showed a clear linear relationship with data sharing frequency, as illustrated in Figure 12.

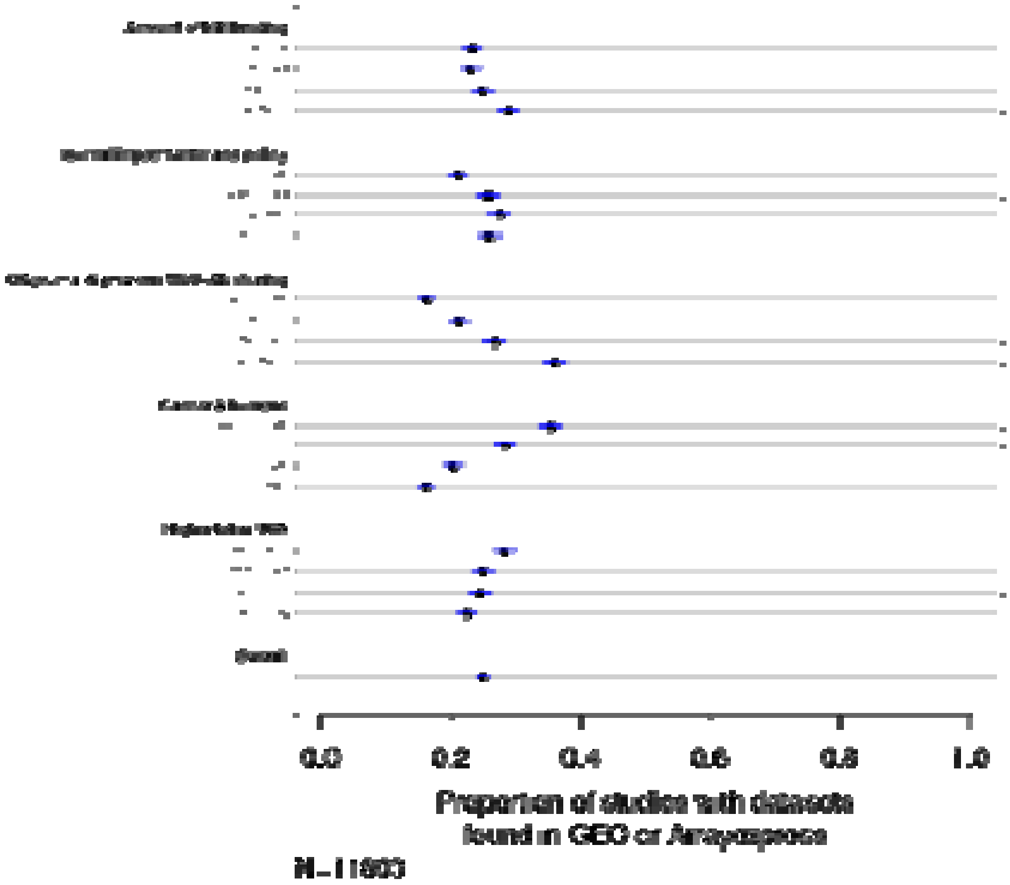


Figure 12: Association between shared data and second-order factors

Percentage of studies with shared data is shown for each quartile for each factor. Univariate analysis.

All five of the second-order factors were associated with data sharing in multivariate logistic regression, $p<0.001$.The increase in odds of data sharing is illustrated in Figure 13, as each factor in the model is moved from its 25th percentile value to its 75th percentile value.

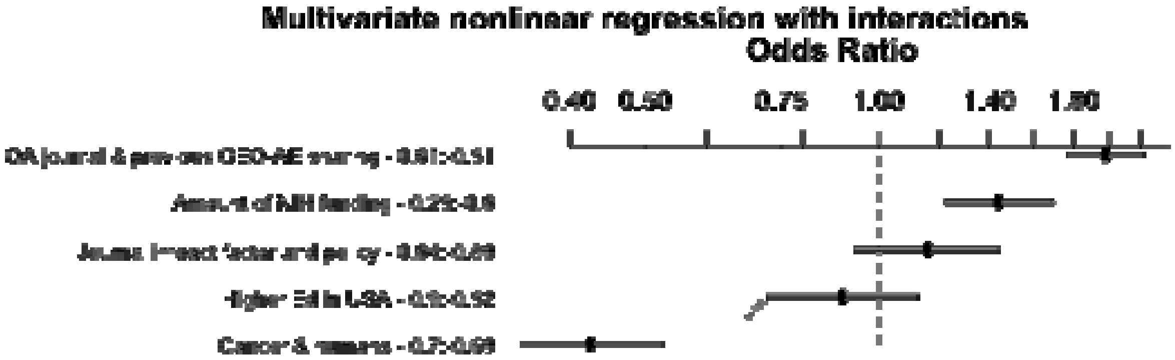


Figure 13: Odds ratios of data sharing for second-order factor, multivariate model
Odd ratios are calculated as factor scores are each varied from
their 25th percentile value to their 75th percentile value.
Horizontal lines show the 95% confidence intervals of the odds ratios.

Finally, to understand which of these factors is most predictive of data sharing behaviour, we performed supervised hierarchical clustering using our second-order factors. Splits on “OA journal & previous GEO-AE sharing” and “Cancer & Humans” were clearly the most informative, so we simply split these two factors at their medians and looked at the data sharing prevalence. As shown in Table 12, studies that scored high on the “OA journal & previous GEO-AE sharing” factor and low on the “Cancer & Humans” factor were almost three times as likely to share their data, compared to a “Cancer & Humans” study published without a strong “OA journal & previous GEO-AE sharing” background.

Table 12: Data sharing prevalence by two second-order factors
95% confidence intervals in brackets.

number of studies with shared data/ number of studies	**Above the median value for the factor “Cancer & Humans”**	**Below the median value for the factor “Cancer & Humans”**	**Total**
Above the median value for the factor “OA and previous GEO-AE sharing”	626/2614 = 24% [22%, 26%]	1193/3187 = **37% [36%, 39%]**	1819/5801 = 31% [30%, 33%]
Below the median value for the factor “OA and previous GEO-AE sharing”	428/3187 = **13% [12%, 15%]**	654/2615 = 25% [23%, 27%]	1082/5802 = 19% [18%, 20%]
Total	1054/5801 = 18% [17%, 19%]	1847/5802 = 32% [31%, 33%]	**2901/11603 = 25% [24%, 26%]**

5.4 DISCUSSION

This study explored the association between attributes of a published experiment and the probability that its raw data was shared in a publicly accessible database. We found that 25% of studies that perform gene expression microarray experiments have deposited their raw research data in a primary public repository. The proportion of studies that shared their gene expression datasets increased over time, from less than 5% in early years, before mature standards and repositories, to over 30% in 2009. Many factors derived from an experiment’s topic, impact, funding, publishing, institutional, and authorship environments were associated with the probability of data sharing. In particular, authors publishing in an open access journal, or with a history of sharing and reusing shared gene expression microarray data, were most likely to share their data, and those studying cancer or human subjects were least likely to share.

Although the current results should be considered preliminary, it is disheartening to discover that datasets of human and cancer studies have particularly low rates of data sharing. This sort of data is surely some of the most valuable for reuse, to the extent that it can help confirm, refute, advance, and train scientists in bench-to-bedside translational research. Further research will be required to understand the interplay of an investigator's motivation, opportunity, and ability that result in a low rate of data sharing in these studies [50, 190]. We can make some guesses: As is appropriate, concerns about privacy of human subjects' data undoubtedly affect a researcher's willingness and ability (perceived or actual) to share raw study data. We do not presume to recommend a proper balance between privacy and the societal benefit of data sharing, but we do feel strongly that researchers should seriously consider the re-identification risk of their data on a study-by-study basis [191], evaluate the risks and benefits across the wide range of stakeholder interests [45], and consider an ethical framework to make these difficult decisions [192]. Data-sharing rates could also be low for reasons other than privacy. Cancer researchers may perceive their field as particularly competitive, or cancer studies may have relatively strong links to industry– two attributes previously associated with data withholding [193, 194].

NIH funding levels are associated with increased prevalence of data sharing, though the overall probability of sharing remains low. Data sharing is infrequent even in studies funded by grants clearly covered by the NIH Data Sharing Policy, such as those that receive more than one million dollars per year and awarded or renewed since 2006. This result is consistent with reports that the NIH Data Sharing Policy is often not taken seriously because compliance is not enforced. [50]

We are intrigued that publishing in an open access journal, previously sharing gene expression data, and previously reusing gene expression data were associated with data sharing outcomes. The results are consistent with the results of our pilot study, in which we found a strong association between "author experience" and data sharing rates [195]. More research is required to understand the drivers behind the association. Does the factor represent an attitude towards "openness" by the decision-making authors? Does the act of sharing data lower the perceived effort of sharing data again? Does it dispel fears induced by possible negative outcomes from sharing data? To what extent does recognizing the value of shared data through data reuse motivate an author to share his or her own datasets?

People often wonder whether the attitude towards data sharing varies with age. Although we were not able to capture author age, we did estimate the number of years since first and last authors had published their first paper. Our analysis suggests that first authors with many years in the field are less likely to share data than those with fewer years of experience, but no such association for last authors. More work is needed to confirm this finding given the confounding factor of previous data-sharing experience.

Gene expression publications associated with Stanford University have a very high level of data sharing. The true level is actually much higher than that reflected in our study: Stanford University hosts a public microarray repository, and many articles that did not have a dataset link from GEO or ArrayExpress do mention submission to the Stanford Microarray Database. If one were looking for a community on which to model best practices for data sharing adoption, Stanford would be a great place to start.

Analyzing data sharing through bibliometric and data-mining attributes has several advantages: We can look at a very large set of studies and attributes, our 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. Our filters for identifying microarray creation studies do not have perfect precision, so we may have included some non-data-creation studies in our analysis. Because studies that do not create data will not have data deposits, their inclusion alters the composition of what we consider to be studies that create but do not share data. Furthermore, our method for detecting data deposits overlooks data deposits that are missing PubMed identifiers in GEO and ArrayExpress, so our dataset misclassifies some studies that did in fact share their data as non-data-sharing.

We made decisions to facilitate analysis, such as assuming that PubMed identifiers were monotonically increasing with publication date and using the current journal data-sharing policy as a surrogate for the data-sharing policy in place when papers were published. These decisions may have introduced errors.

Missing data may have obscured important information. For example, articles published in journals with policies that we did not examine had a lower rate of data sharing than articles published in journals whose “Instructions to Authors” policies we did quantify. It is likely that a more comprehensive analysis of journal data-sharing policies would provide additional insight. Similarly, the data we included on funders was limited: We only included funding information on NIH grants. Inclusion of more funders would help us understand the general role of funder policy and funding levels.

The Author-ity system provides accurate author publication histories: A previous evaluation on a different sample found that only 0.5% of publication histories erroneously included more than one author, and about 2% of clusters contained a partial inventory of an author’s publication history due to splitting a given author across multiple clusters [185]. However, because the lumping does not occur randomly, our attributes based on author publication histories may have included some bias. For example, the documented tendency of Author-ity to erroneously lump common Japanese names[185] may have confounded our author-history variables with author-ethnicity.

Previous work [193] found that investigator gender was correlated with data withholding. It is important to look at gender in multivariate analysis, since male scientists are more likely than women to have large NIH grants[196]. We found little evidence that gender of the first or last author was associated with data sharing, although we recognize limitations in our approach to determining gender. The Baby Name Guesser algorithm empirically estimates gender by analyzing popular usage on the internet. Although coverage across names from all ethnicities seems quite good, we were less able to determine gender for Asian names. This may have confounded our gender analysis, and our “gender not found” variable might have served as an unexpected proxy for author ethnicity.

In previous work we used h-index and a-index metrics to measure “author experience” for both the first and last author (In biomedicine, customarily, the first and last authors make the largest contributions to a study and have the most power in publication decisions.). A recent paper [197] suggests that a raw count of number of papers and number of citations is functionally equivalent to the h-index and a-index, so we used the raw counts in this study for computational simplicity. Our reliance on citations from PubMed Central (to enable scripted data collection) meant that older studies and those published in areas less well represented in PubMed Central were characterized by an artificially low citation count.

We believe our large sample of 11,603 studies captured a fairly diverse and representative subset of gene expression microarray studies, though our method of obtaining it through full-text query may have introduced a slight bias towards open access journals, as we discussed in Chapter 3.

This study did not consider directed sharing, such as peer-to-peer data exchange or sharing within a defined collaboration network, and thus underestimates the amount of data sharing in all its forms.

Furthermore, this study underestimated public sharing of gene expression data on the Internet. It did not recognize data listed in journal supplementary information, on lab or personal web sites, in specialized domains, or in institutional repositories (including the well-regarded and well-populated Stanford Microarray Database). Our study methods did not acknowledge deposits into the Gene Expression Omnibus or ArrayExpress, unless the database entry was accompanied by a citation to the research paper, complete with PubMed identifier. Finally, our 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 (http://www.ncbi.nlm.nih.gov/geo/info/faq.html).

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 our raw data. Nonetheless, we believe the aggregate data does support relative trends.

Finally, in regression studies it is important to remember that associations do not imply causation. It is possible, for example, that receiving a high level of NIH funding and deciding to share data are not causally related, but rather result from the exposure and excitement inherent in a “hot” subfield of study.

We plan to continue analyzing this data. In the spirit of the topic, we have made our raw data available online and encourage others to use it and report their findings. We hope these analyses will contribute to a deeper understanding of information behavior around research data sharing and eventually a culture that embraces the full potential of research output.

6.0 CONCLUSIONS

Aims 1, 2, and 3 were successfully completed, as described in the previous chapters. Here I summarize my findings, describe my contributions and their impact to date, suggest future work, and share some personal reflections.

6.1 SUMMARY

The purpose of this project was not to assess all data sharing behavior in biomedical research, but rather to explore three aspects of such an evaluation:

- Aim 1: Does sharing have benefit for those who share?
- Aim 2: Can sharing and withholding be systematically measured?
- Aim 3: How often is data shared? What predicts sharing? How can we model sharing behavior?

To begin, we analyzed the citation history of 85 clinical trials published between 1999 and 2003. Almost half of the trials had shared their microarray data publicly on the internet. Publicly available data was significantly (p=0.006) associated with a 69% increase in citations, independently of journal impact factor, date of publication, and author country of origin.

Digging deeper into data sharing patterns required methods for automatically identifying data creation and data sharing. Data creation is usually only communicated in a published study's full-text article. Because full text is increasingly queryable through portals such as PubMed Central, Highwire Press, and Google Scholar, we proposed a method to derive full-text queries from analysis of the open access literature. The derived full-text query found 56% of data-creation studies in our gold standard, with 90% precision. Next, we established that searching the two predominant, public, centralized gene expression microarray databases for biomedical literature PubMed identifiers retrieved 77% of associated publicly-accessible datasets.

We used these methods to identify 11603 publications that created gene expression microarray data and estimated that the authors of at least 25% of these publications deposited their data in the predominant public databases. We collected a wide set of variables about these studies and derived 15 factors that describe their authorship, funding, institution, publication, and domain environments. Most factors were associated with the prevalence of data sharing. In second-order analysis, authors with a history of sharing and reusing shared gene expression microarray data were most likely to share their data, and those studying human subjects and cancer were least likely to share.

6.2 CONTRIBUTIONS, IMPLICATIONS, AND FUTURE WORK

The goal of this project has been accomplished: useful evidence on data sharing patterns has been collected through methods that can be applied broadly, repeatably, and cost-effectively. In this section, I summarize the contributions of this project, reactions to the portions that have already been published, suggest a few paths to confirm the preliminary results and extend the analysis, and speculate about implications of the results should they be confirmed.

6.2.1 Contributions

This research work has made several contributions in the form of papers and associated datasets. Several of these have been met with a warm reaction, suggesting they have made a valuable contribution to ongoing dialog about scientific data sharing:

- an assessment of citation benefits of data sharing, published in PLoS ONE

    - Peter Suber, in Open Access News: “Many studies have shown a correlation between OA *articles* and citation impact. I believe this is the first study to document a similar correlation between OA *data* and citation impact.”
    - viewed over 13000 times at PLoS ONE
    - 45 citations from items in Google Scholar, including citations from research articles, books, and editorials
- an award-winning proposal (Thomson-Reuters Dissertation Proposal Scholarship for 2009), openly available online
    - used as a case-study in a PhD-level course at the School of Information Studies, McGill University
- a generalizable approach for developing practical full-text queries for use in established academic literature portals, to be submitted for publication
    - in use by a colleague at the National Core for Neuroethics at the University of British Columbia
- an evaluation of the precision, recall, and bias of using PubMed identifiers to find publicly available gene expression microarray datasets, accepted for publication
- an estimate of the prevalence and patterns of gene expression microarray dataset sharing and preliminary models of data sharing behavior, to be submitted for publication
- a publicly available dataset associating microarray study publications with data sharing status
- open source Python data collection code and R-project statistical analyses

6.2.2 Findings

6.2.2.1 Data sharing is associated with an increased citation rate

Based on 85 cancer clinical trials, we found that publications that made their datasets publicly available received 69% more citations than similar publications that did not share their data. Several editorials have cited this evidence when debuting stricter data sharing policies, suggesting this finding has been helpful for those trying to promote data sharing.

Before an estimate of the association between data sharing and citation rate can have profound implications, however, the estimates need to be confirmed. Ideally it would be confirmed with a larger dataset, more covariates, and different methods across several domains and datatypes. As a first step towards this ambitious goal, I plan to use the dataset and covariates collected in this project to investigate the association between the data sharing choices and citation rates of the 11603 gene expression microarray data-creation studies. Future work will be needed to adapt the automated retrieval methods for use outside biomedicine and gene expression microarray data.

I hypothesize that the association between data sharing and citation rate will be confirmed, though I suspect the citation benefit will be smaller than the initial estimate of 69%. My guess is that cancer clinical trial data might be reused more than datasets of non-human organisms, since bioinformaticians may wish to demonstrate their novel tools and methods are applicable to translational research. I also expect, given the current reuse patterns for gene expression microarray data, that as the number of gene expression microarray datasets continues to increase over time, any given dataset is reused less often. Furthermore, the initial estimate calculation did not include potentially important covariates for predicting citation rate, such as level of NIH funding – including these variables may decrease the estimated association between data sharing and citation rate.

I also hypothesize that there are domains and datatypes for which there is no citation benefit for sharing data. In some areas, the cultural norm is to cite an accession number rather than the originating paper. In others, typical reuse involves a very broad analysis across all data items in the database: it is impossible to cite all associated papers.

It is important to note that we do not understand how motivating a citation benefit of a given size would be to individual authors. Furthermore, an estimate of citation benefit is just one aspect of potential benefits to individual investigators for sharing data.

To present a complete picture, this finding should be integrated with other individual benefits, individual costs, societal benefits, and societal costs.

6.2.2.2 Data creation studies can be identified through full-text queries

We described and evaluated a method to identify articles that create gene expression datasets using open access literature full text as training data and full-text portals as an execution environment.

How useful will this method be, outside of this study? Identifying data creation studies could be useful for investigators looking for data to reuse, for those monitoring the adoption of various research methods, and for extracting evidence types for biocurators.

The most important implication of this work, however, is in the general process we used. Most research in automated retrieval presupposes that the target literature can be downloaded and preprocessed prior to query. Unfortunately, this is not a practical or maintainable option for most users due to licensing restrictions, website terms of use, and sheer volume. Scientific article full text is increasingly queryable through online portals such as PubMed Central, Highwire Press, Scirus, and Google Scholar. Recognizing that these full-text portals can be used for broad systematic retrieval of the biomedical literature based on words and phrases in article full text, particularly when queries are developed, refined, and evaluated by applying machine learning techniques to open access articles, potentially opens up large areas of research and application.

Further research could increase the impact of this approach. A review is needed to describe the scope and breadth of full-text proxy engines. The methods presented here could easily be offered to the general public as an openly-available web service. Derived queries could be improved through application of more advanced text mining techniques. Finally, the methods will have to be refined for domains without well-organized portals like PubMed Central and Highwire Press.

6.2.2.3 Datasets can be identified by their PubMed identifiers

We described and evaluated a method to identify articles that shared gene expression microarray datasets in centralized repositories, using PubMed identifiers. The method is not novel, but knowing the recall and bias may encourage adoption of this method by others. We hope to combine this method and others like it in a web service to help researchers find datasets for reuse.

Unfortunately, this method is difficult to apply to datatypes without centralized databases and to domains not covered by MEDLINE. Future research is needed to determine mechanisms for assessing dataset quality.

6.2.2.4 Many attributes are correlated with data sharing behaviour

We collected a large dataset and found that many attributes were correlated with data sharing behaviour, particularly a history of sharing and reusing shared gene expression microarray data and a focus on human subjects and cancer. These results are preliminary: Confirmation is needed before any of the associations inform policy or decisions.

The immediate implications of this study are those of a proof of concept and published dataset: many new avenues of research. Structural equation modeling can be used to explore causality within the variables. The environmental factors can be further examined and perhaps applied in new contexts. A deeper look into journal and funder policies could be used to explore the direct impacts that their policies have on data sharing rates. The dataset, perhaps supplemented with semi-structured interviews, could be used to understand the relationship between capabilities and inclinations for the data producing investigators.

6.2.3 The next frontier

This study has focused on data sharing. I plan to turn, next, to the study of data reuse. Who reuses data? When? Why? Who doesn’t? Which datasets are most likely to be reused? How many datasets could be reused but aren’t? Why aren’t they? What can we do about it? What should we do about it?

6.3 CODE AND DATA AVAILABILITY

The code and data behind this project are available at http://www.researchremix.org.

6.4 HOPE

I hope this research project will contribute to a deeper understanding of data sharing behavior and eventually more effective dissemination of research output. More generally, I hope this work facilitates and inspires an increased focus on using research methods to study and inform the practice of research. We owe it to ourselves as scientists, as tax-payers, and as patients to pursue biomedical research as effectively as possible. It is only by questioning our assumptions, considering alternatives, and evaluating our choices and results that we can choose methods and practices are most effective for achieving our desired outcomes.

APPENDIX

UNIVARIATE SHARING PATTERNS ON ORIGINAL VARIABLES

The appendix includes a 5-part figure (divided at page breaks) illustrating the association between the probability that a study shares its gene expression microarray dataset and each of our independent variables that describe the study environment.

Overall prevalence of data sharing was 25%. The frequency of data sharing is shown for each quartile for continuous variables. Horizontal lines illustrate 95% confidence intervals of the data sharing frequencies.

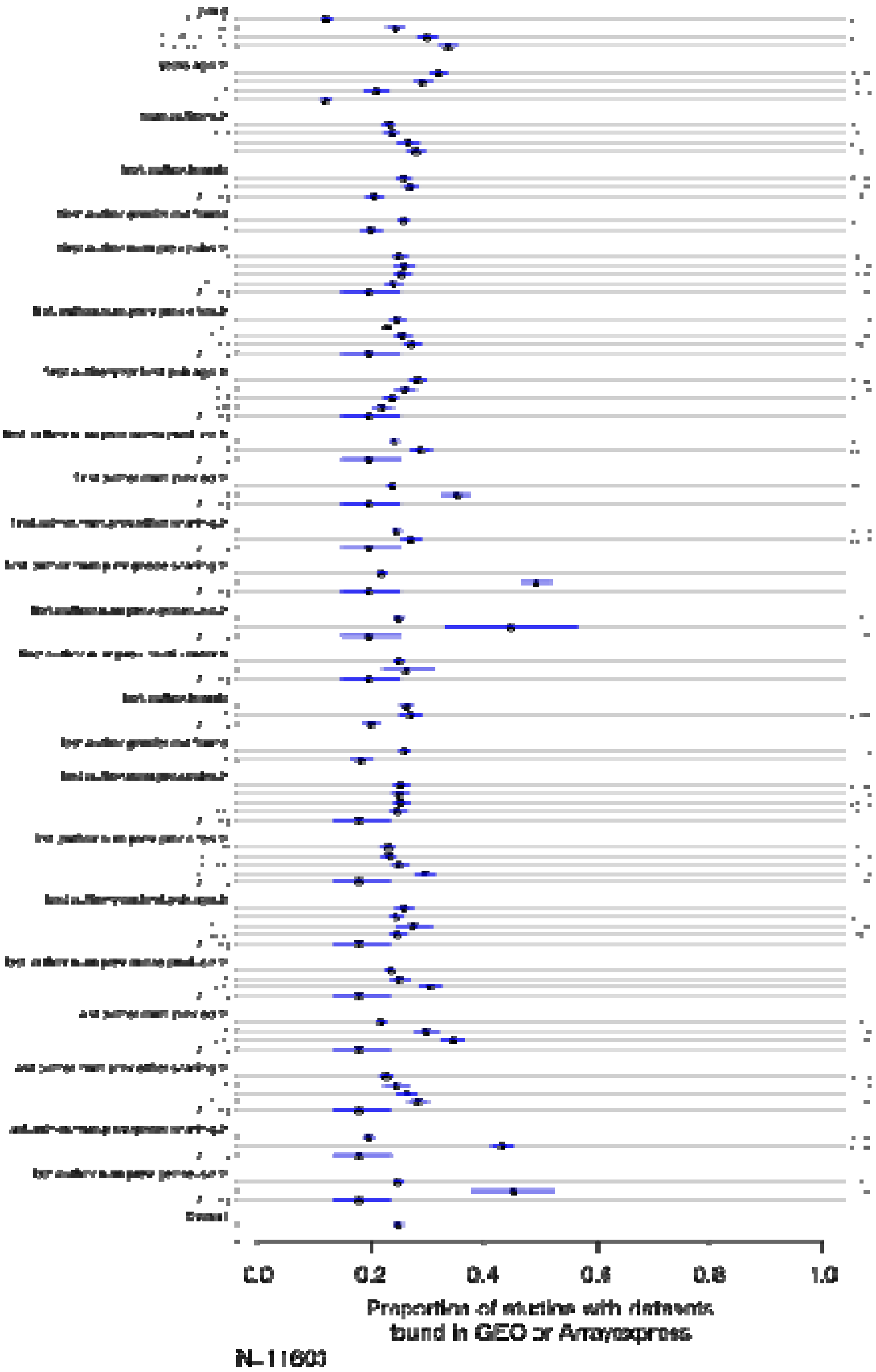


Figure 14: Association between shared data and original independent variables The frequency of data sharing is shown for each quartile for continuous variables. Horizontal lines illustrate 95% confidence intervals of the data sharing proportions.

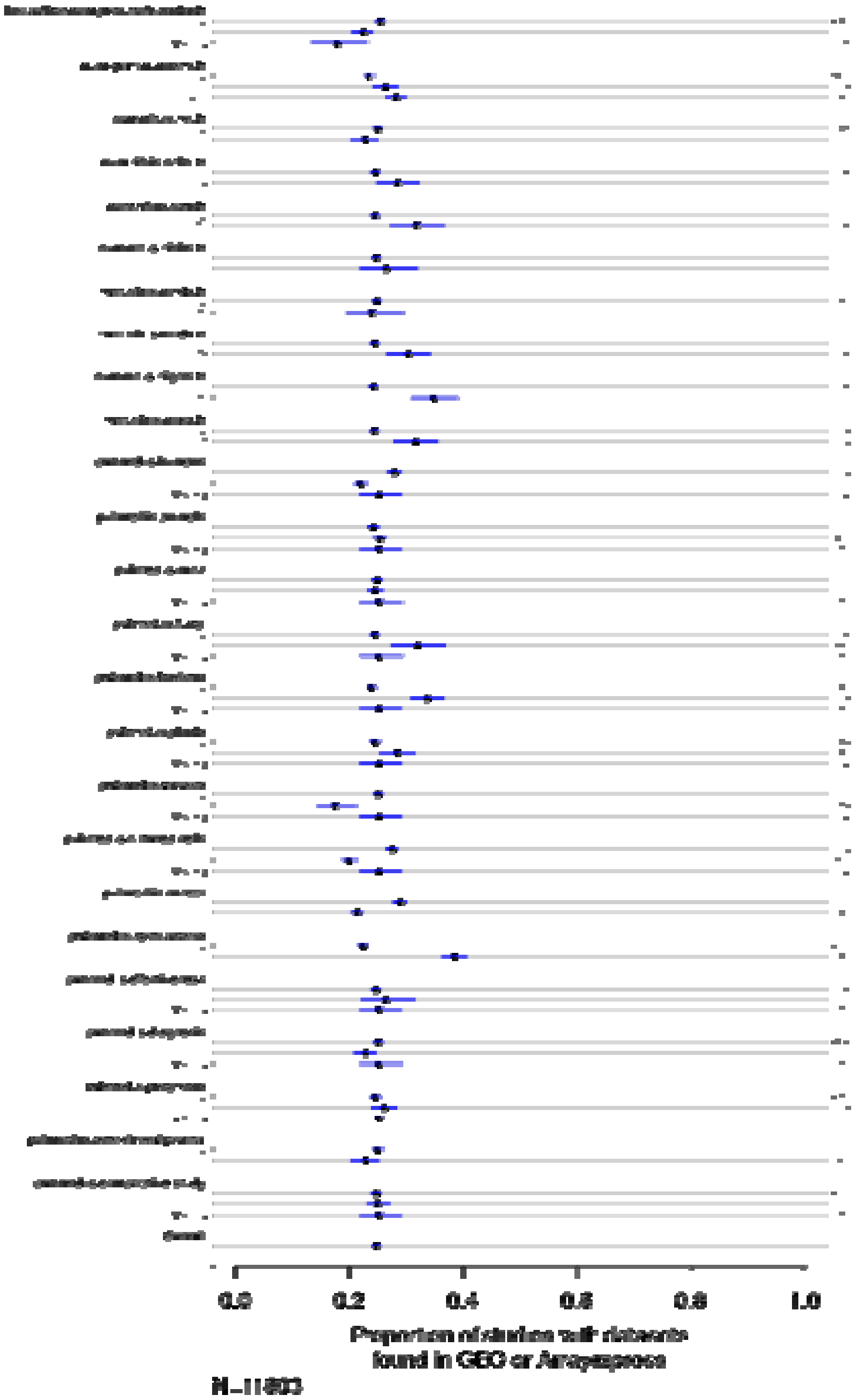


Figure 14 (continued)

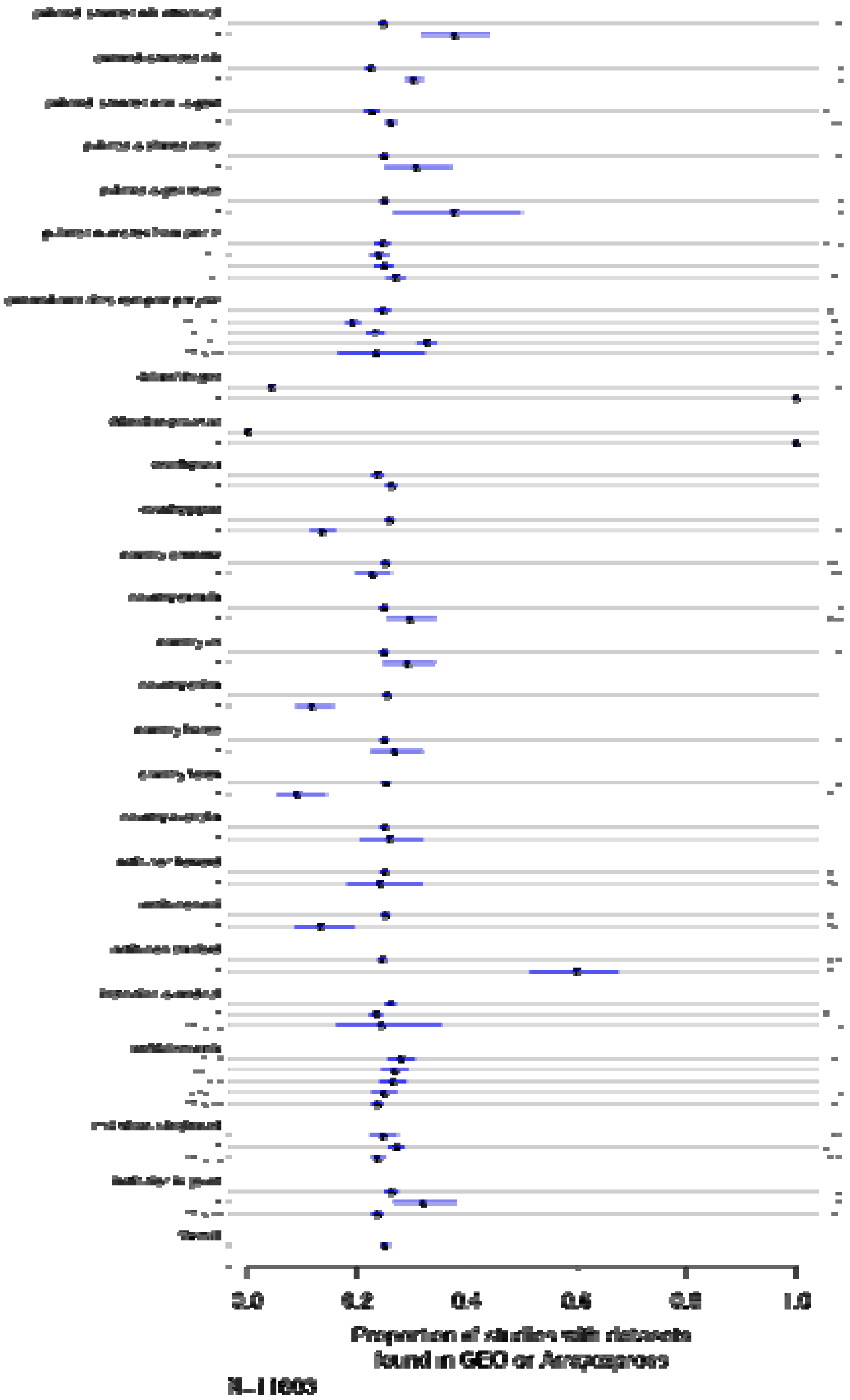


Figure 14 (continued)

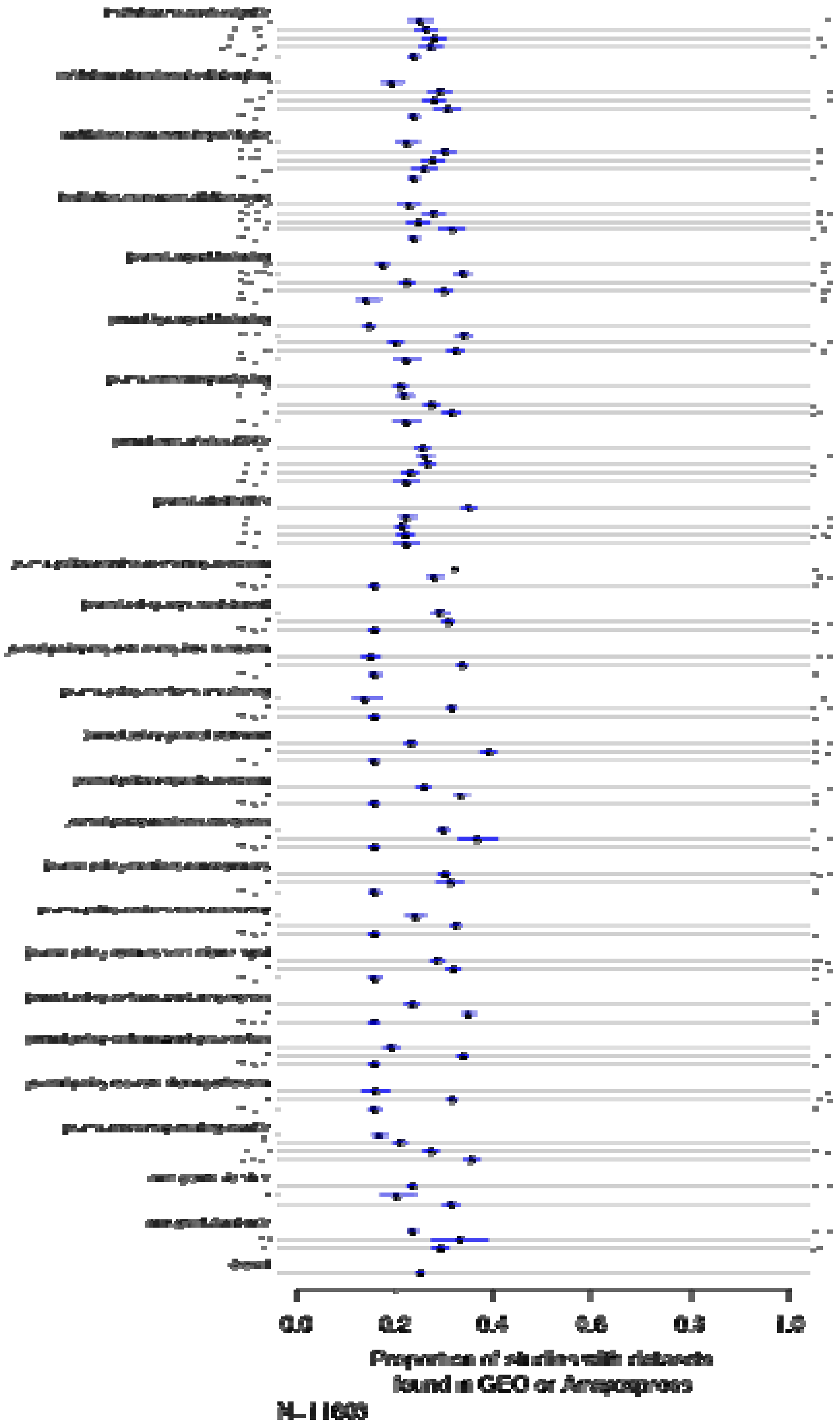


Figure 14 (continued)

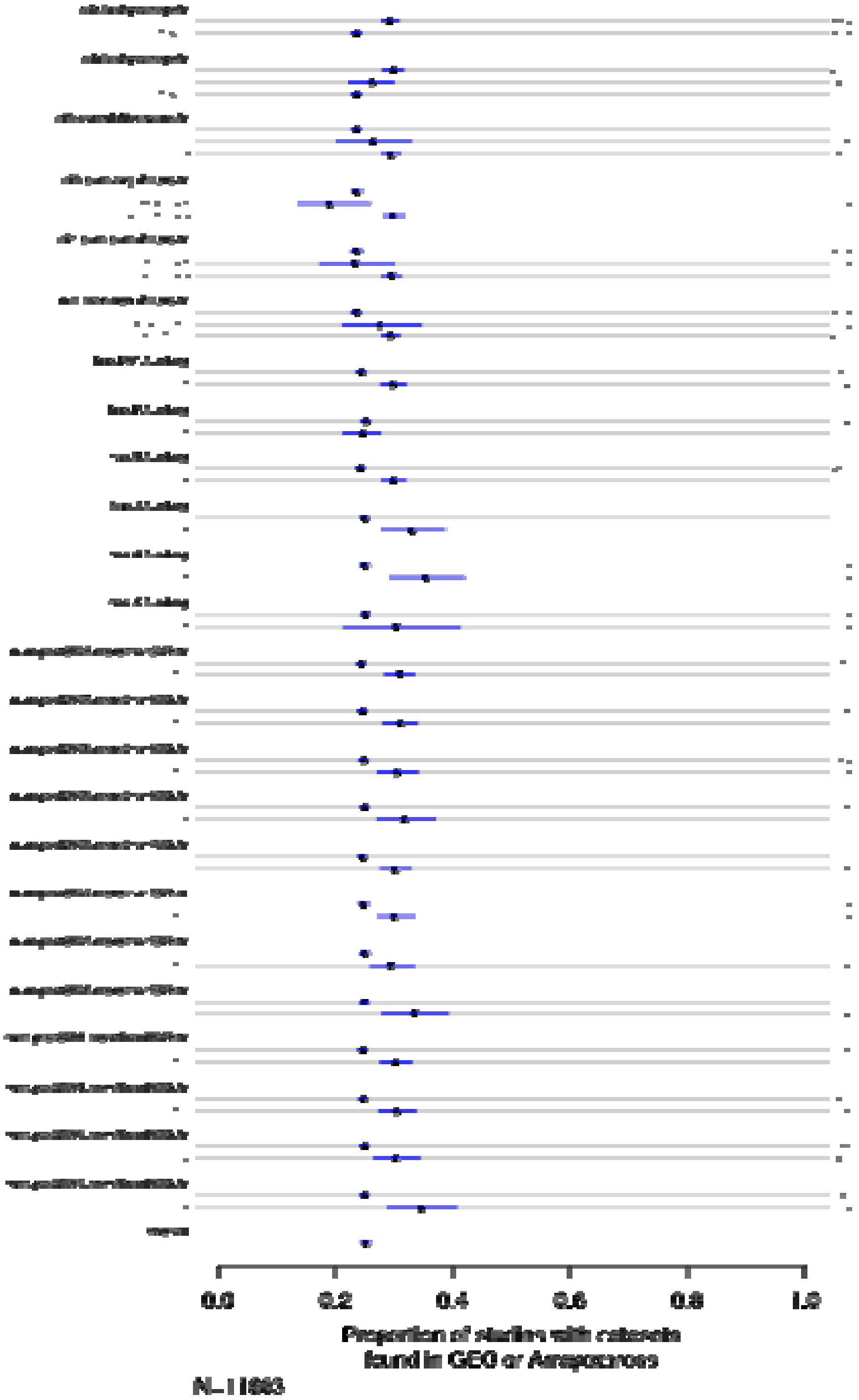


Figure 14 (continued)

BIBLIOGRAPHY

1. Merton RK: **The Sociology of Science: Theoretical and Empirical Investigations**. *booksgooglecom* 1973.

2. Gass A: **Open Access As Public Policy**. *PLoS Biology* 2004, **2**(10):e353.

3. Vickers A: **Whose data set is it anyway? Sharing raw data from randomized trials**. *Trials* 2006, **7**:15.

4. Santos C, Blake J, States D: **Supplementary data need to be kept in public repositories**. *Nature* 2005, **438**(7069):738.

5. Evangelou E, Trikalinos T, Ioannidis J: **Unavailability of online supplementary scientific information from articles published in major journals**. *FASEB J* 2005, **19**(14):1943-1944.

6. Wren JD: **URL decay in MEDLINE--a 4-year follow-up study**. *Bioinformatics* 2008, **24**(11):1381-1385.

7. Sullivan M: **Controversy Erupts Over Proteomics Studies**. *Ob Gyn News* 2005.

8. Liotta L, Lowenthal M, Mehta A, Conrads T, Veenstra T, Fishman D, Petricoin E: **Importance of communication between producers and consumers of publicly available experimental data**. *J Natl Cancer Inst* 2005, **97**(4):310-314.

9. Merton RK, Sills DL, Stigler SM: **The Kelvin dictum and social science: an excursion into the history of an idea**. *J Hist Behav Sci* 1984, **20**(4):319-331.

10. Ochsner SA, Steffen DL, Stoeckert CJ, McKenna NJ: **Much room for improvement in deposition rates of expression microarray datasets**. *Nature Methods* 2008, **5**(12):991.

11. Noor MA, Zimmerman KJ, Teeter KC: **Data Sharing: How Much Doesn't Get Submitted to GenBank?** *PLoS Biol* 2006, **4**(7).

12. Piwowar HA, Day RS, Fridsma DB: **Sharing detailed research data is associated with increased citation rate**. *PLoS ONE* 2007, **2**(3).

13. Blumenthal D, Campbell EG, Gokhale M, Yucel R, Clarridge B, Hilgartner S, Holtzman NA: **Data withholding in genetics and the other life sciences: prevalences and predictors**. *Acad Med* 2006, **81**(2):137-145.

14. Fienberg S, Martin M, Straf M: **Sharing research data**. Washington DC: National Academy Press; 1985.

15. McCain K: **Mandating Sharing: Journal Policies in the Natural Sciences**. *Science Communication* 1995, **16**(4):403-431.

16. Piwowar H, Chapman W: **A review of journal policies for sharing research data**. In: *ELPUB.* Toronto; 2008.

17. NIH: **NOT-OD-03-032: Final NIH Statement on Sharing Research Data**. In.; 2003.

18. NIH: **NOT-OD-08-013: Implementation Guidance and Instructions for Applicants: Policy for Sharing of Data Obtained in NIH-Supported or Conducted Genome-Wide Association Studies (GWAS)**. 2007.

19. Cech T: **Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences**; 2003.

20. Kakazu KK, Cheung LW, Lynne W: **The Cancer Biomedical Informatics Grid (caBIG): pioneering an expansive network of information and tools for collaborative cancer research**. *Hawaii Med J* 2004, **63**(9):273-275.

21. **New models of collaboration in genome-wide association studies: the Genetic Association Information Network**. *Nat Genet* 2007, **39**(9):1045-1051.

22. Mailman M, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L *et al*: **The NCBI dbGaP database of genotypes and phenotypes**. *Nat Genet* 2007, **39**(10):1181-1186.

23. Geschwind DH: **Sharing gene expression data: an array of options**. *Nat Rev Neurosci* 2001, **2**(6):435-438.

24. Rodriguez M, Bollen J, Van de Sompel H: **A Practical Ontology for the Large-Scale Modeling of Scholarly Artifacts and their Usage**. In: *International Conference on Digital Libraries.* Vancouver; 2007.

25. Martone ME, Gupta A, Ellisman MH: **E-neuroscience: challenges and triumphs in integrating distributed data from molecules to brains**. *Nat Neurosci* 2004, **7**(5):467-472.

26. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, Detmer DE, Expert P: **Toward a national framework for the secondary use of**

health data: an American Medical Informatics Association White Paper. *J Am Med Inform Assoc* 2007, **14**(1):1-9.

27. Zerhouni E: **Medicine. The NIH Roadmap**. *Science* 2003, **302**(5642):63-72.

28. Nass S, Stillman B: **Large-Scale Biomedical Science: Exploring Strategies for Future Research**: National Academy Press; 2003.

29. **The Cancer Biomedical Informatics Grid (caBIG): infrastructure and applications for a worldwide research community**. *Medinfo* 2007, **12**(Pt 1):330-334.

30. Grethe JS, Baru C, Gupta A, James M, Ludaescher B, Martone ME, Papadopoulos PM, Peltier ST, Rajasekar A, Santini S *et al*: **Biomedical informatics research network: building a national collaboratory to hasten the derivation of new understanding and treatment of disease**. *Stud Health Technol Inform* 2005, **112**:100-109.

31. Sinnott RO, Macdonald A, Lord PW, Ecklund D, Jones A. **Large-scale data sharing in the life sciences: Data standards, incentives, barriers and funding models.**

32. **NIH Data Sharing Policy and Implementation Guidance** [http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm]

33. **How to encourage the right behaviour**. *Nature* 2002, **416**:1.

34. Ball CA, Brazma A, Causton H, Chervitz S, Edgar R, Hingamp P, Matese JC, Parkinson H, Quackenbush J, Ringwald M *et al*: **Submission of microarray data to public repositories**. *PLoS Biol* 2004, **2**(9).

35. Geer RC, Sayers EW: **Entrez: Making use of its power**. *Briefings in Bioinformatics* 2003.

36. Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, Doherty D, Forsberg K, Gao Y, Kashyap V *et al*: **Advancing translational research with the Semantic Web**. *BMC Bioinformatics* 2007, **8**(Suppl 3).

37. Li K, Chen C, Wu T, Wen C, Tang CY: **BioPortal: A Portal for Deployment of Bioinformatics Applications on Cluster and Grid Environments**. *LECTURE NOTES IN COMPUTER SCIENCE* 2007.

38. Piwowar HA, Becich MJ, Bilofsky H, Crowley RS: **Towards a data sharing culture: recommendations for leadership from academic health centers**. *PLoS Medicine* 2008, **5**(9):e183.

39. Buetow K: **Cyberinfrastructure: Empowering a "Third Way" in Biomedical Research**. *Science* 2005, **308**(5723):821-824.

40. Butler D: **Data sharing: the next generation**. *Nature* 2007, **446**(7131):10-11.

41. Bradley J: **Open Notebook Science Using Blogs and Wikis**. *Available from Nature Precedings* 2007, http://dx.doi.org/10.1038/npre.2007.39.1.

42. **Social software**. *Nat Meth* 2007, **4**(3):189-189.

43. Altman M, King G: **A proposed standard for the scholarly citation of quantitative data**. *D-Lib Magazine* 2007, **13**(3/4).

44. Wren JD, Grissom JE, Conway T: **E-mail decay rates among corresponding authors in MEDLINE. The ability to communicate with and request materials from authors is being eroded by the expiration of e-mail addresses**. *EMBO Rep* 2006, **7**(2):122-127.

45. Foster M, Sharp R: **Share and share alike: deciding how to distribute the scientific and social benefits of genomic data**. *Nat Rev Genet* 2007, **8**(8):633-639.

46. Campbell P: **Controversial Proposal on Public Access to Research Data Draws 10,000 Comments**. *The Chronicle of Higher Education* 1999:A42.

47. Melton GB: **Must Researchers Share their Data?** *Law and Human Behavior* 1988, **12**(2):159-162.

48. Gleditsch NP, Metelitis C: **The Replication Debate**. *International Studies Perspectives* 2003, **4**(1):72-79.

49. McCullough BD, McGeary KA, Harrison TD: **Do Economics Journal Archives Promote Replicable Research?** *Canadian Journal of Economics* 2008 41(4):1496:1420.

50. Tucker J: **Motivating Subjects: Data Sharing in Cancer Research**. 2009:1-261.

51. Reidpath DD, Allotey PA: **Data sharing in medical research: an empirical investigation**. *Bioethics* 2001, **15**(2):125-134.

52. Campbell EG, Clarridge BR, Gokhale M, Birenbaum L, Hilgartner S, Holtzman NA, Blumenthal D: **Data withholding in academic genetics: evidence from a national survey**. *JAMA* 2002, **287**(4):473-480.

53. Ventura B: **Mandatory submission of microarray data to public repositories: how is it working?** *Physiol Genomics* 2005, **20**(2):153-156.

54. Lee C, Dourish P, Mark G: **The human infrastructure of cyberinfrastructure**. Computer Supported Cooperative Work 2006: 483-492.

55. Ryu S, Ho S, Han I: **Knowledge sharing behavior of physicians in hospitals**. *Expert Systems With Applications* 2003 25(1):113-122.

56. Seonghee K, Boryung J: **An analysis of faculty perceptions: Attitudes toward knowledge sharing and collaboration in an academic institution**. *Library* 2008, **30**(4):282-290.

57. **PDBj News Letter**. In: *Volume 7, March 2006* <*http://wwwpdbjorg/NewsLetter/newsletter_vol7_epdf>:* 2006.

58. Henrick K, Feng Z, Bluhm W, Dimitropoulos D, Doreleijers J, Dutta S, Flippen-Anderson J, Ionides J, Kamada C, Krissinel E *et al*: **Remediation of the protein data bank archive**. *Nucleic Acids Research* 2008, **36**(Database issue).

59. Plint AC, Moher D, Morrison A, Schulz K, Altman DG, Hill C, Gaboury I: **Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review**. *Med J Aust* 2006, **185**(5):263-267.

60. Pienta A: **1R01LM009765-01 Barriers and Opportunities for Sharing Research Data**. 2007.

61. Zimmerman A: **Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists**. 2003.

62. Eysenbach G: **Citation advantage of open access articles**. *PLoS Biol* 2006, **4**(5):e157.

63. Wren JD: **Open access and openly accessible: a study of scientific publications shared via the internet**. *Bmj* 2005, **330**(7500):1128.

64. McKechnie L, Goodall GR, Lajoie-Paquette D: **How human information behaviour researchers use each other's work: a basic citation analysis study**. *Information Research* 2005, 10(2).

65. Patsopoulos NA, Analatos AA, Ioannidis JP: **Relative citation impact of various study designs in the health sciences**. *Jama* 2005, **293**(19):2362-2366.

66. Lokker C, McKibbon A, McKinlay J, Wilczynski N, Haynes B: **Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study**. *BMJ* 2008, 336(7645).

67. Fu L, Aliferis C: **Models for predicting and explaining citation count of biomedical articles**. *AMIA Annual Symposium proceedings* 2008:222-226.

68. Torvik VI, Weeber M, Swanson DR, Smalheiser NR: **A probabilistic similarity metric for Medline records: A model for author name disambiguation**.

Journal of the American Society for Information Science and Technology 2005, 56(2):140-158.

69. Lautrup BE, Lehmann S, Jackson AD: **Measures for measures**. *Nature* 2006 444:1003-1004.

70. Hirsch JE: **An index to quantify an individual's scientific research output**. *Proceedings of the National Academy of Sciences* 2005 102(46):16569-16572.

71. Hendrix D: **An analysis of bibliometric indicators, National Institutes of Health funding, and faculty size at Association of American Medical Colleges medical schools, 1997-2007**. *Journal of the Medical Library Association : JMLA* 2008, **96**(4):324-334.

72. Adler R, Ewing J, Taylor P: **Citation Statistics**. *Statistical Science* 2009, 24(1):1-14.

73. Stringer M, Sales-Pardo M, Nunes Amaral L, Scalas E: **Effectiveness of Journal Ranking Schemes as a Tool for Locating Information**. *PLoS ONE* 2008, **3**(2):e1683.

74. Taylor M, Perakakis P, Trachana V: **The siege of science**. *ESEP* 2008, **8**:17-40.

75. Coleman A: **Assessing the value of a journal beyond the impact factor**. *Journal of the American Society for Information Science and Technology* 2007, **58**(8):1148-1161.

76. Rodriguez M, Bollen J, Van de Sompel H: **A Practical Ontology for the Large-Scale Modeling of Scholarly Artifacts and their Usage**. *International Conference on Digital Libraries* 2007, Vancouver.

77. Bakkalbasi N, Bauer K, Glover J, Wang L: **Three options for citation tracking: Google Scholar, Scopus and Web of Science**. *Biomedical Digital Libraries* 2006, **3**:7.

78. Eysenbach G, Trudel M: **Going, going, still there: using the WebCite service to permanently archive cited web pages**. *J Med Internet Res* 2005, **7**(5):e60.

79. West R, McIlwaine A: **What do citation counts count for in the field of addiction? An empirical evaluation of citation counts and their link with peer ratings of quality**. *Addiction* 2002, **97**(5):501-504.

80. Jensen L, Saric J, Bork P: **Literature mining for the biologist: from information retrieval to biological discovery**. *Nature Reviews Genetics* 2006, **7**(2):119-129.

81. Hearst M, Divoli A, Guturu H, Ksikes A, Nakov P, Wooldridge M, Ye J: **BioText Search Engine: beyond abstract search**. *Bioinformatics* 2007, **23**(16):2196-2197.

82. Karamanis N, Seal R, Lewin I, McQuilton P, Vlachos A, Gasperin C, Drysdale R, Briscoe T: **Natural Language Processing in aid of FlyBase curators**. *BMC Bioinformatics* 2008, **9**(1).

83. Siddharthan A, Teufel S: **Whose idea was this, and why does it matter? Attributing scientific work to citations**. In: *Proceedings of NAACL/HLT-07: 2007*; 2007.

84. Marco C, Kroon F, Mercer R: **Using Hedges to Classify Citations in Scientific Articles**. In: *Computing Attitude and Affect in Text: Theory and Applications.* 2006: 247-263.

85. Eales J, Pinney J, Stevens R, Robertson D: **Methodology capture: discriminating between the" best" and the rest of community practice**. *BMC Bioinformatics* 2008, 9:359.

86. Rekapalli HK, Cohen AM, Hersh WR: **A comparative analysis of retrieval features used in the TREC 2006 Genomics Track passage retrieval task**. *AMIA Annual Symposium proceedings* 2007:620-624.

87. Yoo S, Choi J: **Reflecting all query aspects on query expansion**. *AMIA Annual Symposium proceedings* 2008:1189.

88. Abdalla R, Teufel S: **A bootstrapping approach to unsupervised detection of cue phrase variants**. In: *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL: 2006*: Association for Computational Linguistics; 2006: 921-928.

89. Melton GB, Hripcsak G: **Automated detection of adverse events using natural language processing of discharge summaries**. *Journal of the American Medical Informatics Association : JAMIA* 2005, **12**(4):448-457.

90. Harder M: **How Do Rewards and Management Styles Influence the Motivation to Share Knowledge?** In: *Center for Strategic Management and Globalization.* SMG Working Paper; 2008.

91. Samieh H, Wahba K: **Knowledge Sharing Behavior From Game Theory And Socio-Psychology Perspectives**. Hawaii International Conference on System Sciences 2007.

92. Cabrera A, Collins W, Salgado J: **Determinants of individual engagement in knowledge sharing**. *The International Journal of Human Resource Management* 2006.

93. Siemsen E, Roth A, Balasubramanian S: **How motivation, opportunity, and ability drive knowledge sharing: The constraining-factor model**. *Journal of Operations Management* 2007.

94. Wstebro T, Michela J, Zhang X: **The Survival of Innovations: Patterns and Predictors**. *manuscript* 2001.

95. Kolekofski K: **Beliefs and attitudes affecting intentions to share information in an organizational setting**. *Information & Management* 2003, **40**(6):521-532.

96. The EMBL Data Library and GenBank(R) staff: **A new system for direct submission of data to the nucleotide sequence data banks**. *Nucleic Acids Research* 1987, **15**(18):front matter.

97. **Microarray policy**. *Nat Immunol* 2003, **4**(2):93.

98. Seglen P: **Why the impact factor of journals should not be used for evaluating research**. *BMJ* 1997, **314**(7079):498-502.

99. Diamond AM, J.: **What is a Citation Worth?** *The Journal of Human Resources* 1986, **21**(2):200-215.

100. Ntzani E, Ioannidis J: **Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment**. *Lancet* 2003, **362**(9394):1439-1444.

101. Sherlock G, Boussard H, Kasarskis A, Binkley G, Matese J, Dwight S, Kaloper M, Weng S, Jin H, Ball C *et al*: **The Stanford Microarray Database**. *Nucleic Acids Res* 2001, **29**(1):152-155.

102. Edgar R, Domrachev M, Lash A: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**. *Nucleic Acids Res* 2002, **30**(1):207-210.

103. Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, Coulson R, Farne A, Lara G, Holloway E, Kapushesky M *et al*: **ArrayExpress--a public repository for microarray gene expression data at the EBI**. *Nucleic Acids Res* 2005, **33**(Database issue):D553-555.

104. Ikeo K, Ishi-i J, Tamura T, Gojobori T, Tateno Y: **CIBEX: center for information biology gene expression database**. *C R Biol* 2003, **326**(10-11):1079-1082.

105. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: **ONCOMINE: a cancer microarray database and integrated data-mining platform**. *Neoplasia* 2004, **6**(1):1-6.

106. Weale A, Bailey M, Lear P: **The level of non-citation of articles within a journal as a measure of quality: a comparison to the impact factor**. *BMC Med Res Methodol* 2004, **4**:14.

107. Patsopoulos N, Analatos A, Ioannidis J: **Relative Citation Impact of Various Study Designs in the Health Sciences**. *JAMA: The Journal of the American Medical Association* 2005, **293**(19):2362.

108. R Development Core Team: **R: A Language and Environment for Statistical Computing**. In*.* Vienna, Austria: ISBN 3-900051-07-0; 2008.

109. Brazma A, Robinson A, Cameron G, Ashburner M: **One-stop shop for microarray data**. *Nature* 2000, **403**(6771):699-700.

110. Antelman K: **Do Open Access Articles Have a Greater Research Impact?** *College and Research Libraries* 2004, **65**(5):372-382.

111. Swan A, Brown S: **Authors and open access publishing**. *Learned Publishing* 2004, **17**(3):219-224.

112. Cases D, Higgins G: **How can we investigate citation behavior?: a study of reasons for citing literature in communication**. *J Am Soc Inf Sci* 2000, **51**(7):635-645.

113. Theologis A, Davis R: **To give or not to give? That is the question**. *Plant Physiol* 2004, **135**(1):4-9.

114. Popat S, Houlston R: **Re: Reporting recommendations for tumor marker prognostic studies (REMARK)**. *J Natl Cancer Inst* 2005, **97**(24):1855; author reply 1855-1856.

115. Ball C, Sherlock G, Brazma A: **Funding high-throughput data sharing**. *Nat Biotechnol* 2004, **22**(9):1179-1183.

116. Riley R, Abrams K, Sutton A, Lambert P, Jones D, Heney D, Burchill S: **Reporting of prognostic markers: current problems and development of guidelines for evidence-based practice in the future**. *Br J Cancer* 2003, **88**(8):1191-1198.

117. Kyzas P, Loizou K, Ioannidis J: **Selective reporting biases in cancer prognostic factor studies**. *J Natl Cancer Inst* 2005, **97**(14):1043-1055.

118. Check E: **Proteomics and cancer: Running before we can walk?** *Nature* 2004, **429**(6991):496.

119. Ball C, Awad I, Demeter J, Gollub J, Hebert J, Hernandez-Boussard T, Jin H, Matese J, Nitzberg M, Wymore F *et al*: **The Stanford Microarray Database**

accommodates additional microarray platforms and data formats. *Nucleic Acids Res* 2005, **33**(Database issue):D580-582.

120. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball C, Causton H *et al*: **Minimum information about a microarray experiment (MIAME)-toward standards for microarray data**. *Nat Genet* 2001, **29**(4):365-371.

121. McShane L, Altman D, Sauerbrei W, Taube S, Gion M, Clark G: **Reporting recommendations for tumor marker prognostic studies (REMARK)**. *J Natl Cancer Inst* 2005, **97**(16):1180-1184.

122. Spellman P, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M *et al*: **Design and implementation of microarray gene expression markup language (MAGE-ML)**. *Genome Biol* 2002, **3**(9):RESEARCH0046.

123. Gupta D, Saul M, Gilbertson J: **Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research**. *Am J Clin Pathol* 2004, **121**(2):176-186.

124. Matsubayashi M, Kurata K, Sakai Y, Morioka T, Kato S, Mine S, Ueda S: **Status of open access in the biomedical field in 2005**. *Journal of the Medical Library Association : JMLA* 2009, **97**(1):4-11.

125. Barrett T, Troup D, Wilhite S, Ledoux P, Rudnev D, Evangelista C, Kim I, Soboleva A, Tomashevsky M, Edgar R: **NCBI GEO: mining tens of millions of expression profiles--database and tools update**. *Nucleic Acids Res* 2007, **35**(Database issue).

126. Piwowar HA, Chapman WW: **Recall and bias of retrieving gene expression microarray datasets through PubMed identifiers**. *Discovery and Collaboration* 2010**, [accepted].**

127. NCBI: **PubMed Help, Stopwords**. *http://wwwncbinlmnihgov/bookshelf/brfcgi?book=helppubmed∂=pubmedhelp&rendertype=table&id=pubmedhelpT43 (Archived by WebCite at http://wwwwebcitationorg/5o3fDEbFh)* 2010.

128. Beall J: **The Weaknesses of Full-Text Searching**. *The Journal of Academic Librarianship* 2009, **34**(5):438-444.

129. Bernal-Delgado E, Fisher ES: **Abstracts in high profile journals often fail to report harm**. *BMC Medical Research Methodology* 2008, **8**(1):14.

130. Shah P, Perez-Iratxeta C, Bork P: **Information extraction from full text scientific articles: Where are the keywords?** *BMC Bioinformatics* 2003.

131. Schuemie M, Weeber M, Schijvenaars B, van Mulligen E, van der Eijk C, Jelier R, Mons B, Kors J: **Distribution of information in biomedical abstracts and full-text publications**. *Bioinformatics* 2004, **20**(16):2597-2604.

132. Hemminger B, Saelim B, Sullivan P, Vision T: **Comparison of full-text searching to metadata searching for genes in two biomedical literature cohorts**. *Journal of the American Society for Information Science and Technology* 2007, **58**(14):2341-2352.

133. Lin J: **Is searching full text more effective than searching abstracts?** *BMC Bioinformatics* 2009, **10**:46.

134. Muller H, Kenny E, Sternberg P: **Textpresso: an ontology-based information retrieval and extraction system for biological literature**. *PLoS Biol* 2004, **2**(11).

135. Garten Y, Altman RB: **Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text**. *BMC Bioinformatics* 2009, **10 Suppl 2**:S6.

136. Fink JL, Kushch, Sergey, Williams, Parker R, Bourne, Philip E: **BioLit: integrating biological literature with databases**. *Nucleic Acids Research* 2008, **36**(Web Servier issue):W385-W389.

137. Rubin D, Thorn C, Klein T, Altman R: **A statistical approach to scanning the biomedical literature for pharmacogenetics knowledge**. *J Am Med Inform Assoc* 2005, **12**(2):121-129.

138. Poulter GL, Rubin DL, Altman RB, Seoighe C: **MScanner: a classifier for retrieving Medline citations**. *BMC Bioinformatics* 2008, **9**(1):108.

139. PubMed Central: **PMC Open Access Subset**. *http://wwwncbinlmnihgov/pmc/about/openftlisthtml (Archived by WebCite at http://wwwwebcitationorg/5o3elVXa9)* 2009.

140. Verspoor K, Cohen KB, Hunter L: **The textual characteristics of traditional and Open Access scientific journals are similar**. *BMC Bioinformatics* 2009, **10**(1):183.

141. PubMed Central: **PubMed Central Journals.** [http://www.ncbi.nlm.nih.gov/pmc/journals/ (Archived by WebCite at http://www.webcitation.org/5lcmBT8aU)]

142. Wu T, Pottenger W: **A semi-supervised active learning algorithm for information extraction from textual data**. *Journal of the American Society for Information Science and Technology* 2005, **56**(3):258-271.

143. Carpenter B: **Phrasal queries with LingPipe and Lucene: ad hoc genomics text retrieval**. *NIST Special Publication: SP* 2004:500-261.

144. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis C: **Text categorization models for high-quality article retrieval in internal medicine**. *J Am Med Inform Assoc* 2005, **12**(2):207-216.

145. Taylor C, Field, D, Sansone, SA, Aerts, J, Apweiler, R, Ashburner, M, Ball, CA, Binz, PA, Bogue, M, Booth, T: **Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project**. *Nature Biotechnology* 2008, **26**(8):889-896.

146. Shah NH, Rubin DL, Espinosa I, Montgomery K, Musen MA: **Annotation and query of tissue microarray data using the NCI Thesaurus**. *BMC Bioinformatics* 2007, **8**:296.

147. Butte AJ, Chen R: **Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics**. *AMIA Annu Symp Proc* 2006:106-110.

148. Williams-Devane C, Wolf M, Richard A: **Towards a public toxicogenomics capability for supporting predictive toxicology: Survey of current resources and chemical indexing of experiments in GEO and ArrayExpress**. *Toxicol Sci* 2009.

149. Dudley J, Butte AJ: **Enabling integrative genomic analysis of high-impact human diseases through text mining**. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing* 2008:580-591.

150. Lin Y-A, Chiang A, Lin R, Yao P, Chen R, Butte AJ: **Methodologies for extracting functional pharmacogenomic experiments from international repository**. *AMIA Annual Symposium proceedings* 2007:463-467.

151. Djebbari A, Karamycheva S, Howe E, Quackenbush J: **MeSHer: identifying biological concepts in microarray assays based on PubMed references and MeSH terms**. *Bioinformatics* 2005, **21**(15):3324-3326.

152. Butte AJ, Kohane IS: **Creation and implications of a phenome-genome network**. *Nature Biotechnology* 2006, **24**(1):55-62.

153. Korbel J, Doerks T, Jensen L, Perez-Iratxeta C, Kaczanowski S, Hooper S, Andrade M, Bork P: **Systematic Association of Genes to Phenotypes by Genome and Literature Mining**. *PLoS Biol* 2005, **3**(5).

154. Scherf M, Epple A, Werner T: **The next generation of literature analysis: integration of genomic analysis into text mining**. *Brief Bioinform* 2005, **6**(3):287-297.

155. Tanabe L, Scherf U, Smith L, Lee J, Hunter L, Weinstein J: **MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling**. *Biotechniques* 1999, **27**(6):1210-1214, 1216-1217.

156. **Free PubMed Facelifts: Alternative Interfaces to an Essential Database**. In: *University of Manitoba Info-Rx Newsletter.* vol. March 27, 2007: URL:http://myuminfo.umanitoba.ca/index.asp?sec=857&too=100&dat=3/26/2007&sta=3&wee=5&eve=8&npa=12437. (Archived by WebCite at http://www.webcitation.org/5ibUYdS3X); 2007.

157. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M *et al*: **ArrayExpress--a public database of microarray experiments and gene expression profiles**. *Nucleic Acids Res* 2007, **35**(Database issue).

158. **EUtils Python library** [http://sourceforge.net/projects/eutils/]

159. Morales M, R Development Core Team, R-help listserv community and especially Duncan Murdoch: **sciplot: Scientific Graphing Functions for Factorial Designs**.

160. Harrell FE, contributions from many other users: **Hmisc: Harrell Miscellaneous**. 2007.

161. Bhat T, Bourne P, Feng Z, Gilliland G, Jain S, Ravichandran V, Schneider B, Schneider K, Thanki N, Weissig H *et al*: **The PDB data uniformity project**. *Nucleic Acids Res* 2001, **29**(1):214-218.

162. **Compete, collaborate, compel**. *Nat Genet* 2007, **39**(8):931.

163. Piwowar H, Day R, Fridsma D: **Sharing detailed research data is associated with increased citation rate**. *PLoS ONE* 2007, **2**(3):e308.

164. NIH: **NOT-OD-08-013: Implementation Guidance and Instructions for Applicants: Policy for Sharing of Data Obtained in NIH-Supported or Conducted Genome-Wide Association Studies (GWAS)**. 2007.

165. **Time for leadership**. *Nat Biotech* 2007, **25**(8):821-821.

166. **Got data?** *Nat Neurosci* 2007, **10**(8):931-931.

167. The GAIN Collaborative Research Group: **New models of collaboration in genome-wide association studies: the Genetic Association Information Network**. *Nat Genet* 2007, **39**(9):1045-1051.

168. Hedstrom M: **Producing Archive-Ready Datasets: Compliance, Incentives, and Motivation**. *IASSIST Conference 2006: Presentations* 2006.

169. Giordano, R: **The Scientist: Secretive, Selfish, or Reticent? A Social Network Analysis.** *e-Social Science 2007*.

170. Hedstrom M, Niu J: **Research Forum Presentation: Incentives to Create "Archive-Ready" Data: Implications for Archives and Records Management**. *Society of American Archivists Annual Meeting* 2008.

171. Niu J: **Incentive study for research data sharing. A case study on NIJ grantees,** icd.si.umich.edu/twiki/pub/ICD/LabGroup/fieldpaper_6_25.pdf

172. Lowrance W: **Access to Collections of Data and Materials for Heath Research: A report to the Medical Research Council and the Wellcome Trust**. 2006.

173. University of Nottingham: **JULIET: Research funders' open access policies**. In*.*

174. 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* 2003, **54**(10):926-938.

175. McCullough BD, McGeary KA, Harrison TD: **Do Economics Journal Archives Promote Replicable Research?** *Canadian Journal of Economics* 2008, **41**(4):1406-1420.

176. Ryu S, Ho SH, Han I: **Knowledge sharing behavior of physicians in hospitals**. *Expert Systems With Applications* 2003, 25(1):113-122.

177. Bitzer J, Schrettl W, Schröder PJH: **Intrinsic motivation in open source software development**. *Journal of Comparative Economics* 2007.

178. Kim J: **Motivating and Impeding Factors Affecting Faculty Contribution to Institutional Repositories**. *Journal of Digital Information* 2007, **8**(2).

179. Warlick S, Vaughan K: **Factors influencing publication choice: why faculty choose open access**. *Biomed Digit Libr* 2007, **4**(1):1.

180. Kuo F, Young M: **A study of the intention–action gap in knowledge sharing practices**. *Journal of the American Society for Information Science and Technology* 2008, **59**(8):1224-1237.

181. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: **Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression**. *Proc Natl Acad Sci U S A* 2004, **101**(25):9309-9314.

182. Hrynaszkiewicz I, Altman D: **Towards agreement on best practice for publishing raw clinical trial data**. *Trials* 2009, **10**(1):17.

183. **Microarray standards at last**. *Nature* 2002, **419**(6905):323.

184. Yu W, Yesupriya A, Wulf A, Qu J, Gwinn M, Khoury MJ: **An automatic method to generate domain-specific investigator networks using PubMed abstracts**. *BMC medical informatics and decision making* 2007, **7**:17.

185. Torvik V, Smalheiser N: **Author Name Disambiguation in MEDLINE**. *Transactions on Knowledge Discovery from Data* 2009:3(3):11.

186. Bird S, Loper E: **Natural Language Toolkit**. 2006, http://nltk.sourceforge.net/.

187. Theus M, Urbanek S: **Interactive Graphics for Data Analysis: Principles and Examples (Computer Science and Data Analysis)**: Chapman & Hall/CRC; 2008.

188. Harrell FE: **Regression Modeling Strategies**: Springer; 2001.

189. Gorsuch RL: **Factor Analysis, Second Edition**: Psychology Press; 1983.

190. Siemsen E, Roth A, Balasubramanian S: **How motivation, opportunity, and ability drive knowledge sharing: The constraining-factor model**. *Journal of Operations Management* 2008, **26**(3):426-445.

191. Malin B, Karp D, Scheuermann RH: **Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research**. *J Investig Med* 2010, **58**(1):11-18.

192. Navarro R: **An ethical framework for sharing patient data without consent**. *Inform Prim Care* 2008, **16**(4):257-262.

193. Blumenthal D, Campbell E, Anderson M, Causino N, Louis K: **Withholding research results in academic life science. Evidence from a national survey of faculty**. *JAMA* 1997, **277**(15):1224-1228.

194. Vogeli C, Yucel R, Bendavid E, Jones L, Anderson M, Louis K, Campbell E: **Data withholding and the next generation of scientists: results of a national survey**. *Acad Med* 2006, **81**(2):128-136.

195. Piwowar HA, Chapman WW: **Public Sharing of Research Datasets: A Pilot Study of Associations**. *Journal of Informetrics* 2010, 4(2):148-156.

196. **Gender Differences in Major Federal External Grant Programs.** The RAND Coproration; 2005.

197. Bornmann L, Mutz R, Daniel H-D: **Do we need the hindex and its variants in addition to standard bibliometric measures?** *Journal of the American Society for Information Science and Technology* 2009, **60**(6):1286-1289.