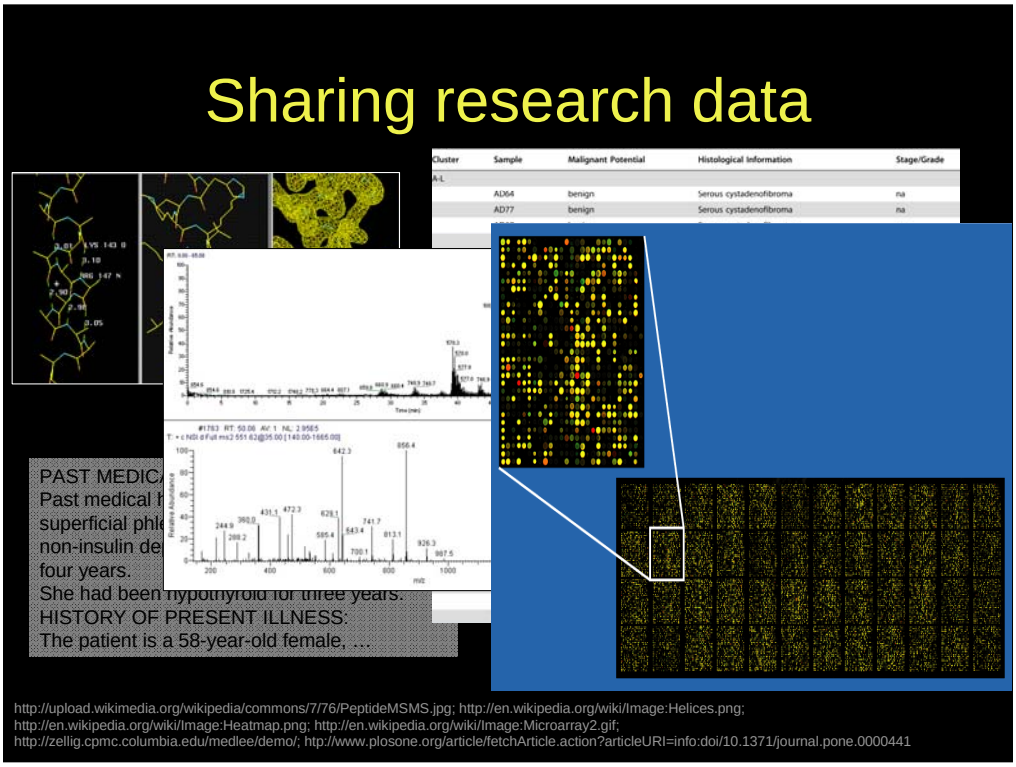
Sharing Detailed Research Data is Associated with Increased Citation Rate

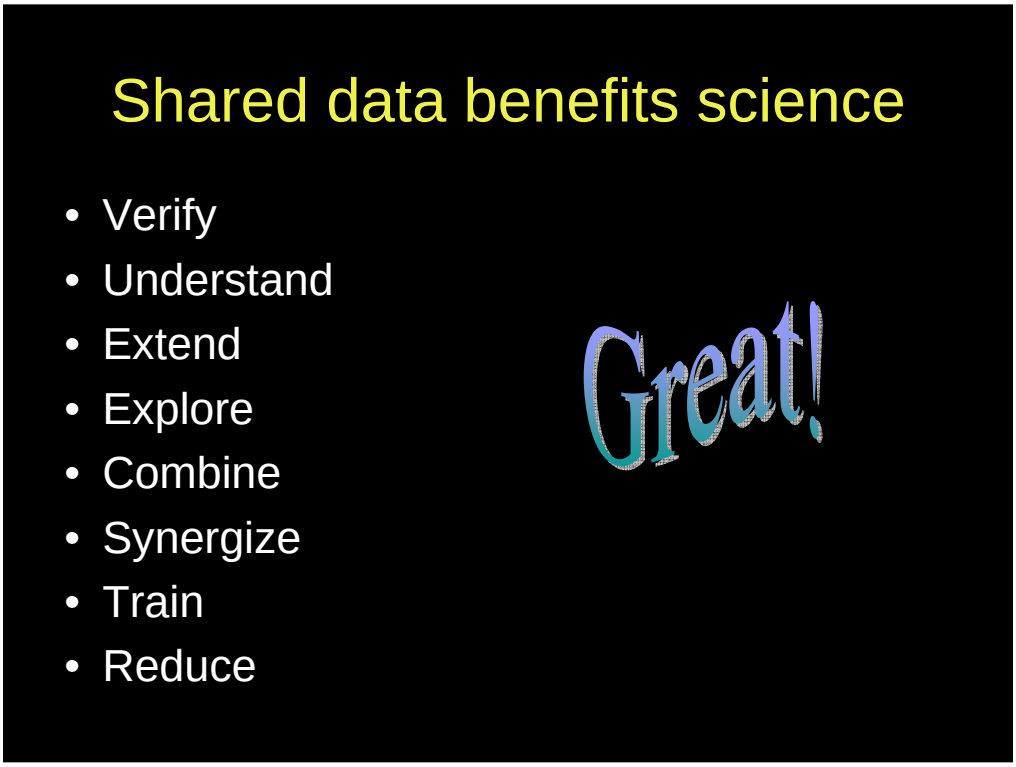
Heather Piwowar
Roger Day and Douglas Fridsma
University of Pittsburgh

Published in PLoS ONE, March 27 2007
Funded by NLM Training Grant
Presented at NLM Trainee Conference, June 27 2007

Hello everyone. My name is Heather Piwowar, and I'm from the University of Pittsburgh. Today I'm going to be talking about sharing research data.



Authors have a choice. When scientists like us do research and we collect or compute individual data points – perhaps 3d protein coordinates, or de-identified hospital reports, or patient clinical trial covariates or protein spectra or gene expression microarray values – we have a choice. We can make these detailed primary data points available when we publish our research, or we can keep them to ourselves.



Whenever we choose to share, the whole scientific community benefits. Many of the opportunities fo reuising data are obvious, but let me recount a few since it isn't something that most of us spend time thinking about.

Verify

Understand

Extend

Explore

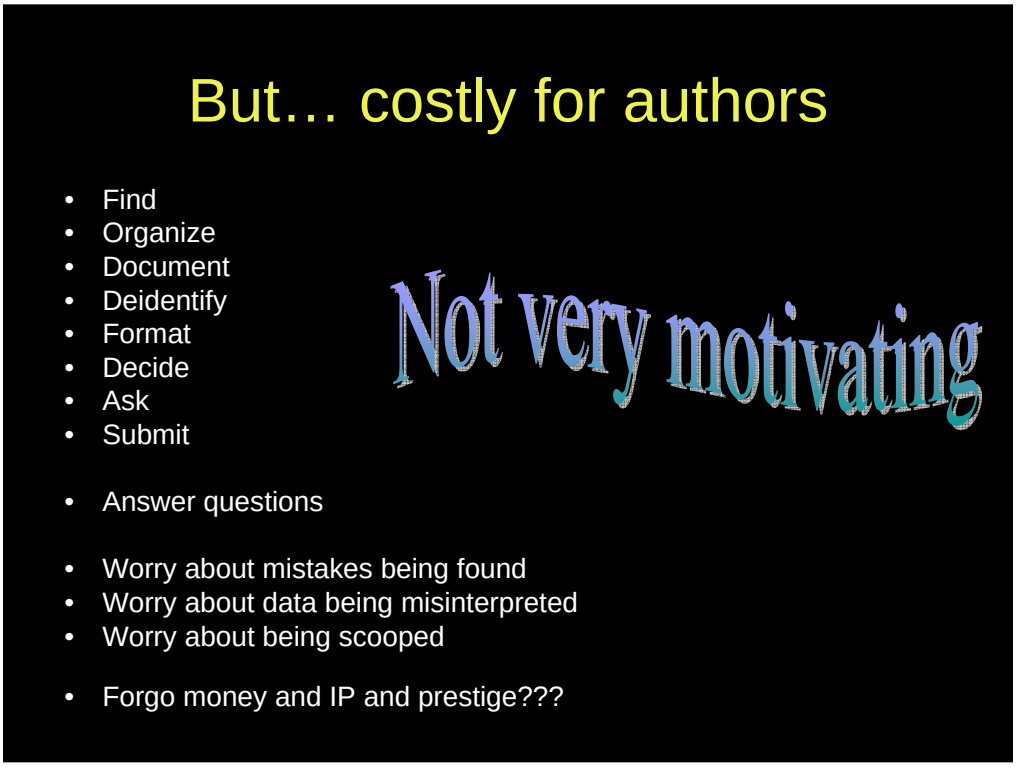
Combine

Synergize

Train

Reduce resource use and fraud

So that is great. But.



Except for helping to fund and maintain some public databases,almost all of the COST of sharing data isn't born by te whole community. The costs are felt by the authors each time they share their data. **The steps I'm going to outline will feel very familiar to those** of you who have made your data sets publicly available in the past.

Authors have to find their data, which in our busy and disorganized lives isn't always as easy as it should be. Then we have to format them, document, ….

Find

Organize

Document

Deidentify

Format

Decide

Ask *patients, IRBs, funders, co-authors*

Submit

Answer questions

Concurrently with this, we are maybe

Worry about mistakes being found

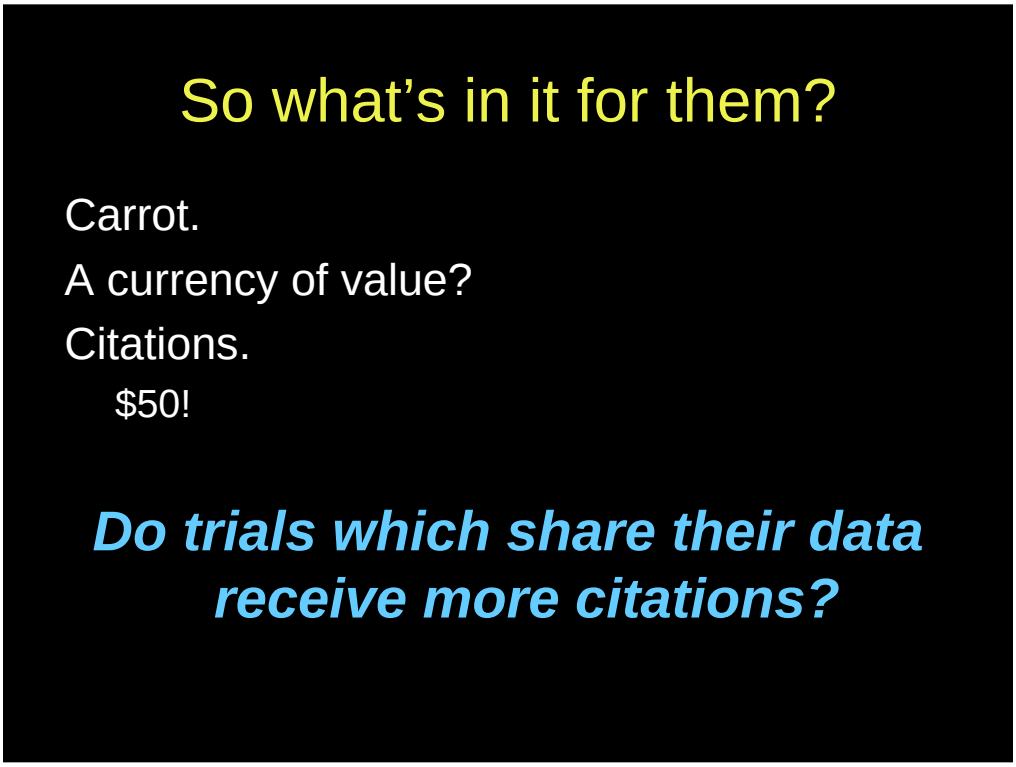
Worry about data being misinterpreted

Worry about being scooped

And potentially afraid

Forgo money and *intellctual property* and prestige???

Not very motivating.



So what is in it for the authors? What are some carrots we can offer, in addition to altruism? There are sticks already – some funders including the NIH and some journals, particularly Nature and Science, make it mandatory to share some types of data. But if we want to appeal to human nature and offer incentives, what can we do?

One idea is that Academics value citations. Although they are imperfect as measures of scientific value , they are nonetheless used as a proxy for scientific contribution. In fact, a study done in the 1980s looked at the correlation between promotions and academic salary increases. They estimate that each citation indirectly led to a $50 raise.

Believe that or not, but increasing citations has indeed been identified as a motivator for people to publish in open access journals, and we believe it may also encourage authors to make data available.

Therefore, this study addresses the question:

Do trials which share their data receive more citations?

Methods

Cancer Microarray Trials

Ntzani and Ioannidis identified 85 trials published 1999-2003

Citations

ISI Web of Science Citation Index, citations from 2004-2005

Data availability

Publisher and lab websites, microarray databases, WayBack Internet Archive, Oncomine

Statistics

Multivariate linear regression

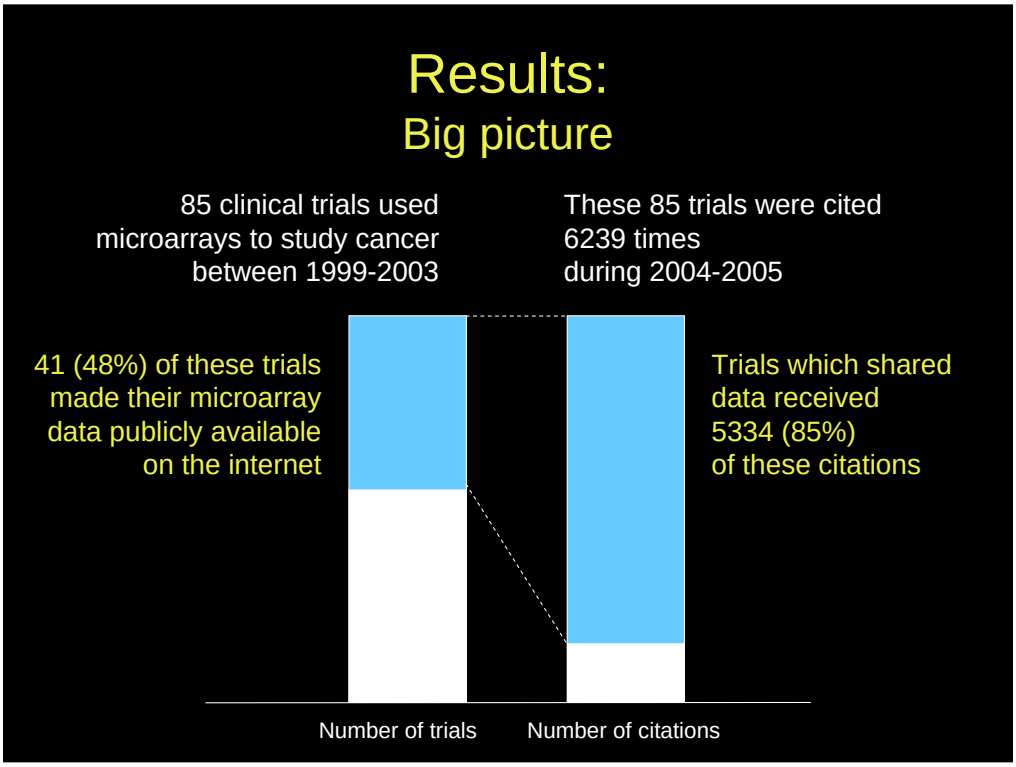
Briefly, to answer this question, we followed the citation history of 85 microarray trials. These publications included all of the studies which tried to associate microarray gene expression with cancer outcomes between 1999 and 2003. We scoured the internet for the datasets, and found the Oncomine database at the University of Michigan particularily helpful. Finally, we used multivariate linear regression to investigate the relationship between data availablility and number of citations, independently of other factors which are known to affect citation rate.

Results:
Eligible trials

- 85 trials
- 41 (48%) made data available
- Various locations:
  - Lab websites (28)
  - Publisher websites (4)
  - SMD (6)
  - GEO (6)
  - GEDP (2)
- 6239 total citations

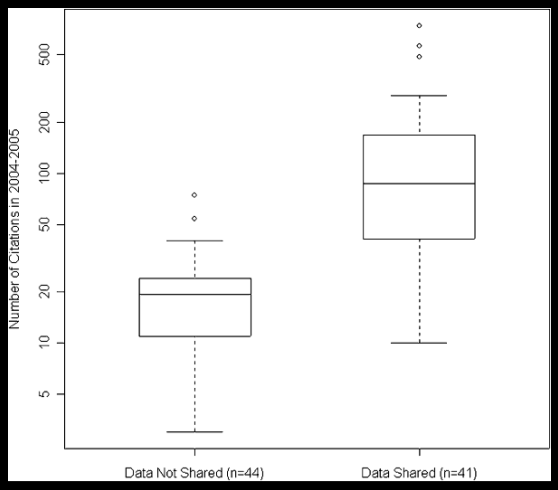
Here's what we found. Of the 85 trials, 41 made their data available. Most datasets were found on lab websites, but some were also as supplementary info in journal websites or at the Stanford Microarray Database, The Gene Expression Omnibus… you can see a link on Pubmed to GEO for the papers which have data in that database, or other centralized datastores.

There were over 6000 total citations.



And, indeed, the 48% of trials which made their data available received 85% of the total citations… clearly more than their fare share.

Results:
Distribution of citation counts



From: Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308

A more detailed view can be seen on this diagram. The publications which did not make their data available had a median of 20 citations, whereas on the right, the publications which did make their data available had a median of about 100 citations. The Y axis is a log scale, in case you can't see that in the back.

Results:
Multivariate regression

	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

From: Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308

Since it is possible that perhaps all of the publications which made their data available were also published in big-name journals and thus the higher number of citations is simply due to journal prestige rather than the data availability, we did a multivariate regression including journal impact as well as other factors. The results were that making data available was associated with a 70% increase in the number of citations, independently of other covariates.

Limitations

- Outliers
  - Subset analysis of lower profile papers
- Complex timing
  - Additional analysis of citations within 24 months
- Association does not imply causation
  - Could be common cause

It is of course necessary to mention some of the limitations of this study. Our cohort included the Golub dataset. Has anybody heard of that? Yup. It is famous. They made their data available, and many people use it. They were the first ones to use microarrays with clinical cancer outcomes, and as such they have received a phenominial amount of citations, many more than normal papers do. Because of this and other outliers in our sample, we did perform a subset analysis on the lower-profile papers and our results were similar to those I've shown.

A second limitation comes from complex timing. The microarray field has been growing and maturing over the last 8 years. A paper published today receives many more citations than one published in 1999 simply because more scientists are working in this area. At the same time, any given paper has its own natural citation trajectory in time: it takes a few months before it receives any citations, and after a few years its citations begin to dwindle again. Our primary analysis looked at citations in 2004 and 2005, but because of this complex timing, we also had a look at citations within 24 months of each paper publication. Again, the results were similar.

Finally and most importantly, as we all know, association does not imply causation. An association may instead be due, for example, to a common cause. It is possible that large, well-funded, clinically relevant trials are more apt to share their data because they are relatively well funded, however their large number of citations have nothing to do with data availability but are rather because the trial addresses an important issue. Nonetheless, we speculate that a number of the citations are indeed caused by the data sharing, but that hasn't been shown as part of this study.

Data sharing help on the way

- Free, centralized databases
  - SMD, GEO, ArrayExpress
- Standards
  - MIAME, CONSORT
- Tools
  - De-id, caBIG
- Community
  - Journals, Funders, Organizations, Blogs

In the interest of time, I'll let you read for yourselves about tools and initiatives underway to make data sharing easier for authors.

Conclusions

- 70% increase in citation impact for trials which make data available
- Result holds for lower-profile publications
- Hopefully a motivation for authors to share data and thus maximize its usefulness

So, in conclusion, we have found that sharing raw research datapoints is associated with about a 70% increase in citations, a currency of value to authors. This is the first study to demonstrate such an association, and we hope it will provide a motivation for researchers to share data with one another.

For more information

- Participate in the discussion on this paper at PLoS ONE
- Check out blogs on Open Access, Open Data, Open Notebook Science
  - Peter Suber's Open Access News blog
  - Wikipedia: "Open Data"
  - Nature Editorial: May 3, 2007
- Contact Heather Piwowar for further discussion and enthusiasm! hpiwowar@alumni.pitt.edu

If you would like more information on this topic, please have a look at our published paper at PLoS ONE (which is a great open access journal, by the way), participate in the ongoing dialog about open data and open science, and please contact me.

Thank you

- Peter Suber's blog: "Open Access News"
- Wikipedia: "Open Data"
- Nature Editorial: May 3, 2007

I support Open Data
and share my literature, code, and data whenever possible.

Long term research interest:
data reuse as an underutilized informatics resource

Questions?

I do appreciate your attention during this last session of our conference. I'd like to thank the NLM for its generous funding, and also thanks to each and every one of you who have made your detailed research data available in the past, or will do so in the future.