Interview With a Data Sucker

07 Sep 2016

By Rafa Irizarry

A few months ago Jill Sederstrom from ASH Clinical News interviewed me for this article on the data sharing editorial published by the The New England Journal of Medicine (NEJM) and the debate it generated. The article presented a nice summary, but I thought the original comprehensive set of questions was very good too. So, with permission from ASH Clinical News, I am sharing them here along with my answers.

Before I answer the questions below, I want to make an important remark. When writing these answers I am reflecting on data sharing in general. Nuances arise in different contexts that need to be discussed on an individual basis. For example, there are different considerations to keep in mind when sharing publicly funded data in genomics (my field) and sharing privately funded clinical trials data, just to name two examples.

In your opinion, what do you see as the biggest pros of data sharing?

The biggest pro of data sharing is that it can accelerate and improve the scientific enterprise. This can happen in a variety of ways. For example, competing experts may apply an improved statistical analysis that finds a hidden discovery the original data generators missed. Furthermore, examination of data by many experts can help correct errors missed by the analyst of the original project. Finally, sharing data facilitates the merging of

datasets from different sources that allow discoveries not possible with just one study.

Note that data sharing is not a radical idea. For example, thanks to an organization called The MGED Soceity, most journals require all published microarray gene expression data to be public in one of two repositories: GEO or ArrayExpress. This has been an incredible success, leading to new discoveries, new databases that combine studies, and the development of widely used statistical methods and software built with these data as practice examples.

The NEJM editorial expressed concern that a new generation of researchers will emerge, those who had nothing to do with collecting the research but who will use it to their own ends. It referred to these as "research parasites." Is this a real concern?

Absolutely not. If our goal is to facilitate scientific discoveries that improve our quality of life, I would be much more concerned about "data hoarders" than "research parasites". If an important nugget of knowledge is hidden in a dataset, don't you want the best data analysts competing to find it? Restricting the researchers who can analyze the data to those directly involved with the generators cuts out the great majority of experts.

To further illustrate this, let's consider a very concrete example with real life consequences. Imagine a loved one has a disease with high mortality rates. Finding a cure is possible but only after analyzing a very very complex genomic assay. If some of the best data analysts in the world want to help, does it make any sense at all to restrict the pool of analysts to, say, a freshly minted masters level statistician working for the genomics core that generated the data? Furthermore, what would be the harm of having someone double check that analysis?

The NEJM editorial also presented several other

concerns it had with data sharing including whether researchers would compare data across clinical trials that is not in fact comparable and a failure to provide correct attribution. Do you see these as being concerns? What cons do you believe there may be to data sharing?

If such mistakes are made, good peer reviewers will catch the error. If it escapes peer review, we point it out in post publication discussions. Science is constantly self correcting.

Regarding attribution, this is a legitimate, but in my opinion, minor concern. Developers of open source statistical methods and software see our methods used without attribution quite often. We survive. But as I elaborate below, we can do things to alleviate this concern.

Is data stealing a real worry? Have you ever heard of it happening before?

I can’t say I can recall any case of data being stolen. But let’s remember that most published data is paid for by tax payers. They are the actual owners. So there is an argument to be made that the public’s data is being held hostage.

Does data sharing need to happen symbiotically as the editorial suggests? Why or why not?

I think symbiotic sharing is the most effective approach to the repurposing of data. But no, I don’t think we need to force it to happen this way. Competition is one of the key ingredients of the scientific enterprise. Having many groups competing almost always beats out a small group of collaborators. And note that the data generators won’t necessarily have time to collaborate with all the groups interested in the data.

In a recent blog post, you suggested several possible data sharing guidelines. What would the advantage be of having guidelines in place in help guide the data sharing process?

I think you are referring to a post by Jeff Leek but I am happy to answer. For data to be generated, we need to incentivize the endeavor. Guidelines that assure patient privacy should of course be followed. Some other simple guidelines related to those mentioned by Jeff are:

1. Reward data generators when their data is used by others.
2. Penalize those that do not give proper attribution.
3. Apply the same critical rigor on critiques of the original analysis as we apply to the original analysis.
4. Include data sharing ethics in scientific education

One of the guidelines suggested a new designation for leaders of major data collection or software generation projects. Why do you think this is important?

Again, this was Jeff, but I agree. This is important because we need an incentive other than giving the generators exclusive rights to publications emanating from said data.

You also discussed the need for requiring statistical/computational co-authors for papers written by experimentalists with no statistical/computational co-authors and vice versa. What role do you see the referee serving? Why is this needed?

I think the same rule should apply to referees. Every paper based on the analysis of complex data needs to have a referee with statistical/computational expertise. I also think biomedical journals publishing data-driven research should start adding these experts to their editorial boards.

I should mention that NEJM actually has had such experts on their editorial board for a while now.

Are there certain guidelines would feel would be most critical to include?

To me the most important ones are:

1. The funding agencies and the community should reward data generators when their data is used by others. Perhaps more than for the papers they produce with these data.

2. Apply the same critical rigor on critiques of the original analysis as we apply to the original analysis. Bashing published results and talking about the "replication crisis" has become fashionable. Although in some cases it is very well merited (see Baggerly and Coombes work for example) in some circumstances critiques are made without much care mainly for the attention. If we are not careful about keeping a good balance, we may end up paralyzing scientific progress.

You mentioned that you think symbiotic data sharing would be the most effective approach. What are some ways in which scientists can work symbiotically?

I can describe my experience. I am trained as a statistician. I analyze data on a daily basis both as a collaborator and method developer. Experience has taught me that if one does not understand the scientific problem at hand, it is hard to make a meaningful contribution through data analysis or method development. Most successful applied statisticians will tell you the same thing.

Most difficult scientific challenges have nuances that only the subject matter expert can effectively describe. Failing

to understand these usually leads analysts to chase false leads, interpret results incorrectly or waste time solving a problem no one cares about. Successful collaboration usually involve a constant back and forth between the data analysts and the subject matter experts.

However, in many circumstances the data generator is not necessarily the only one that can provide such guidance. Some data analysts actually become subject matter experts themselves, others download data and seek out other collaborators that also understand the details of the scientific challenge and data generation process.

Related Posts

Not So Standard Deviations Episode 24 - 50 Minutes of Blathering 16 Oct 2016

Should I make a chatbot or a better FAQ? 14 Oct 2016

The Dangers of Weighting Up a Sample 12 Oct 2016