Peng House 4012

http://www.statsblogs.com/2014/04/01/this-is-how-an-important-scientific-debate-is-being-used-to-stop-epa-regulation/

This is how an important scientific debate is being used to stop EPA regulation

April 1, 2014
By Roger Peng

(This article was originally published at Simply Statistics, and syndicated at StatsBlogs.)

Environmental regulation in the United States has protected human health for over 40 years. Since the Clean Air Act was enacted in 1970, levels of outdoor air pollution have dropped dramatically, changing the landscape of once heavily-polluted cities like Los Angeles and Pittsburgh. A 2011 cost-benefit analysis conducted by the U.S. Environmental Protection Agency estimated that the 1990 amendments to the CAA prevented 160,000 deaths and 13 million lost work days in the year 2010 alone. They estimated that the monetary benefits of the CAA were 30 times greater than the costs of implementing the regulations.

The benefits of environmental regulations like the CAA significantly outweigh their costs. But there are still costs, and those costs must be borne by someone. The burden is usually put on the polluters, such as the automobile and power generation industries, which have long fought any notion of air pollution regulation as a threat to their existence.

> [COMMENT: Power companies are highly regulated and are generally given a specific "return on investment". It is natural to think that they pay for pollution cleanup. By law any costs they have are passed on to the public. Economists mostly agree that any business costs are passed on their customers.]

Initially, as air pollution and health studies were still emerging, opponents of regulation often challenged the science itself, claiming flaws in the methodology, the measurements, or the interpretation. But when study after study demonstrated a connection between outdoor air pollution and a variety of health problems, it became increasingly difficult for critics to mount a credible challenge. Lawsuits are another tactic used by industry, with one case brought by the American Trucking Association going all the way to the U.S. Supreme Court.

The latest attack comes from the House of Representatives in the form of the Secret Science Reform Act, or H.R. 4102.

> [COMMENT: the word "attack" seems a bit strong as you and coauthors called for environmental epidemiology data to be made publicly available.]
>
> R. D. Peng, F. Dominici, and S. L. Zeger, Commentary: Reproducible epidemiologic research, American Journal of Epidemiology 163 (2006), 783–789.

In summary, the *proposed bill requires that every scientific paper cited by the EPA to justify a new rule or regulation needs to be reproducible.* What exactly does this mean? To answer that question we need to take a brief diversion into some recent important developments in statistical science.

The idea behind reproducibility is simple. *All the data used in a scientific paper and all the computer code used to analyze that data should be made available to other researchers and the public.* It may be surprising that much of this data actually isn't already available. The primary reason most data isn't available is because, until recently, most people didn't ask scientists for their data.

> [COMMENT: I and a number of others have asked for air pollution data sets and have been uniformly refused access to data. In fact, you in a 2006 paper said data should be made available.]

> [COMMENT: there have been systematic requests for data sets. Even when the scientist signed a form saying they would provide the data sets used in a paper, 2/3s of the time data was not made available. Lack of access to data sets is wide-spread and a known problem.]

The data was often small and collected for a specific purpose so other scientists and the general public just weren't that interested. If a scientist were interested in checking the truth of a claim, she could simply repeat the experiment in her lab to see if the claim could be replicated.

> [COMMENT: Even experimental biology papers are proving surprisingly difficult to reproduce. See the Begley and Ellis Nature paper.]

The nature of science has changed quickly over the last three decades. There has been an explosion of data, fueled by the decreasing cost of data collection technologies and computing power. At the same time, increased access to sophisticated computing power has let scientists conduct more sophisticated analyses on their data. The massive growth in data and the increasing sophistication of the analyses has made communicating what was done in a scientific study more complicated.

> [COMMENT: In epidemiology, data is often put into SAS data sets and analysis done with SAS code. It should be easy to provide these data sets and analysis code. Yet data and code are most often not made available.]

The traditional medium of journal publications has proven to be inadequate for describing the important details of a data analysis. As a result, it has been said that scientific articles are merely the "advertising" for the research that was conducted.

> [COMMENT: Most journal now allow for Supplemental Material that is not limited in size. There are repositories, datadryad.org, where data can be deposited for a nominal fee. Journals are catching up.]

The real research is buried in the data and the computer code actually used to compute the results. Journals have traditionally not required that data or computer code be published along with papers. As a result, many important details may be lost and prevent key studies from being fully reproducible.

The explosion of data has also made completely replicating a large study by an independent scientist much more difficult and costly. A large study is expensive to conduct in the first place; there is usually little appetite or funding to repeat it. The result is that much of published scientific research cannot be reproduced by other scientists because the necessary data and analytic details are not available to others.

> [COMMENT: Again, technology, SAS data sets and SAS code, have been available since the mid 1970s. Currently it is relatively easy to give a scientists SAS code and SAS data sets.]

The scientific community is currently engaged in a debate over how to improve reproducibility across all of science. You might be tempted to ask, why not just share the data? Even if we could get everyone to agree with that in principle, it's not clear how to do it.

Imagine if everyone in the U.S. decided we were all going to share our movie collections, and suppose for the sake of this example that the movie industry did not object. How would it work? Numerous questions immediately arise. Where would all these movies be stored? How would they be transferred from one person to another? How would I know what movies everyone else had? If my movies are all on the old DVD format, do I need to convert them to some other format before I can share? My Internet connection is very slow, how can I download a 3 hour HD movie? My mother doesn't use computers much, but she has a great movie collection that I think others should have access to. What should she do? And who is going to pay for all of this? While each question may have a reasonable answer, it's not clear what is the optimal combination and how you might scale it to the entire country.

> [COMMENT: This come across as "Let's just agree that it is impossible." Before the IP owners got upset, music and movies were easily shared with servers. NCBI GEO is a repository for a massive number of biology experiments, many with massive amounts of data. The technology for sharing data appears to be readily available.]

Some of you may recall that the music industry had a brilliant sharing service that essentially allowed everyone to share their music collections. It was called Napster. Napster solved many of the problems raised above except for one -- they failed to survive. So even when a decent solution is found, there's no guarantee that it will always be there.

> [COMMENT: The music industry did not have "a brilliant sharing service" Napster was originally run by music pirates. They were put out of action by legal action by copyright owners. The technology worked. Services like iTunes, owned by industry work very well.]

> Wikipedia: "Napster is a name given to two music-focused online services. It was originally founded as a pioneering peer-to-peer file sharing Internet service that emphasized sharing audio files, typically music, encoded in MP3 format. The original company ran into legal difficulties over copyright infringement, ceased operations and was eventually acquired by Roxio. In its

second incarnation Napster became an online music store until it was acquired by Rhapsody from Best Buy[1] on December 1, 2011."

As outlandish as this example may seem, minor variations on these exact questions come up when we discuss how to share scientific data. The volume of data being produced today is enormous and making all of it available to everyone is not an easy task. That's not to say it is impossible. If smart people get together and work constructively, it is entirely possible that a reasonable approach could be found. But at this point, a credible long-term solution has yet to emerge.

[COMMENT: The enemy of the good is the perfect. NCBI GEO is working well. NSF datadryad.org is working well.]

This brings us back to the Secret Science Reform Act. The latest tactic by opponents of air quality regulation is to force the EPA to ensure that all of the studies that it cites to support new regulations are reproducible.

[COMMENT: There is pejorative language here, "tactic by opponents of air quality."]

A cursory reading of the bill gives the impression that the sponsors are genuinely concerned about making science more transparent to the public. But when one reads the language of the bill in the context of ongoing discussions about reproducibility, it becomes clear that the sponsors of the bill have no such goal in mind.

[COMMENT: Laws are laws and leave behind the intent of the writers of the law. Is the law sensible and good as WRITTEN? We are supposed to be a nation of laws, not subject to some person deciding in an ad hoc way, oh, you can have the data because you agree with me and you cannot because I think you have bad intentions.]

The purpose of H.R. 4102 is to prevent the Environmental Protection Agency from proposing new regulations.

[COMMENT: The EPA purports to be science-based. Making data sets available is part of science as you acknowledge. The man on the street would first think that if science data is not available then the EPA has something to hide.]

The EPA develops rules and regulations on the basis of scientific evidence. For example, the Clean Air Act requires EPA to periodically review the scientific literature for the latest evidence on the health effects of air pollution. The science the EPA considers needs to be published in peer-reviewed journals.

[COMMENT: The EPA makes much of "peer review". Peer review is no guarantee that a claim

made in a scientific paper will replicate. See Feinstein, Science 1988. See Prinz, Nature 2011. See also Begley and Ellis, Nature 2012. That peer reviewed science replicates so poorly is one of the reasons that reproducibility has become a critical issue.]

This makes the EPA a key consumer of scientific knowledge and it uses this knowledge to make informed decisions about protecting public health. What the EPA is not is a large funder of scientific studies. The entire budget for the Office of Research and Development at EPA is roughly $550 million (fiscal 2014), or less than 2 percent of the budget for the National Institutes of Health (about $30 billion for fiscal 2014). This means EPA has essentially no influence over the scientists behind many of the studies it cites because it funds very few of those studies. [COMMENT: $550M is not a small amount of money in the air pollution universe. A feed-back loop can develop. University researchers are smart and they know the policy goals of the EPA. Research, data and analysis, can be slanted. Large, complex data set are notoriously difficult to analyze. Any small bias can distort the resulting claims. If the result support EPA policy there is an increased likelyhood of more grants.] The best the EPA can do is politely ask scientists to make their data available. If a scientist refuses, there's not much the EPA can use as leverage.

[COMMENT: My understanding is that it is EPA policy to not require that data sets used in research they fund be given to the EPA. POLICY. Joe Cecil pointed out in a book chapter, The role of legal policies in data sharing, 1986, that an agency can shield itself from FOI by not taking position of the data. Someone wanting data cannot "reach through" the agency to the scientist that was paid by that agency to do the work. Is the EPA policy a reaction to the Shelby Amendment or Data Access Act, 1999.Can you imagine an industrial company funding research and as a matter of policy refusing a copy of the data sets they paid for? Can you imagine the FDA relying on "peer reviewed" studies to clear a new drug without access to the data?]

The latest controversy to come up involves the Harvard Six Cities study published in 1993. This landmark study found a large difference in mortality rates comparing cities with high and low air pollution, even after adjusting for smoking and other factors. The House committee has been trying to make the data for this study publicly available so that it can ensure that regulations are "backed by good science". However, the Committee has either forgotten or never knew that this particular study has been fully reproduced by independent investigators. In 2005, independent investigators found that they were "...able to reproduce virtually all of the original numerical results, including the 26 percent increase in all-cause mortality in the most polluted city (Stubenville, OH) as compared to the least polluted city (Portage, WI). The audit and validation of the Harvard Six Cities Study conducted by the reanalysis team generally confirmed the quality of the data and the numerical results reported by the original investigators."

[COMMENT: So the data is electronically available. The scientists are sure of the results. So why not make the data available?]

It would be hard to find an air pollution study that has been subject to more scrutiny than the Six Cities studies. Even if you believed the Six Cities study was totally wrong, its original findings have been replicated numerous times since its publication, with different investigators, in different populations,

using different analysis techniques, and in different countries.

> [COMMENT: The EPA contracted with NISS in the mid 1990s to look at the question of air pollution and deaths. NISS concluded that the claim was not supported. NISS never got another dime from the EPA.]

> P. Styer, N. McMillan, F. Gao, J. Davis, and J. Sacks, Effect of outdoor airborne particulate matter on daily death counts, Environ Health Perspect 103 (1995), 490–497.

If you're looking for an example where the science was either not reproducible or not replicable, sorry, but this is not your case study.

Ultimately, it is clear that the sponsors of this bill are cynically taking advantage of a genuine (but difficult) scientific debate over reproducibility to push a political agenda. Scientists are in agreement that reproducibility is important, but there is no consensus yet on how to make it happen for everyone. By forcing the EPA to ensure reproducibility of the science on which it bases regulation, lawmakers are asking the EPA to solve a problem that the entire scientific community has yet to figure out.

> [COMMENT: "cynically taking advantage" again is very pejorative language. Given your 2006 paper, are you now saying only the good guys should have access to data?]

> [COMMENT: Again, the enemy of the good is the perfect. For key studies, provide the SAS code and the SAS data sets. That would be a good start and indicate that the EPA actually wants transparency.]

The end result of passing a bill like H.R. 4102 is that the EPA will be forced to stop proposing any new regulation, handing a major victory to opponents of air quality standards and dealing a major blow to public health in the U.S.