Young Testimony

Testimony of Committee on Science, Space and Technology
Fostering Quality Science at EPA

I am Dr. S. Stanley Young.

I am the Assistant Director for Bioinformatics at the National Institute of Statistical Sciences, NISS. NISS is a not-for profit, non-governmental statistics organization. NISS' mission is to identify, catalyze and foster high-impact, cross- disciplinary research involving the statistical sciences. I am also the CEO of Omicsoft Corporation a company that designs software.

I graduated from North Carolina State University, BS, MES and a PhD in Statistics and Genetics.

I've worked in the pharmaceutical industry on all phases of pre-clinical research, first at Eli Lilly and then at GlaxoSmithKline. I've authored or co-authored over 60 papers and book chapters including six "best paper" awards. I co-authored a highly cited book, *Resampling-Based Multiple Testing,* which deals with false positives among other things. I have three issued patents. I conduct research in the area of data mining.

I am a Fellow of the American Statistical Association and the American Association for the Advancement of Science. I am an adjunct professor of statistics at North Carolina State University, the University of Waterloo and the University of British Columbia.

Today I am here to speak to making data sets used in papers supporting regulation by the EPA publicly available. It is just good science to have data used in papers public. A claim may be made. Is it plausible? If the data is not available, then the claim is effectively "trust me" science.

You might think, the claim is made in a peer reviewed journal, surely that makes it right. Peer review only says that the work meets the standards of the discipline and that on the face of it, the claims are plausible. Scientists doing peer review essentially never ask for the data set, they look for obvious things to correct and agree or not that the claims make some sense.

How often do claims prove false or dramatically less pronounced than in the original paper? Ioannidis, 2005, showed that for medical observational studies, claims fail about 80 percent of the time. I have kept informal count of claims coming from medical observational studies and then tested in randomized clinical trials. Over 90% of the claims have failed to replicate. Yes, 90% failure rate. I refer you to a recent paper covering these findings, Young and Karr (2011).

There are a number of technical and systems reasons for the high failure rate, which I will not deal with here. I will say that the work of congress and the work of regulatory agencies often depend on valid science. With the best of intentions, and incorrect scientific claims, you can make spectacularly bad decisions. To give a historical, medical example, two very large observational studies made the claim that Vitamin E will protect against heart attacks.

Several very large randomized clinical trials did not support those claims. Hundreds of millions of dollars were spent on the RCTs.

My goal here is to suggest several things that can be done to improve the situation. Any regulation that depends on epidemiology studies, e.g. formaldehyde, should make data public. The ACS CPS II database that is being relied upon for air pollution regulations should be public.

It makes sense to separately fund data generation and data analysis separately. One group collects and stages the data and posts it. Separate groups of scientists can be funded to analyze the data. Interested scientists can analyze the data. Scientists can become vested in the claims they derive from a data set. One group of scientists should not "own" a data set.

Making efficient the running of science is a good way forward. Science is much more efficient if scientists have access to the data used to make claims. One scientist can make a claim and another can say, let's examine the data and see if the claim is supported. Maybe there is a problem. For example, a Duke University study that lead to clinical trials was discovered to have data staging errors. Perhaps, the statistical analysis strategy is flawed. I examined a data set where a claim was made that eating breakfast cereal would make a boy baby more likely. Examination of the data showed the claim was the result of a flawed statistical analysis strategy. Evidence from medical observational studies indicates that claims most often fail to replicate. Environmental epidemiology studies are just as subject to error.

On publication of a paper, were research is funded by the EPA, the data should be made public. When the EPA proposes a regulation based on science, it should name the papers it is depending on and it should make data sets used in those papers publicly available. The agency should want to move forward based on good science. Congress should want the EPA regulations based on good science. The EPA would be more efficient if the entire scientific process is utilized. Congress would then depend not only on the EPA but the normal operating of science. Claims are more likely to be valid and the resulting policy sensible. Let normal science help in the vetting process. Make the data available.