Benefits and Importance of Data-Sharing

Tony Cox

tcoxdenver@aol.com

March 18, 2014
Current situation

- Epidemic of false-positive/irreproducible results in peer-reviewed literature

### Trials and Errors: Why Science Is Failing Us


Every year, nearly $30 billion is invested in biomedical research in the US, all of it aimed at bringing a cure to the ravages of disease. Photo: Mario Alpo

On November 20, 2009, researchers at Pfizer—the largest pharmaceutical company in the world—held a meeting with investors at the firm’s research center in Groton, Connecticut. Jeff Kansler, then CFO of Pfizer, began the presentation with an upbeat assessment of the company’s efforts to bring new drugs to market. He cited “exciting approaches” to the treatment of Alzheimer’s disease, fibromyalgia, and asthma. But that was just a warm-up. Kansler was more excited about a new drug called tootoot, which had recently entered Phase III clinical trials, the last step before filing for FDA approval. He confidently declared that tootoot was “one of the most important compounds of our generation.”

Kansler’s enthusiasm was understated. The potential market for the drug was enormous. Like Pfizer’s blockbuster medication, Lipitor—the most widely prescribed branded pharmaceutical in America—tootoot was designed to treat the cholesterol pathway. Although cholesterol is an essential component of cell membranes, high levels of the compound have been consistently associated with heart disease. The accumulation of the pale yellow substance in arterial walls leads to inflammation. Choices of white blood cells gather around these “plaques,” which leads to even more inflammation. The end result is a blood vessel clogged with atheroma.

Lipitor works by inhibiting an enzyme that plays a key role in the production of cholesterol in the liver. In particular, the drug lowers the level of high-density lipoprotein (HDL), or so-called “bad” cholesterol. In recent years, however, scientists have begun to focus on a separate part of the cholesterol pathway, the one that produces high-density lipoprotein. One function of HDL is to transport excess LDL back to the liver, where it is broken down. Therefore, HDL is a “junior of the family” cleaning up the gunky mess of the modern diet, which is why it is often referred to as “good cholesterol.”

And this return to the heartbreak. It was designed to block a protein that converts HDL cholesterol into its more sinister sibling. LDL. In theory, this would cure one cholesterol problem, creating a surplus of...
Why so many false positives?

1. Model selection biases
2. Multiple comparisons bias
3. Model form specification bias
4. Model over-fitting bias → narrow confidence intervals
5. Model uncertainty bias
6. Biases due to residual confounding
7. Biases from exposure estimation errors
8. Confirmation bias in interpreting results
9. Data selection biases
10. Study selection and publication biases

Opening data to scrutiny and new analysis can help the interested scientific community to detect and reduce these biases.
Current situation

• In a recent survey of three professional societies involved in risk assessment, “69 percent said it was ‘very important’ to have access to the underlying raw data for the most critical studies in order to do their own independent analysis of the results.”

• However, “only 36 percent said that having this access was often or always the case” (Butterworth, 2013).
Wider access to key data can help

Hypotheses:

1. Open access to key data (and models) $\rightarrow$ greater scrutiny $\rightarrow$ more careful research and interpretation of results $\rightarrow$ increased trustworthiness of published results

   - Restore lost credibility, trust and value for science and peer-review

2. Better (and more frequent) follow-up and extensions $\rightarrow$ increased value from research investments
Some possible pros and cons of data transparency for key studies

<table>
<thead>
<tr>
<th>Con</th>
<th>Pro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burden on researchers, chilling effect on original research in public interest</td>
<td>Good scientific method facilitates data sharing and results-checking, alternative analyses</td>
</tr>
<tr>
<td>Individual privacy and confidentiality could be compromised</td>
<td>Technical methods to address these are well developed</td>
</tr>
<tr>
<td>Unscrupulous or motivated investigators might use reanalyses and alternative analyses to delay needed actions and mislead the public and policy makers</td>
<td>J.S. Mill, <em>On Liberty</em>. Unattacked and undefended conclusions are often not worth having (not trustworthy)</td>
</tr>
</tbody>
</table>
Technical methods to protect privacy and confidentiality are well developed.
What can journals and others do?

• Supplemental online information to provide access to original data and analyses
• Open data access policies
• Improvements in public data archiving for scientific research (e.g., www.plosbiology.org/article/info%3Adoi%2F10.1371%2Fjournal.pbio.1001779).
• Example (3-14): “[A]ll PLOS journals will require that all manuscripts have an accompanying data availability statement for the data used in that piece of research” (http://blogs.plos.org/biologue/2014/02/03/opening-up-data-access-not-just-articles/).
• Sponsors: www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTDV030690.htm
Current situation: Untrustworthy results undermine science

Science advances on a foundation of trusted discoveries. Reproducing an experiment is one important approach that scientists use to gain confidence in their conclusions. Recently, the scientific community was shaken by reports that a troubling proportion of peer-reviewed preclinical studies are not reproducible. Because confidence in results is of paramount importance to the broad scientific community, we are announcing new initiatives to increase confidence in the studies published in Science. For preclinical studies (one of the targets of recent concern), we will be adopting recommendations of the U.S. National Institute of Neurological Disorders and Stroke (NINDS) for increasing transparency.* Authors will indicate whether there was a pre-experimental plan for data handling (such as how to deal with outliers), whether they conducted a sample size estimation to ensure a sufficient signal-to-noise ratio, whether samples were treated randomly, and whether the experimenter was blind to the conduct of the experiment. These criteria will be included in our author guidelines.
Key points

Making key data (and models) used to support decisions widely available for scrutiny can potentially create important benefits:

1. Improve original research quality and trustworthiness
   1. Reduce false-positive claims
   2. Invite alternative modeling methods and interpretations
   3. Replace faith (in original researchers) with reason based on data analysis

2. Improve follow-up research quality
   1. Enable others to test, improve, or extend causal models/assumptions
   2. Example: Air pollution health effects research

3. Making data available is not (or should not be) burdensome
4. It is technically feasible to protect privacy/confidentiality

Proposal: Sharing data upon request should be the rule, not the exception, for data used to support public policy decisions
Thanks
References

Appendix:
Supplementary Materials
Dockery (2002): "The results could not be more clear: Reducing particulate air pollution reduces the number of respiratory and cardiovascular-related deaths immediately."

A parable: Dublin intervention study

Wittmaack 2007, Pelucchi et al., Cox 2013: Air pollution reduction has no detectable effect on mortality rate trend

www.ncbi.nlm.nih.gov/pubmed/17365039
2012: Action taken

New Smoky Coal Ban Regulations will bring Cleaner Air, Fewer Deaths and can help efficiency 03/09/12

Phil Hogan T.D., Minister for the Environment, Community and Local Government, today (3 September 2012) announced that he has now given effect to new consolidating ‘smoky coal ban’ regulations.

Minister Phil Hogan T.D., said “Research has indicated that the smoky coal ban introduced in Dublin in 1990 resulted in up to 350 fewer deaths...per year. It has clearly been effective in reducing air pollution with proven benefits for human health and our environment and has led to improved quality of life in cities and towns where the ban applies. I believe that it’s now time to take steps to ensure that those proven benefits are preserved and safeguarded, and are extended more widely by updating the main provisions of the ban to reflect the more recent expansion of many of our urban areas and to ensure its continued effectiveness in mitigating harmful emissions caused by the burning of smoky coal.”


2013: Research re-done using control groups (Dockery)

NEW HEI RESEARCH REPORT
Did the Irish Coal Bans Improve Air Quality and Health?

...In contrast to the earlier study, there appeared to be no reductions in total mortality or in mortality from other causes, including cardiovascular disease, that could be attributed to any of the bans. That is, after correcting for background trends, similar reductions were seen in ban and non-ban areas.

The study by Dockery and colleagues shows that accounting for background trends in mortality can be crucial, since the earlier Dublin study appears likely to have overestimated the effects of the 1990 coal ban on mortality rates from diseases that were already declining for other reasons.

http://pubs.healtheffects.org/getfile.php?u=929
Lessons

• Confident expert judgments and peer-reviewed published interpretations about causality are *not* sound substitutes for more formal, objective causal analysis.

• Testing for potential causality using available data can be straightforward and highly informative:
  – Consider trends
  – Test for effects
  – Use control groups
Quantitative Evaluation of Multiplicity in Epidemiology and Public Health Research

Kenneth J. Ottenbacher

Epidemiologic and public health researchers frequently include several dependent variables, repeated assessments, or subgroup analyses in their investigations. These factors result in multiple tests of statistical significance and may produce type 1 experimental errors. This study examined the type 1 error rate in a sample of public health and epidemiologic research. A total of 173 articles chosen at random from 1996 issues of the American Journal of Public Health and the American Journal of Epidemiology were examined to determine the incidence of type 1 errors. Three different methods of computing type 1 error rates were used: experiment-wise error rate, error rate per experiment, and percent error rate. The results indicate a type 1 error rate substantially higher than the traditionally assumed level of 5% ($\alpha < 0.05$). No practical or statistically significant difference was found between type 1 error rates across the two journals. Methods to determine and correct type 1 errors should be reported in epidemiologic and public health research investigations that include multiple statistical tests. Am J Epidemiol 1998;147:615–19.

bias (epidemiology); probability; research design; significance tests

Levin noted recently, "Multiple comparisons are a very common feature—and, indeed, very often a necessity—in epidemiologic and public health research" (1, p. 628). He went on to discuss various procedures used to protect against type 1 errors, including the commonly used Bonferroni method and a procedure developed by Holm (2). Aickin and Gensler (3) argue that the Holm-adjusted $p$ value should be routinely used to reduce the type 1 error rate in studies involving multiple statistical tests.

Problems involving multiple statistical testing of hypotheses in health care and medical research arise for the following reasons: 1) the repeated analysis of accumulating data; 2) the use of multiple dependent measures; and 3) the analysis of data from subgroups (4). All three of these practices are common in public health and epidemiologic research. For example, Godfrey (5) demonstrated that researchers frequently present and analyze means from several groups within the same study. She found that the most common method of statistically comparing several means involved the use of multiple $t$ tests. Godfrey correctly argued that the use of univariate statistical procedures to analyze the results of studies containing multiple contrasts was inappropriate. Her analysis revealed that of 50 articles examined from the New England Journal of Medicine, a majority (54 percent) used improper univariate statistical procedures to analyze differences between subgroup means.