Data Torturing

Some years ago I was completing my undergraduate degree and was working in Dr. Allen's lab. I very much wanted to go on to graduate school and was looking to my project—a fairly sophisticated one I was doing with a post-doc—to get me admitted to the graduate program I had in mind. I was also counting on Dr. Allen, who was pretty well known in the field, to write me a nice letter of recommendation.

Unfortunately, my project yielded no significant data. I wasn't terribly bothered by this as I felt the results were still reportable—maybe even, given the nature of the research, "significant" in a nonstatistical way. Dr. Allen, however, could not have disagreed more.

He said I would have to write a report on this project and that he wouldn't have a report coming out of his lab with no statistically significant findings. He told me to go on a "fishing expedition" and perform a bunch of ad hoc analyses in the hopes of finding some correlations that were statistically significant. Because I was taking a statistics course at the same time, I knew what Dr. Allen was asking me to do was experimentally unsound and, therefore, unethical.

What a dilemma: I needed to turn in the report to graduate, and I needed Dr. Allen's letter of recommendation. If I questioned him on it, I was afraid he'd throw me out of his lab. I talked to my post-doc for advice, and she agreed that the fishing expedition, also termed "data torturing," was unsound. But she wouldn't go to Dr. Allen on my behalf.

So, I decided not to make any waves. I managed to find a few correlations that were statistically significant, wrote my report, and graduated with a letter of recommendation from Dr. Allen. I rationalized the whole thing by thinking, "Who was I to be questioning someone who had been doing research for so many more years than I?" Also, Dr. Allen never struck me as a devious person. I strongly suspect he didn't think there was anything wrong with the fishing expedition. Of course, the primary reason I went along was fear of retaliation and damage to my career prospects. I'd imagine there are many individuals who can relate stories similar to mine. What should institutions do to prevent this kind of thing from happening?

Expert Opinion

Why is data torturing ethically problematic? In a word, because neither the reported data nor the explanations or hypotheses the data torturer offers are all that trustworthy. As James Mills explained in a now classic 1993 article, either the data have been manipulated to fit a preferred or favored hypothesis or, as in the reported case above, the investigator "pores over the data until a 'significant association' is found between variables and then devises a biologically plausible hypothesis to fit the association.."[1, p. 1196] So, as in the above example, the investigator generates a post-hoc or a posteriori hypothesis to explain correlations that may or may not be generated by chance, even though they have been determined statistically significant at the P value of 0.05. (A P value of 0.05 means that there is a 5 percent chance that a reported difference occurring between two groups was actually due to chance (making it a false positive finding) or, alternatively, that there is a 95 percent chance that a reported difference between two groups is real and not due to chance.)

In the above case, both the data eventually selected for comparison as well as the hypothesis forwarded to explain their associations are more created than experimentally derived. Neither were experimentally planned or resulted from a primary hypothesis, which explains why this kind of data manipulation is sometimes referred to as a "fishing expedition"—one never knows what he or she will find.

One might object, though, that if the findings—no matter which—correlate at a P value of 0.05, then significance is significance and no harm is done. But such an objection obscures what usually occurs on the fishing expedition. Typically, what "opportunistic" (as Mills calls them) data torturers do is first generate dozens of categories or subgroups and then survey their associations or co-occurrences. But if one is analyzing dozens if not hundreds of such possible associations, certain ones might indeed demonstrate a P value less than 0.05 but only because no statistical adjustments were made for the multiple comparisons. In other words, the more one creates subgroups so that one can keep generating comparisons among them, the more one improves the chance that some of these associations will satisfy the P value of 0.05. But of those that do, one will not know which co-occur because of chance (i.e., due to all the comparisons) or which correlations are real, i.e., would be replicated by additional experiments. Mills's explanation of this deserves a lengthy quotation:

> [A]n honest exploratory study should indicate how many comparisons were made…most experts agree that large numbers of comparisons will produce apparently statistically significant findings that are actually due to chance. The data torturer will act as if every positive result confirmed a major hypothesis. The honest investigator will limit the study to focused questions, all of which make biologic sense. The cautious reader should look at the number of 'significant' results in the context of how many comparisons were made.(p.)

Nevertheless, Mills's observations were not universally accepted. Some months after his article appeared, Douglas Dix complained that some of the greatest discoveries of the modern era (e.g., Einstein's quantum theory, Mendelian genetics, the structure of the double helix) were generated in precisely the a posteriori fashion that Mills repudiates.[2] Indeed, one might say that the experiments to be conducted at the Large Hadron Collider near Geneva, Switzerland, which will examine collisions of sub atomic particles, is yet another instance of data in search of a theory.[3]

However, a glaring difference between the data torturing in the above example and the Mendelian/Quantum/Double Helix/Hadron Collider examples is that the former was done in apparent desperation, while the latter look to a long history of scientific inquiry, the plausibility of whose aims and rationales were under constant development, scrutiny and investigation. Indeed, in the above example, we have no antecedent notion of which data we're looking for. All we know is that we are looking for correlations that are statistically significant and around which we will try to fashion some kind of hypothesis that one hopes is scientifically plausible. None of this occurs against an historical background of a community of scientists doing hypothesis testing, data accumulation, analysis of data trending, public discussions at professional forums or in professional publications, and so forth. In the above example, Dr. Allen seems more interested in maintaining the reputation of his lab than in advancing the cause of generalizable knowledge. The data torturer is mostly trying to get away with his or her professional respectability intact.

Since Mills's article appeared, clinical trials are especially set up rigorously with specified questions and measurable end-points listed *before* the trial begins. Primary publications report mainly, if not exclusively, on this data. They may report some subgroup analyses, but clinical research has become very strict in clearly reporting any subgroup analyses as such, and not recommending treatment decisions based on subgroup data only. This is not to say that subgroup analyses are inherently bad. They often suggest new hypotheses and form the bases for a new clinical trial. In fact, they have opened up the field of "individualized" treatment where, for example, breast cancer patients are now treated based on specific features of the tumor and the patient. But this only became possible with carefully constructed *prospective* trials of the subgroups, which is the only way to ensure statistically significant results. Indeed, the FDA does not approve drugs based only on subset analyses, but requires prospective randomized trials.

Returning to the above dilemma, we believe it would have been acceptable if the report included the methods and results of this fishing expedition and suggested that perhaps a somewhat different hypothesis (or a different experimental design) might fit both the suggestive data resulting from testing the original hypothesis and also the statistically significant results teased out of the data by this fishing expedition and how they were found. The writer would be ethically bound to report why the results of the fishing expedition should be taken with caution and were not the result of an experiment carefully designed to ensure that they are trustworthy. But the dilemma contributor was directed by Dr. Allen to come up with a modified, after the fact hypothesis that would fit the statistically significant correlations and to misrepresent that this was the hypothesis that they were testing all along.

This is ethically troublesome, because by misrepresenting the process, the report contributes to a false understanding of the way 1) science proceeds, 2) the likelihood of success in connection with any one experiment, and 3) the challenges of formulating a hypothesis and designing an experiment that yields unambiguous results. This tainting of the "process data" ripples through the scientific community in a way that generates false expectations of students, research funders, scientists themselves, and the public. It contributes to a vicious cycle of motivation and temptation to do as was done here, which, in turn, encourages the reporting of results that are not properly qualified and hence untrustworthy and undermines the pursuit of scientific "truth."

In conclusion, here are some of Mills' recommendations for assessing allegedly statistically significant findings :

- Did the reported findings result from testing a primary hypothesis or an a posteriori hypothesis?
- Does the hypothesis have good supporting data from previous studies? Does it use theoretical insights and an examination of previously reported data?
- Have data been reported for all groups in the study or were certain study groups excluded from analysis and why?
- Is the reported finding consistent across a wide range of values, or does it only apply to a selected group and none of the rest?
- Are the cutoff points for laboratory studies reasonable and justifiable or are they selected because they allow the results to be significant?

- Was the effect of multiple comparisons discussed and statistically managed?
- How many significant results were reported relative to the number of comparisons made?
- Was the research outcome defined before collecting the data?

A final point: If there are many ways to understand a study's reported findings, there are probably many ways to tweak or "torture" the data.

References:

1. Mills J. Data torturing. *New England Journal of Medicine*, 1993;329:1196-1199.

2. Dix D. Correspondence. More on data torturing. *New England Journal of Medicine*, 1994;330:861-862.

3. Overbye D. Officials set timetable for getting particle collider back on track. *The New York Times*, Feb. 16, 2009, available at http://www.nytimes.com/2009/02/17/science/17cern.html?_r=1.

Recommended reading:

Huff D. *How to lie with statistics*. New York: W.W. Norton, 1954.