The NEW ENGLAND JOURNAL of MEDICINE

SOUNDING BOARD
◄ Previous Volume 329:1196-1199 October 14, 1993 Number 16 Next ►

Data Torturing

COMMENTARY
- Letters
- Letters

TOOLS & SERVICES
- Add to Personal Archive
- Add to Citation Manager
- Notify a Friend
- E-mail When Cited

MORE INFORMATION
- PubMed Citation

"If you torture your data long enough, they will tell you whatever you want to hear" has become a popular observation in our office. In plain English, this means that study data, if manipulated in enough different ways, can be made to prove whatever the investigator wants to prove. Unfortunately, this is generally true. Because every investigator wants to present results in the most exciting way, we all look for the most dramatic, positive findings in our data. When this process goes beyond reasonable interpretation of the facts, it becomes data torturing. The unfortunate result of torturing data is the dissemination of incorrect information to the research community and to patients.

It is impossible to tell how widespread data torturing is. Like other forms of torture, it leaves no incriminating marks when done skillfully., and like other forms of torture, it may be difficult to prove even when there is incriminating evidence. In this article I shall describe some of the telltale signs of data torturing in the hope of teaching readers to recognize its more obvious forms and of helping editors to eradicate it. My comments are based largely on actual cases; the details have been changed to protect the (possibly) innocent.

There are two major types of data torturing. In the first, which I term "opportunistic" data torturing, the perpetrator simply pores over the data until a "significant" association is found between variables and then devises a biologically plausible hypothesis to fit the association. The second, or "Procrustean," type of data torturing is performed by deciding on the hypothesis to be proved and making the data fit the hypothesis (Procrustes, a robber in Greek mythology, made all his victims fit the length of his bed by stretching or cutting off their legs).

Opportunistic Data Torturing

To understand how opportunistic data torturing works, it is necessary to understand the assumptions that underlie significance testing. In simple terms, significance tests are used to determine whether observed differences between groups, such as medically and surgically treated patients, are greater than one would expect to occur by chance. If survival rates in the two groups differed by 5 percent or 10 percent, for example, how would we know whether the difference was due to chance? For fairly arbitrary reasons, we usually say that a result is not due to chance if the P value is less than 0.05. A P value of 0.05 means that there is a 5 percent chance that we will conclude that the two groups differ when they actually do not (called a type I error). In other words, there is a 95 percent probability that we will correctly conclude that there is no difference when no difference is present. But when many independent tests are performed, that 95 percent probability of a correct conclusion drops drastically. For example, by simple probability calculations it can be shown that for two tests the probability that the "significant" differences found by the investigators will reflect true differences is 90 percent (0.95 x 0.95). For 20 tests, it is only 36 percent (0.95^{20}). Thus, the data torturer can find significant results when none exist simply by making multiple comparisons.

One slightly fictionalized example of opportunistic data torturing is a study of parents' occupational exposures as a risk factor for birth defects in their offspring. Seven major categories of occupational exposure were identified. When no significant relation between these categories and birth defects was found for either the mothers or the fathers (14 comparisons), the categories were split into 64 separate occupations for the mothers and 80 separate occupations for the fathers. Not surprisingly, the authors then found "significant" associations with birth defects. Although the authors mentioned that some positive results could have occurred by chance, the differences were treated as real. The probability that all their "significant" findings were real? Three in 10,000 (0.95^{158}).

It must be a great comfort to practitioners of this technique to know that 1 of every 20 independent comparisons they make will yield a "significant" result ($P<0.05$) if -- and this is critical -- they ignore the need to adjust for multiple comparisons. When this type of data torturing is done well, it may be impossible for readers to tell that the positive association did not spring from an a priori hypothesis.

Procrustean Data Torturing

Procrustean data torturing, or manipulating the data so that they prove the desired hypothesis, requires selective reporting. It can take several forms. First, exposure may be redefined in a way that strengthens the association. One study of adverse effects of oral contraceptives on the outcome of pregnancy defined exposure as presumed use within 600 days before a delivery or miscarriage; the choice of an inappropriately extended period to define exposure produced a positive result by including women not actually exposed during pregnancy. Second, study subjects whose experiences do not support the hypothesis may be dropped. For example, the report on a cancer-therapy trial might include outcomes only for subjects who survive more than three months, on the grounds that earlier deaths were inevitable and unrelated to the experimental therapy. In fact, these deaths could have resulted from toxic effects of the agent being tested. Third, disease outcomes may be lumped together, split, or dropped altogether to produce the desired results. In the cancer trial, for instance, the investigators' original intention might have been to look at differences in survival according to six-month intervals. But if no significant differences were found, the data could be reanalyzed according to longer or shorter time intervals until a significant difference was found. The authors could then report only the significant difference. Finally, normal ranges for laboratory results may be altered (although this must be done with care when common tests are reported). Of course, all these methods of selective reporting require the suppression of contradictory data.

Procrustean data torturing is more difficult to carry out than opportunistic data torturing, but its results are often more believable if one starts with a popular hypothesis. It is also more destructive, because it may produce results that are seen as definitive proof of the hypothesis, whereas opportunistic data torturing is often viewed as only hypothesis generation.

Clues to Data Torturing

Data torturing can rarely be proved. There are, however, clues that should arouse the reader's suspicion.

In the case of opportunistic data torturing (the search for chance associations), the reader must ask, Is this a chance finding with an a posteriori hypothesis concocted to give it credibility, or is this an honest hypothesis-generating study? Tukey[1] points out the need for exploratory studies using "theoretical insights and exploration of past data." Hypothesis-generating studies (sometimes referred to somewhat contemptuously as "fishing expeditions") should be identified as such. To warrant further exploration, findings from such studies should be biologically plausible. If the fishing expedition catches a boot, the fishermen should throw it back, not claim that they were fishing for boots. If a finding has good data from animal studies or related human studies to support it, it is unlikely to have resulted from opportunistic data torturing. If it has neither biologic plausibility nor supporting data, it should be viewed with a jaundiced eye.

Similarly, an honest exploratory study should indicate how many comparisons were made. Although there is disagreement about how (or even whether) to adjust for multiple comparisons,[1,2] 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. In the occupational-exposure study described earlier, nine "significant" findings were reported. Given that 158 comparisons were made, eight of those nine results could easily have occurred by chance.

Identifying Procrustean data torturing (in which the data are made to fit the hypothesis) also requires asking the right questions:

Why were study subjects dropped? One recent study of the health effects of exposure to heat dropped one of the four categories of exposure, changing a nonsignificant effect to a significant effect. One should suspect data torturing whenever subjects are dropped without a clear reason, or when a large proportion of subjects are excluded for any reason.

Does the classification of exposure and disease make sense? Statements such as "We studied those with at least five years of exposure to lead smelters, or those with blood lead levels of 50 µg per deciliter or higher" should raise questions. Why were the data on subjects with shorter exposure or less elevated lead levels not reported? Is it because they did not fit the hypothesis?

Are the cutoff points for laboratory studies reasonable and customary? Some of the bolder data torturers will argue that the clustering of subjects' test values at the upper range of normal is evidence of a pathologic state. Others will take advantage of the lack of a well-established cutoff point to select the point that makes their data produce the most significant results. A study of AIDS could use various CD4 cell counts as cutoff points, then report the one that shows the most impressive effect. The presence of a dose-response relation is evidence that the reported effect is genuine, not the result of arbitrary classification. If a diabetic woman's risk of miscarriage increases 5 percent for each 1 percent increase in her glycosylated hemoglobin level, the association is not likely to be due to data torturing. The key is that the effect is consistent across a wide range of values.

Is the rationale for the subgroup analyses convincing? If a drug works only in women over 60 years of age, the savvy reader should suspect a chance finding. Remember that two sexes, multiple age groups, and different clinical features such as stages of disease make it possible for the investigators to examine the data in many different ways.

Is there a clear biologic mechanism that could account for an effect in one subgroup but not in others, or were multiple comparisons made in order to produce positive results? "The study drug produced significantly increased survival at 18 months" may mean that there were no significant differences in survival at any of the other five periods examined.

In the same vein, it is important to ask whether the data have been censored. As I noted above, looking only at the group that survived at least three months after starting treatment may disguise the fact that the drug under study caused a substantial number of deaths in the first three months.

P Values and Confidence Intervals

Volumes have been written on the misapplication of statistical tests. I will say just a few words about the misuse of P values and confidence intervals. P values give the reader a sense of how likely an observation is to be due to chance, but they can be abused by investigators who make multiple comparisons without adjusting the standard for significance. Confidence intervals offer more information. Technically, a 95 percent confidence interval tells the reader that if the same study were done 100 times, with subjects from the same population pool, 95 of the 100 confidence intervals would contain the true relative risk (or whatever was being estimated in the study). Confidence intervals are thus valuable indicators of the precision of an estimate and the likely values of a measure (such as the relative risk of disease) within the population; a 95 percent confidence interval of 3.2 to 6.5 for a relative risk clearly defines an increased risk. A 95 percent confidence interval extending from 0.9 to 6.5 suggests a positive effect, but it is still within the realm of chance findings because the P value is greater than 0.05. Certainly, a confidence interval extending from 0.2 to 11.6 is merely an imprecise estimate. Yet such a confidence interval is sometimes used as evidence for high relative risk because the lower limit of an imprecise estimate can only approach zero, whereas the upper limit can increase without bounds.

Can Data Torturers Be Stopped?

Many, if not all, of these data-torturing techniques have been familiar to experts for years[3,4,5]. Some were described in the aptly titled book How to Lie with Statistics,[4] published nearly 40 years ago. Unfortunately, little has been done to alert the medical community to these abuses, or to eradicate them.

How can data torturing be prevented? It cannot. However, journals can demand information from authors that will discourage it:

Did the reported findings result from testing a primary hypothesis of the study? If not, was the secondary hypothesis generated before the data were analyzed?

What was the rationale for excluding various subjects from the analysis?

Were the following determined before looking at the data: definition of exposure, definition of an outcome, subgroups to be analyzed, and cutoff points for a positive result?

How many statistical tests were performed, and was the effect of multiple comparisons dealt with appropriately?

Are both P values and confidence intervals reported?

And have the data been reported for all subgroups and at all follow-up points?

Honest answers to these questions will make it much easier for editors, reviewers, and readers to separate honest data analysis from data torturing. Until such steps are taken, we shall remain at the mercy of those who are driven to produce positive findings by fair means or foul.

James L. Mills, M.D., M.S.
National Institute of Child Health and Human Development
Bethesda, MD 20892

I am indebted to Drs. Barry Graubard, Mark Klebanoff, and John Clemens for their valuable advice, and to Ms. Diane Wetherill for assistance in the preparation of the manuscript.

References

1. Tukey JW. We need both exploratory and confirmatory. Am Stat 1980;34:23-5.
2. Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology 1990;1:43-46. [Medline]
3. Wallis WA, Roberts HV. The nature of statistics. New York: Free Press, 1965.
4. Huff D. How to lie with statistics. New York: W.W. Norton, 1954.
5. Feinstein AR. Clinical biostatistics. St. Louis: C.V. Mosby, 1977.

Related Letters:

More on Torturing Data
Stryer D. B., Browner W., Newman T., Dix D., Schneiderman M. A., De Geeter F., Mills J. L.
Extract | Full Text
N Engl J Med 1994; 330:861-862, Mar 24, 1994. **Correspondence**

Guns and Homicide in the Home
Litaker D., Blackman P. H., *Saint Louis UniversitySt. Louis, MO 63108* , LeClaire J. E., *Saint Louis UniversitySt. Louis, MO 63108* , Gillette R. D., *Saint Louis UniversitySt. Louis, MO 63108* , Baranello P., *Saint Louis UniversitySt. Louis, MO 63108* , Beckmann C. R.B., *Saint Louis UniversitySt. Louis, MO 63108* , Pipas J. M., *Saint Louis UniversitySt. Louis, MO 63108* , Kellermann A. L., Somes G., Rivara F. P., *Saint Louis UniversitySt. Louis, MO 63108* , Kassirer J. P., *Saint Louis UniversitySt. Louis, MO 63108*
Extract | Full Text
N Engl J Med 1994; 330:365-368, Feb 3, 1994. **Correspondence**

This article has been cited by other articles:

- Lang, T. (2007). Documenting Research in Scientific Articles: Guidelines for Authors: 2. Reporting Hypothesis Tests. *Chest* 131: 317-319 [Full Text]
- Lang, T. (2006). Documenting research in scientific articles: guidelines for authors: reporting research designs and activities.. *Chest* 130: 1263-1268 [Full Text]
- Young, J., Graham, P., Blakely, T. (2006). Modeling the Relation between Socioeconomic Status and Mortality in a Mixture of Majority and Minority Ethnic Groups. *Am J Epidemiol* 164: 282-291 [Abstract] [Full Text]
- Vardy, J., Tannock, I. F. (2004). Quality of cancer care. *Ann Oncol* 15: 1001-1006 [Abstract] [Full Text]
- Chan, A.-W., Hrobjartsson, A., Haahr, M. T., Gotzsche, P. C., Altman, D. G. (2004). Empirical Evidence for Selective Reporting of Outcomes in Randomized Trials: Comparison of Protocols to Published Articles. *JAMA* 291: 2457-2465 [Abstract] [Full Text]
- CROW, T.J. (2003). Obstetric Complications and Schizophrenia. *Am. J. Psychiatry* 160: 1011-1012 [Full Text]
- Blum, R. S. (2002). Legal Considerations in Off-Label Medication Prescribing. *Arch Intern Med* 162: 1777-1779 [Full Text]
- Rethman, M., Hujoel, P. P., Drangsholt, M. T., Spiekerman, C. F., DeRouen, T. A. (2002). Pre-existing cardiovascular disease and periodontitis: a follow-up study.. *J. Dent. Res.* 81: 372-373 [Full Text]
- Wu, C. L., Fleisher, L. A. (2000). Outcomes Research in Regional Anesthesia and Analgesia. *Anesth. Analg.* 91: 1232-1242 [Full Text]
- Li, F. P. (1999). Cancer Control in Susceptible Groups: Opportunities and Challenges. *JCO* 17: 719-719 [Abstract] [Full Text]
- Stryer, D. B., Browner, W., Newman, T., Dix, D., Schneiderman, M. A., De Geeter, F., Mills, J. L. (1994). More on Torturing Data. *NEJM* 330: 861-862 [Full Text]
- Litaker, D., Blackman, P. H., Saint Louis UniversitySt. Louis, MO 63108, , LeClaire, J. E., Saint Louis UniversitySt. Louis, MO 63108, , Gillette, R. D., Saint Louis UniversitySt. Louis, MO 63108, , Baranello, P., Saint Louis UniversitySt. Louis, MO 63108, , Beckmann, C. R.B., Saint Louis UniversitySt. Louis, MO 63108, , Pipas, J. M., Saint Louis UniversitySt. Louis, MO 63108, , Kellermann, A. L., Somes, G., Rivara, F. P., Saint Louis UniversitySt. Louis, MO 63108, , Kassirer, J. P., Saint Louis UniversitySt. Louis, MO 63108, (1994). Guns and Homicide in the Home. *NEJM* 330: 365-368 [Full Text]
- Altman, D G (1994). The scandal of poor medical research. *BMJ* 308: 283-284 [Full Text]