

The Role of Data Access in Scientific Replication

John C. Bailar III

Presented at: Access to Research Data: Risks and Opportunities.
Committee on National Statistics, National Academy of Sciences,

October 17, 2003

I am grateful to Eleanor Singer, Chris Mackie, and the Committee on National Statistics for the opportunity to talk here about some matters that have interested me, and sometimes concerned me, for many years.

To this point, the emphasis here has been on microdata in the federal statistical system, with special attention to concerns about balancing the release of useful data against the concerns of persons identifiable from the data. I will make a sharp change in focus, to data generated by non-governmental sources, primarily in academia, and balancing the concerns of those who generate the data against the public interest in broader use. As noted yesterday, federal agencies typically have an interest in expanding access to the data they generate, while academics typically want to restrict access. The state of understanding about this aspect of academic research is nowhere near as advanced as thinking about federal microdata, so my comments will necessarily be much less focused.

My more specific topic is the role of data access in scientific replication. Eleanor talked yesterday about the risks and opportunities of increased access, and these are much in evidence in considerations about privately generated data. This broad topic immediately raises several equally broad issues. Most of these can be cast as a conflict between legitimate but competing values. One such conflict is that our society has a strong interest in protecting the privacy and confidentiality of persons who are the subjects of scientific study, but it also has a strong interest in assuring that scientific findings and interpretations are as close to correct as the state of the art allows. Another is that much research information is of both personal and proprietary value, but

it is often far from feasible for each potential user of some data to negotiate directly with either the person to whom the data apply or the holder of the data set about terms for its use. A third is that data are the stock in trade for most research scientists (with important exceptions such as mathematics and scientific theory), and scientific rewards are based almost exclusively on the generation and interpretation of data, but there is also a strong and legitimate interest in both understanding the social return from public investments and rewarding good science (whatever that may mean is a special context); this creates pressures for broad access to the data. Each of these conflicts reaches deep into questions about data access in scientific replication.

To cast these matters in terms of legitimate but competing values is to convert them from issues of science to issues of ethics and, inevitably, law. Yesterday I heard mostly about law, though the paper by Seltzer and Anderson was an exception. I have heard ethics defined as behavior when values are in conflict, and I will use that definition here. But to say that these matters of data access are issues of ethics and law is to say that there are no "right" answers. Subjective views about what both ethics and law should require are much in evidence here. Of course, nearly all scientists do try to be both ethical and lawful, but they are not often trained in these fields, just as ethicists and lawyers often do not have a deep understanding of the workings of science or the social context in which science is pursued. This lack of understanding may contribute much to the debate, sometimes acrimonious, about who has what rights to obtain and use scientific data. I do not myself have an expert's understanding of either ethics or law, but I will try to deal with the issues in a way that will further our dialog about data access. George Duncan, in his discussion, touched on the necessary trade-offs in the ethical balancing of competing values.

I will begin with several propositions that seem to me to be unobjectionable, though I am old enough to know that others may find good reason to object.

First, there are substantial costs and risks to science, and to knowledge generally, from curtailing access to data (where data should be broadly defined to include such things as computer programs, intermediate

computations, observations as recorded in the original lab notebooks, data points excluded as outliers or for other reasons, statistical analyses run but set aside in favor of other analyses, direct access to measuring apparatus for the purpose of checking calibrations, and on, and on). Few working research investigators would be happy to give away their data, and perhaps fewer would be happy to give away these intermediate products of their investigations, especially if the products of their work, other than published analyses, findings and conclusions, were to be examined by hostile interests bent on destroying the credibility of the findings. Nor is hostile scrutiny likely to advance the state of the science; necessary defenses against mischief may well retard progress by diverting time and other valuable resources from more productive work, and it may even discourage the best scientists from engaging in certain kinds of work that could lead to loss of exclusive access to data, or even to unproductive and unpleasant controversy. (I am well aware of the view of many lawyers that hostile examination is the best way to uncover the truth, but I have seen no evidence to support that view, and have in fact seen cross-examination used to destroy the credibility of difficult but perfectly correct science.)

Broad data access also raises questions about being "scooped" by competing investigators. While this is certainly a big concern to investigators, and often a barrier to sharing research data that others may want to use in ways the originator would not use them, it may have little impact in practice, for reasons I will come to later.

A second general point is that people or institutions good at generating data are not always the best at analyzing the data. As a member of two semi-permanent committees that sponsor and review research projects, I have found that the producers of sound, broad research data of general interest often do not know how to go past simple descriptive analyses, so that they and the public miss much of the value of their work.

Third, those who generate data are not always diligent about completing their own work and making the results public. Surely many of us have had experience with delays in publication by colleagues who are known to be

rather far along in some study, but have not properly completed it or put anything in the public domain. It is important here to understand that not every project worth doing is worth detailed reporting. Meta-analysts, in particular, are much concerned about "publication bias", which refers to the greater probability that interesting (generally "positive") findings will be published than sound but negative or unsurprising findings. This is not the venue for a full defense of publication bias, but a few words of support for it may be in order.

Some promising ideas are found to be dead ends; some are overtaken by other approaches to the same problem; some run into operational difficulties such that the data are seriously incomplete or untrustworthy. Nor does every report have to be a full-length original article, or even in the public, peer-reviewed literature. The key concept here is need to know. Who needs to know what about the work? Alternatives to Original Articles include a published abstract and presentations at a national meeting, reporting in brief form in the Introduction or Discussion section of an Original Article on a closely related topic, and even private circulation to a narrow group of persons if it includes all of those likely to have a use for the findings. I agree that research is never properly completed until there has been some report of the results in a form that is accessible to persons who may need to know about it, with enough detail for them to understand what was done, what came of it, and, sometimes, why the work was not completed and published in full. This simply does not always call for an Original Article in a peer-reviewed journal. These matters are of course much complicated by the fact that scientific rewards and recognition, at least in academia, depend almost entirely on publication.

Thus, I see a need for discussion about both ends of the publication spectrum - - research with limited, even private, distribution (perhaps because it does not meet the test for a broader need to know), and things that are dragged into hostile scrutiny where every detail is picked at with the intent is to demolish its credibility rather than to get to the truth. These matters need more discussion than they have received.

It may be useful first to separate replication and/or extension of the analyses from replication of data, which might generally be more satisfactory but may be prohibitively expensive, time-consuming, or difficult. In short, using the original data to replicate the analysis may be best regarded as a substitute for a broader package of new data as well as new analysis.

I will now tighten my focus on the role of data access in scientific replication. I have come to believe that there may often be good reasons for the support of research projects with a large component of data-generation to be sequential, divided between support for data generation and support for analysis.

Separation of data generation from analysis is not generally an issue with official statistics, because they are generally already separated. I will refer almost exclusively to academic and corporate efforts, which may be supported by federal dollars but are not generally under federal control. Data generation would include putting the data in proper condition for use by a range of other investigators, with any needed documentation. There would generally be only one primary data generation project, but several groups might be funded for the analysis.

Before I get to a few specifics, I want to put some limitations on it. First, is that it should be considered case-by-case, not that it be adopted universally. Second, it should be limited to research projects with a relatively large and difficult component of data generation, expensive in public dollars, and of particular public significance. Third, it should be phased in, with a fair amount of experimentation in its application before a more permanent set of policies is adopted.

A two-step approach to data generation and data analysis would still leave the data generators well ahead of others who might want to use the same data because the generators would be familiar with the data, have a running start on writing programs and otherwise preparing the analysis, would not need as much time and effort to write a new grant or contract as would a separate group, etc. I have mentioned such a separation to other scientists, and have often heard that no good scientist would spend his or her time producing data for others to "steal". (Emotions tend to run high when this kind of sharing

comes up.) However, I do not believe that would happen, for the simple reason that if nobody generates data, everybody will soon be out of business. There is also a set of questions about whether an independent investigator would know the things needed to do an analysis correctly, but that seems to me to be a red herring, unless disgruntled originators of data deliberately cause problems. The country is already full of people who make good use of micro data published by others -- for a host of examples look at the statistical programs of a host of US federal agencies, which collect data, do some critical analyses themselves, but tend to be pretty generous in sharing microdata; within the limits of needs to protect personal privacy and business secrets. I have had some personal experience with this two-step approach; many years ago, when I was Director of the Third National Cancer Survey (which evolved into the SEER Program of the National Cancer Institute), and adopted the policy of maximal access to data as a foundation of the program. I have seen no reason to regret that decision.

Of all the public myths about how science is done, one of the broadest and most persistent is that the scientific method rests on replication of critical observations. Straight replication is in fact uncommon, largely, I believe, because no scientist gets much professional credit for straightforward replication unless the findings are critical, there is suspicion of fraud, or there is some other unusual condition such that slavish replication of the methods reported might have some meaning not attached to the first round. Here I exclude replication by an independent investigator for the sole purpose of assuring himself or herself that the original results are correct and that the methods are working properly, as a preliminary to going further in some way.

There are other big and poorly understood barriers to replication of scientific data, including subtle differences in technique. The National Center for Health Statistics is just now trying to determine the number of children without health insurance. Five separate surveys have examined this matter in five different ways, with five different sets of definitions. It is no surprise that they have come up with five different answers. But which is "correct", in terms of what specific users of the estimates need to know?

Difficulties affect the physical sciences, too. The succession of incompatible measurements of the speed of light is well known. In chemistry, I recall hearing about a specific, real, case of a new chemical synthesis in which, spite the best efforts of the originator and the replicator, the latter could not get a specific synthesis to work -- until the two of them followed the protocol side by side and found that the critical difference was in whether a metal stirring rod touched the side of the beaker during mixing.

Other kinds of barriers include the prohibitive cost of some possible replications (time on astronomical instruments, long-term carcinogen bioassays, longitudinal studies of the population (which have still other barriers embedded in them)), the use of all the available material in the first round (including such "material" as the nation's supply of patients with some rare disease), changes in the context of a question with the passage of time (perhaps especially important in the social and medical sciences), changes in the questions themselves with the advance of knowledge in related topics, and a lot of other things.

Replication of data may also have some value in assuring scientific integrity. I was puzzled during the early phases of l'affaire Baltimore by what appeared to be the lack of interest in trying to replicate the studies that had been questioned. I raised this several times with persons on the side of the whistle-blowers, and was assured each time, rather curtly, that replication was not the issue, and would not settle the controversy or satisfy the complainants. This seems to leave the matter in an unsatisfactory state unless I have missed a follow-up.

Overall, replication of data seems to be one of those ideals that get a fair amount of discussion but have little influence on behavior. Perhaps what is most important is that the original investigators publish the background and methods with enough detail and precision for a knowledgeable reader to replicate the study if he had the resources and inclination to do so.

I turn now to replication of analyses, or in many cases, using data generated by others to undertake an analysis that has not been done even once.

Common examples arise in the use of broad, public governmental data such as the US census or the periodic national economic statistics. Our Federal Government publishes voluminous reports, but that is not enough for many important purposes. However, the government seems to have a divided mind about access to micro data. I have seen instances where a contract required that the investigator send all of the data and related materials back to the agency, and prohibited independent use of the data in any way, even with independent funding. In other instances, governmental funding agencies have been very active in promoting broad use of data, including microdata, that they have sponsored. An example is the SEER (Survival, Epidemiology, and End-Results) program of the National Cancer Institute, which contracts with universities, state health departments, and others to collect data for use by themselves, the NCI, and outsiders.

I will mention a couple of examples of disclosures that I have found particularly disturbing. One was the demand, some years ago, for disclosure to the tobacco industry of not just the micro-data but the names of persons surveyed in a tobacco study, so that the industry could harass the original respondents and try to change their answers for use in defense against a suit. I believe that one investigator did in fact release individual identification, but others objected and the objections were upheld. It was apparent, to me at least, that the sole purpose of the exercise would have been to discredit the original study, and probably to set a precedent for more of the same with other studies having outcomes unfavorable to the industry.

The other disturbing example is the sale of many decades of medical information, along with centuries of genealogies, by the government of Iceland to a private company for private profit. I would like to know more about why that government decided to make the sale, and why there was insufficient public outcry to block the sale, but in the absence of such knowledge I find this kind of invasion of privacy horrifying.

The government also has a divided mind depending on the direction of the flow of the data. In some contexts (e.g., the Patriot Act) it acts as though

everyone should rush in to provide everyone else, especially the Government itself, all of the data that might possibly be relevant to any perceived concern. In other cases (e.g., the protection of medical information, or business secrets), strict confidentiality is just dandy and should be locked into law forever (e.g., HIPPA). I have seriously overstated this, of course, to make the point that different users, with different uses in mind, will have different views and approaches regarding access to, and the further distribution of, various kinds of data.

Time does not permit discussion of several more matters of importance regarding access to data and scientific replication. These additional issues include the pros and cons of the Shelby amendment, which requires public disclosure of the original data for any analysis used in support of governmental policy-making or regulation; record linkage, which is critical to many of the concerns of this workshop; means to diminish disclosure, such as those used by the Census Bureau in its public use data sets; and the timing of data access, a large issue with some continuing governmental data series. Further, all of these issues interact in ways that I have not had time to discuss such as the potentially serious effects of record linkage on de-identification.

In the end, risk depends on both the likelihood and the consequences of some adverse outcome. Bill Lowrance, in a small book published some years ago by the NAS Press, wrestled with the concept of risk and finally came to the definition that a thing is safe if its risks are deemed acceptable. That requires knowing both the likelihood and the consequences of some problem, as well as knowing the benefits one hopes to generate. There has been little said here about the consequences of identification and other bad outcomes, though I think they must vary greatly from one area to another. They are surely important to assess.

Finally, I refer interested readers to the wonderful NAS book some years ago by (as I recall) Steve Fienberg, Margaret Martin, and Miron Straf, titled *Sharing Data*. It is full of careful thinking, and has not become in the least outdated by time.