PSYCHIATRIC SERVICES

psychiatryonline
Lifelong Learning in Psychiatry™

LETTERS | May 01, 2003

Large Data Sets Are Powerful

Steven P. Segal, Ph.D.

Psychiatric Services 2003; doi: 10.1176/appi.ps.54.5.745-a

To the Editor: To call a large data set "dangerous" is to reinforce a tendency to evaluate research on the basis of a chosen method rather than on how well the method enables the researcher to answer a question.

The problems with large data sets highlighted by Drake and McHugo—poor-quality data, statistical significance without meaningfulness, multiple tests, and post hoc interpretations—are neither new nor unique to large data sets. In addition, not all administrative data sets are necessarily large. It is difficult to find good data, administrative or otherwise, and to use it appropriately. Researchers must address validity problems or acknowledge limitations. Large administrative data sets can be put to good use, such as in addressing questions whose answers require observations over an extended period. Their value depends on what they are used for.

There might be more confidence in findings derived from large data sets if the science around statistical significance were not so focused on reification of .05.

Statistical significance is a probabilistic statement that helps us to discard good ideas that don't work or to know when to take a finding seriously. Some journals limit reports of significance to .001. Perhaps allowing values of .0001 or greater to be reported would help avoid overvaluation of small effects.

Sample size may be unrelated to frequency of significance testing. When more than one test per sample is used, p values no longer reflect the usual stated probability (<.05) unless procedures are used to adjust for the number of tests. Thus most values in behavioral science journals are approximations. Confidence in science is in replication.

Fitting the question to the data is a problem regardless of sample size. Should we discourage the use of large data sets to answer questions that can be addressed by them? Would the nation's health be better off if data from the long form of the U.S. Census or from the Physicians' Health Study were not available for wide-ranging investigation? Gathering good data is expensive, and even when a data set is mined a great deal, the confirmation of soundly conceived hypotheses provides support for future research.

Methods are tools with acknowledged limitations that are used to answer research questions. Labeling a method as dangerous without reference to how it is used to address a specific research question is as stigmatizing as calling mentally ill persons dangerous. At worst it may be self-serving, similar to the way in which some quantitative researchers devalue qualitative research. It also seems to support a trend of valuing statistical procedures over substantive content.

Recently a journal editor remarked that her publication looked askance at "salami research"—a term that describes use of the same data set to produce numerous research papers, each investigating a different question. This policy represents another form of evaluating research on the basis of the method used rather than on how well the method enables investigation. The same journal also took issue with use of data more than 15 years old, even though the data were used to investigate researchable questions.

These policies set a poor scientific standard. I would rather have a fine piece of salami even if it is old—many prefer their salami aged—than much of the bologna that is passed off as research articles in many journals. Within ethical boundaries, there is no substitute for research that validly addresses a question that needs resolution. Stigmatizing the use of particular research methods takes us in the wrong direction.

Dr. Segal is professor and director of the Mental Health and Social Welfare Research Group in the School of Social Welfare at the University of California, Berkeley.