Thoughts on the Future of Data Sharing

May 01, 2015

 A number of policy changes are occurring that could profoundly affect our science, perhaps in unanticipated ways. Unfortunately, many of these changes are being formulated without sufficient input from psychological scientists, even though components of the policies could threaten the protection of privacy, the data quality, the nature of our samples, and perhaps even the design of studies.

An example is reflected in policies regarding data sharing. In accordance with directives from the United States Congress and the White House, federal agencies such as the National Science Foundation and the National Institutes of Health (NIH) recently have been releasing updated expectations about data sharing. Let me emphasize that I am supportive of data sharing in the abstract and in a number of situations; indeed, data sharing often is very desirable for advancing innovations and scientific discoveries. It is important that people share data if a colleague wishes to validate or replicate published findings or analyze the data in other ways to answer scientific questions. On large projects involving many researchers and research participants as well as large amounts of public funds, it clearly is in the best interest of the science to capitalize on secondary analyses of the data by nonparticipating scientists. It also is quite reasonable, if study participants have signed a consent form that allows for general use of the data, for data in moderately sized studies to be made available for public use after a sufficient amount of time to allow for data preparation and initial key publications.

The latest large-scale policy is the NIH Genomic Data Sharing (GDS) Policy released August 27, 2014. Among the examples of data sets to which the policy applies, *but is not limited to*, are the following:

- Sequence data from more than one gene or region of comparable size in the genomes of more than 1,000 human research participants.
- Sequence data from more than 100 genes or region of comparable size in the genomes of more than 100 human research participants.
- Catalog of more than 100,000 single nucleotide polymorphisms (SNPs) from one or more model

organism species or strains.

The new policy applies to the phenotype and behavioral data that are collected in such projects; thus, some behavioral scientists are undoubtedly already involved in research covered by the policy. Moreover, the various institutes at NIH have the authority to interpret the policy as including data sets with smaller sample sizes and different amounts and types of genomic data if the data are of interest to the institute or center. Of perhaps most importance, this policy is a model for expanding data-sharing efforts dealing with sharing nongenetic/genomic data like what NIH and the Department of Defense's FITBIR (Federal Interagency Traumatic Brain Injury Research) database for traumatic brain injury research currently requires (available at fitbir.nih.gov).

In the Supplemental Information to the NIH Genomic Data Sharing Policy, specific expectations are outlined in regard to data producers' submission of data and the release of these data for use by other scientists. For human genomic data, the expectation for submission of the data is "Project specific; after cleaning and quality control, which is generally within 3 months after data have been generated"; the expectation for release is "Up to 6 months after data submission is initiated or at the time of acceptance of initial publication, whichever occurs first." In addition, and particularly relevant to psychological scientists, the phenotypic data in these studies, which include the kinds of variables that most of us collect, are to be submitted "as analyses are completed" for human data (p. 3).

However, the document further states that phenotype or clinical data are expected to be "submitted to the NIH-designated data repository at the earliest opportunity, but no later than the date of level 2 genomic data submissions (or levels 2 and 3 for GWAS datasets), especially for studies in which all phenotype data have already been gathered." (Levels 2 and 3 data refer to various forms of genomic data to be submitted within 3 months after data have been generated.)

For studies in which phenotype data collections are ongoing and/or may be regularly updated, data files should be submitted to NIH-designated data repositories as early as possible considering the practical needs for ensuring data accuracy; generally speaking, this time should not exceed 3 months after data cleaning begins (p. 4).

The expectation that phenotypic data, even in ongoing longitudinal research, should be submitted within 3 months after data cleaning begins is totally unrealistic in many cases and reflects a lack of understanding regarding the complexity of behavioral/psychological data. Note that once the data are released, even if only months after collection and cleaning, they often can be used by other investigators for publications without any participation by the researchers who designed and conducted the original research.

Such mandates could create problems. First, for many types of behavioral data, quality data generation may be impossible under the expectation of data deposition in a matter of months because the reality of working with such data can mean a year or more is needed to train coders and to actually code and clean the data for analyses. In longitudinal studies with large quantities of behavioral data, this process is ongoing. Moreover, for those researchers conducting longitudinal studies (or a series of studies), data analyses and publications often cannot proceed at a rapid pace (or even commence) until the researchers are finished with data collection, coding, and data cleaning. Thus, without realistic embargo periods that would protect the data-producing scientists by prohibiting other scientists from publishing their data,

other people could easily publish most of data-producing scientists' findings before they themselves have an opportunity to do so. In brief, people who have spent years conceptualizing the research; writing grants to fund the work; conducting the research; and coding, cleaning, and analyzing the data may be systematically disadvantaged in regard to being able to publish the results of their own research.

There are many other pragmatic concerns. For many behavioral measures (e.g., codes of parent–child interactions), the coding systems are complex and involve extensive training. Because of the difficulties with documenting complex measures, users of archived data often may misinterpret or lack sufficient understanding of the data to use them correctly. Furthermore, in the process of analyzing data for publication, coding and data entry errors are often discovered; thus, people who use the data prior to the original investigators may base their analyses on incorrect data. Users of the data also may need the assistance of the original investigators, perhaps over an extended period of time; the cost in time to the original researchers could be great and could continue for years after funding. These obstacles are not insurmountable, but they require attention and targeted funding.

There are additional financial costs to consider. The resources needed for data deposition must be included in the same budget used to conduct the research, but the budget often is insufficient for collecting, cleaning, and analyzing the proposed data after the typical NIH cuts of 10% to 24% across the board when funded. And to receive funding, investigators often are expected to submit a data-sharing plan that is endorsed by their university; if they do not submit the data on time, the university and researcher may suffer the consequence of not having their grant funding released.

Finally, inherent in the recent policies is an expectation that research participants will agree to participate in broad general research that goes beyond the immediate study. Thus, ethical concerns arise related to how to obtain informed consent and the identifiability of participants. In the age of genomics, neuroimaging, the Apple Watch, Fitbit, Facebook, and geospatial and other demographic data, researchers will have difficulties ensuring research participants' privacy and anonymity by merely removing HIPAA identifiers. Moreover, it is likely that many potential study participants, especially those in vulnerable groups (e.g., various racial-ethnic groups; people who use illegal substances or have stigmatizing traits, stigmatizing behaviors, or rare health issues), will be reluctant to participate knowing that their data will be shared with the government and the public (even if the data are "de-identified"); such refusals can introduce systematic biases into our samples. I believe that many parents in my own research would not participate if we could not guarantee that the data will not be released to others outside our research group.

As previously noted, it is likely that the aforementioned 2014 policy will serve as a model for future data sharing of nongenomic data. Thus, before this happens, it is important for our scientific community to educate the policymakers who desire a rapid timetable for data sharing about the nature of our research and to work proactively with the government agencies drafting the details of policies relevant to psychological science. œ