Subject: **Response to REQUEST FOR INFORMATION: PUBLIC ACCESS TO PEER-REVIEWED SCHOLARLY PUBLICATIONS RESULTING FROM FEDERALLY FUNDED RESEARCH**

Date: December 22, 2011 6:38:32 PM EST

This is in response to the RFI: Specifically it is in response to: "Please identify any other items the Task Force might consider for Federal policies related to
public access to peer-reviewed scholarly publications resulting from federally supported research."

As introduction, I am a Professor of Radiology at UCSF, I've been doing research for more than 40 years, and I am currently Prinicple Investigator of the Alzheimer's Disease Neuroimaging Initiative (ADNI) which is the largest grant funded in the world concerning Alzheimer's disease. (AD) ADNI is a multisite longitudinal clinical observational study aimed at validating imaging and biomarkers for diagnosis, early detection, and as inclusion and outcome measures in AD clinical trials. ADNI shares ALL raw, processed, and analyzed data with all qualified scientists in the world through its website UCLA/LONI/ADNI, which (to our knowledge) is unprecedented!. This data sharing has led to almost 300 peer reviewed publications. This experience has led me to be a passionate advocate of widespread sharing of all raw scientific data, after publication.

Although currently, most scientists do not share raw data, one big problem is that for those scientists who wish to share their data, there are no easy ways to do this. There is no federally funded mechanism for sharing raw data. Universities do not provide mechanisms for this. In fact there is not a lot of available software for data sharing, although DATAVERSE from Harvard (Gary King is PI) is an excellent open source tool for data sharing.

I believe that widespread sharing of raw scientific data
which is described more extensively in the attached document, will have numerous benefits, and thus should be a national repository. Specifically there should be a National Raw Scientific Data Repository, where any scientist can deposit their data. The data could easily be linked to publication in PubMed Centra. The cost of such a repository would not be huge. I believe that the benefits could be quite substantial leading to new discoveries, less scientific fraud, and a shift of the scientific culture leading to more cooperation and interaction. All of the benefits and issues are described in more detail in the attached document.

Sincerely

Michael W. Weiner, M.D.

Director, Center for Imaging of Neurodegenerative Diseases (CIND)
http://www.cind.research.va.gov/
San Francisco VA Medical Center
4150 Clement Street (114M), San Francisco CA, 94121 USA

Professor of Medicine, Radiology, Psychiatry, and Neurology
University of California, San Francisco, USA

Principal Investigator: Alzheimer's Disease Neuroimaging Initiative (ADNI)
http://www.adni-info.org/

Principal Investigator: Resource for MRI of Neurodegenerative Diseases
http://www.rrmind.research.va.gov/

Scientific Data Sharing Project
http://scientificdatasharing.com/

Plan to Achieve Widespread Sharing of Scientific Data

Michael W. Weiner M.D., Director, Center for Imaging of Neurodegenerative Diseases
Professor of Medicine, Radiology, Psychiatry, and Neurology, UCSF

Overall Goal: The overall goal of this project is to achieve widespread voluntary sharing of scientific data at the time of publication. The impact of widespread sharing of scientific data will be: to greatly increase the amount of information and knowledge which will be available to all scientists, to accelerate development and facilitate new discoveries of improved diagnostic and therapeutic methods leading to improved health and quality of life for society, and to stimulate the economic sectors which will use this information including the pharmaceutical, biotechnology, chemical, engineering, and computer/internet industries. In addition to the scientific and economic gains, substantial benefits from data sharing are also expected for education, scientific culture and communication. This goal will be achieved by a parallel approach of individual and institutional initiatives as discussed below.

Background: My experience as the Principle Investigator of the Alzheimer's Disease Neuroimaging Initiative (ADNI) has led me to the conclusion that widespread sharing of scientific data can be achieved now and that great scientific and economic benefits will ensue. ADNI is the largest NIH grant funded for Alzheimer's research ($140 million total funding thus far) and all our raw data is immediately shared with all scientists in the world without embargo. The success of this project (more than 160 publications and 80 more submitted) demonstrates the feasibility of this approach and reinforces the success of other projects which share data, e.g. the Human Genome project. Currently data sharing is mostly done by large, well funded multi-investigator projects. There would be great benefit if much more raw data were widely shared, especially data from individual investigators in all fields of biological/medical science (and other areas of science as well). This would be best done at the time of publication, with the raw data being linked to papers in PubMed Central. It should be mentioned that in many fields within the social sciences (e.g. economics and political science), sharing of raw data at time of publication is already widely done, and is de rigueur. Widespread voluntary sharing of raw data will be achieved using two interlinked approaches:

Individual initiatives: Our laboratory at the Center for Imaging of Neurodegenerative Diseases will begin to share data at the time of publication during 2011. The raw data, e.g. individual subject data including numerical maps and images, will be available on a website, and access to the data will be achived using available software (Dataverse) which allows the investigator control over data release. In addition to the raw data, a description of how the raw data was processed and analyzed, leading to the findings in the publication, will be provided. All data sharing will be performed with permission of the Institutional Review Boards and other university and governmental authorities concerned with human subjects and privacy protection. As this is being achieved we will identify other scientific groups who are sharing data and post them on our website scientificdatasharing.com. We will relate our experiences to the following: 1) Other collaborators in ADNI. ADNI scientists will be encouraged to share the raw data of their ADNI papers, and other papers from their laboratories. This should impact the field of Alzheimer's

leading to a greater acceptance of data sharing in the medical imaging and neurology fields. 2) Other faculty in the Department of Radiology at UCSF and our collaborators in Neurology and Psychiatry at UCSF. 3) If there is sufficient interest at UCSF, the Chancellor, Deans, and Department Chairs will be urged to make more widespread voluntary sharing of scientific data a UCSF priority/policy. Such actions would include providing storage space for shared data, and development of policies which would reward data sharing in the hiring and promotion process. The example of UCSF should urge the entire University of California system to encourage data sharing. 4) Other collaborators and colleagues in other universities around the world will learn about the work done at UCSF and in ADNI and will adopt similar policies (evidence of this happening is already available). 5) We will develop and test a "data sharing impact factor" which would allow scientists to cite the utilization by others, of data they collected.

Institutional mechanisms: Efforts are already underway to encourage increasing involvement by the NIH, NSF, and the National Library of Medicine (NLM), to promote and facilitate sharing of scientific data. These efforts will be strengthened as we gather increasing evidence of data sharing by our laborator and others at UCSF and elsewhere. First, the NIH and NSF will be encouraged to emphasize and expand their existing policies concerning data sharing and notify the scientific community of this greater emphasis. Second, the NIH should establish a small group of committed individuals who can help formulate policy in this area and suggest specific steps including generation of budgets to achieve specific goals. One approach to this would be for a few Institutes (such as the NIA, NIBIB, NCRR, NIMH, NINDS, NIAAA/NIDA, Neuroscience Blueprint) to take a leadership role in creating a policy framework that favors open availability of scientific data. Third, and hugely important, would be to establish technical mechanisms for data sharing, such as a national system for storage of all raw scientific data, such as a national data repository or data bank. This can be achieved by the National Library of Medicine, with links on PubMed Central publications to the raw data. Another approach would be repositories supporting universities, foundations or private companies, using systems like Dataverse. The advantage of an NLM national repository is that its long term exisitance would not be in doubt. Fourth, means should be developed at NIH and NSF to incentivize scientists and institutions to share their raw data. This could be done: 1) by requesting reports in non competitive reviews, competitive reviews and/or new applications; 2) through the grant review process by instructing the reviewers to consider data sharing in assessing priority scores; 3) through special acknowledgements in publications; 4) by providing affordable access to infrastructure, i.e. software and media, which facilitates data sharing. Fifth, the NIH should provide funding for small grants aimed to promote and take advantage of shared data. New methods are already under development for putting pieces of data from different sources together and making a new whole collection of data which is greater than the sum of its parts. Having all the raw data on the internet will enable: 1) data mining: computer methods which troll the internet searching for particular types of information of interest; 2) cloud computing: computer methods which assemble different "clouds" of data and derive new knowledge using the combined information. Knowledge based industries are likely to benefit because of the increased accessibility to large a amounts of raw data - especially the pharmaceutical and health care industry, chemistry, technology, engineering, etc. But ultimately the entire economy would benefit from improved knowledge. New technologies and new companies are expected to be developed to take advantage of the new information being made widely available. It is hard to predict the exact benefits of this type of activity. On the other hand, the costs would probably be relatively modest.

Its important to re-emphasize that the gains to be achieved by promoting widespread sharing of raw scientific data promise to greatly outweigh the relatively small costs involved in developing the necessary infrastructure. There is reason to believe that there will be substantial economic and public

benefits gained by widespread sharing of scientific data, because of the ability to link data sets, and make discoveries not related to the original goals of the data collectors.

In conclusion, we propose a two-fold plan (individual initiatives and institutional mechanisms) to achieve widespread sharing of scientific data which will speed the development of improved diagnostic techniques and treatments for a wide range of disorders, and to facilitate the growth of the knowledge based economy. Of course we would be happy to discuss this in more detail, and present a (modest) budget to implement this plan.