(Saturday) 4 December 2010

Memorandum

To: Curt Meinert

Fr: CLM

Re: Individual patient data sharing

Having written recently on meta-analyses and systematic reviews, I turn now to a related topic – data sharing. I know you have thought a good deal about this, but I also know that you have not committed your thoughts to writing so I am doing it for you. I will be interested to learn if I have captured your thoughts on the subject.

The Final NIH Statement of Policy on Data Sharing, issued 26 February 2003, is reproduced below.

NIH reaffirms its support for the concept of data sharing. We believe that data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health. The NIH endorses the sharing of final research data to serve these and other important scientific goals. The NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers. Starting with the October 1, 2003 receipt date, investigators submitting an NIH application seeking $500,000 or more in direct costs in any single year are expected to include a plan for data sharing or state why data sharing is not possible.

As indicated above, all investigator-initiated applications with direct costs greater than $500,000 in any single year will be expected to address data sharing in their application. Applicants are encouraged to discuss their data sharing plan with their program contact at the time they negotiate an agreement with the Institute/Center (IC) staff to accept assignment of their application as described at http://grants.nih.gov/grants/guide/notice-files/NOT-OD-02-004.html. Applicants are reminded that agreement to accept assignment of applications over $500,000 must be obtained at least six weeks in advance of the anticipated submission date. Instructions related to the data sharing policy as it is applied to applications and proposals responding to a specific Request for Application (RFA) or Request for Proposals (RFP) will be described in the specific solicitation. In some cases, Program Announcements (PA) may request data sharing plans for applications that are less than $500,000 direct costs in any single year. Reviewers will not factor the proposed data-sharing plan into the determination of scientific merit or priority score. Program staff will be responsible for overseeing the data sharing policy and for assessing the appropriateness and adequacy of the proposed data-sharing plan.

NIH recognizes that data sharing may be complicated or limited, in some cases, by institutional policies, local IRB rules, as well as local, state and Federal laws and regulations, including the Privacy Rule. As NIH stated in the March 1, 2002 draft data

sharing statement (http://grants.nih.gov/grants/guide/notice-files/NOT-OD-02-035.html), the rights and privacy of people who participate in NIH-sponsored research must be protected at all times. Thus, data intended for broader use should be free of identifiers that would permit linkages to individual research participants and variables that could lead to deductive disclosure of the identity of individual subjects. When data sharing is limited, applicants should explain such limitations in their data sharing plans.

The final NIH statement on data sharing is largely the same as stated in the March 1, 2002 draft with the following exceptions:

- *The effective start date has been changed from January 1, 2003 to October 1, 2003 receipt date.*
- *This policy applies to applicants seeking $500,000 or more in direct costs in any year of the project period. Such applicants are expected to contact IC program staff prior to submission and are also expected to include a data-sharing plan in their application stating how they will share the data or, if they cannot share the data, why not. Applicants responding to an RFA or RFP will find instructions related to data sharing in the specific announcement.*
- *Several groups and individuals objected to sharing of research data prior to publication. As noted earlier, NIH recognizes that the investigators who collect the data have a legitimate interest in benefiting from their investment of time and effort. We have therefore revised our definition of "the timely release and sharing" to be no later than the acceptance for publication of the main findings from the final data set. NIH continues to expect that the initial investigators may benefit from first and continuing use but not from prolonged exclusive use.*

(http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html)

The policy relates to any NIH application of $500,000 or more in direct costs in any year of requested support. The policy is not specific to trials though, by virtue of the dollar amount specified, the effect is to apply to most NIH-supported trials to the extent that they are likely to exceed the dollar limit. The exceptions are small-scale, single center trials, but even those will be fewer in time as the dollar erodes, sort of like the creep of the alternative minimum tax (AMT) to ever lower income earners.

The policy is appropriately devoid of details on what data are to be shared and on how data are to be shared, but the policy does indicate that the *NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers.* Although the policy does not use the term "public", the implication is that data will be available to the public.

The policy recognizes the right of investigators to be the first to present and publish, but only within limits. "Timely release and sharing" is defined to be *no later than the acceptance for publication of the main findings from the final data set*.

The basic problem with the policy is that it removes the option of choice. "Sharing" implies the existence of a partnership between the giver and the receiver where both stand to gain from the sharing. This form of sharing has gone on for centuries. The NIH policy, however, sucks the life out of "share" and might as well have been titled *NIH Policy on Giving Data Away*.

Sure there may be societal gain in giving data away, but try selling your societal contribution to a promotions committee as a reason for being light on publications.

The policy states that sharing is to be "*no later than the acceptance for publication of the main findings from the final data set*". This policy assumes that publication of the "main findings" is based on the "final data set" and that there is a single "main findings" publication. Consider ADAPT. The first publication came in November 2006 describing results leading to the decision to stop the trial. Since then there have been six major publications and several more in the works. Data collection ended 31 May 2007 and data entry and editing ended 31 October 2008. The final dataset was distributed to study investigators July 2009.

There are risks with IPD data sharing, even with de-identification. Any breach of confidentiality will result in censure by the sharer's IRB. Another risk is to the creditability of the trial if the dataset supplied is error-laden or if the mapping done to create the dataset is errant.

Leastwise you have doubts as to how this goes, revisit the VIGOR trial and the "expression of concern" by NEJM editors three years after results were published because of count discrepancies (NEJM 2005; 353; 2,813-4 and NEJM 2006; 344; 1.193).

So how about some rules to guide IPD sharing?

Rule 1: In NIH applications above the dollar limit, outline plans for data sharing
Comment: The plan should indicate if sharing will be limited to analysis datasets not requiring de-identification. If the plan includes preparation of an IPD de-identified dataset, the plan should indicate when that is likely to occur.

Rule 2: In conduct of the trial, insist on consent forms which inform study subjects that data will reside in a data center and that they will be provided to the collective investigatorship, including investigators located at other participating centers
Comment: In multicenter trials make certain that the consent forms used at the various clinics include that notice.

Rule 3: Write rules as to what constitutes the "final dataset" and outline the conditions that must be satisfied before de-identified IPD are released outside the study
Comment: Rules as to availability of IPD de-identified data should be written early in the course of the trial. Considerations in writing the rules are whether there will be a single "main paper" or multiple "main papers" and the definition of "final dataset".

Rule 4: Have the plan reviewed and approved by study leaders
Comment: People in the coordinating center will need that buy-in when pressed by NIH sponsors for IPD de-identified data.

Rule 5: Ensure that study investigators have access to identified IPD datasets and that they have sufficient time to use them prior to release of de-identified IPD datasets outside the study
Comment: Typically, investigators in clinics are shielded from treatment results. They do not have access to study data until the trial is stopped or finished. Investigator rights of primacy require that they be supplied with study data for their own use before supplying data to others.

Rule 6: Do not provide data for use outside the study without approval of study leaders or without IRB approval

Rule 7: Datasets supplied to investigators in the study, should not contain personal identifiers (such as study subject name, address, social security number, and other numbers directly linked to persons studied)

Rule 8: De-identify data as specified by HIPAA for export to others outside the study
Comment: De-identification of persons studied, their relatives, household members, and employers under HIPAA regulations, involves stripping:

- Names
- Any geocodes that identify an individual household such as street address or Post Office Box Number
- Telephone numbers
- Fax numbers
- Electronic mail addresses
- Social Security Numbers
- Medical record numbers
- Health plan beneficiary identifiers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers and serial numbers, including license plate numbers
- Medical device identifiers and serial numbers
- Web universal resource locators (URL)
- Internet Protocol (IP) address numbers
- Biometric identifiers, including finger and voice prints
- Full face photographic images

Datasets must also be devoid of:

- Geographic subdivisions designations smaller than a state (i.e., county, city, town, precinct)
- 5 or 9 digit ZIP codes (1st three digits allowable in most cases)
- All elements of dates (except year) directly related to an individual, including dates of birth or death, dates of health care services or health care claims (de-identified datasets cannot contain birth dates; file may contain the individual's age expressed in years, months, days, or hours, as appropriate, except for individuals aged 90 or above; such persons to be identified simply as being 90 or above)
- Any other unique identifying number, characteristic, or code that could be used to identify the individual (supplier of data may affix codes to allow user to associate data with persons, provided codes cannot be used to re-identify persons)

AND

Certification by qualified experts that the likelihood of probabilistic identification is nil. The certification requirement has the effect of rending de-identified data useless. Certification of non-identifiability requires "combining and collapsing" (eg, as by combining blocks in a census survey). "Combining and collapsing" render data useless for IPD analyses.

Rule 9: Do not place IPD datasets in public repositories
Comment: Unrestricted public use is precluded because IPD datasets do not meet the requirements for public deposit. The National Human Subjects Protection Advisory Committee recommendations on public use data are as below (28-29 January 2002 meeting; http://www.hhs.gov/ohrp/nhrpac/documents/dataltr.pdf).

When IRBs are asked to authorize public data files from data originally collected with identifiers, the following factors should be considered by the IRB to be certain the data files has been effectively de-identified for analysis by secondary users.

(a) removal of any identifiers of a human subject or of persons named by a human subject
(b) removal of any variables that by definition would serve as surrogates for the identity of a human subject
(c) collapse or combine categories of a variable to remove the possibility of identification due to a human subject being in a small set of persons with specific attributes regarding a variable (for example, due to the infrequency of subjects in a lower or upper range)
(d) collapse or combine variables to provide summary measures to mask what otherwise would be identifiable information
(e) use of statistical methods, where necessary, to add random variation with variables otherwise impossible to mask
(f) removal of any variables that could be linked to identifiers by secondary users

Rule 10: Issue datasets outside the study with use limitations

Rule 11: Do not supply IPD limited use datasets absent a signed statement from the requestor certifying that he/she will not attempt to re-identify people, that use will be limited to the purpose stated in the user request, and that the requestor will not copy or supply data to others not covered in the request

Rule 12: Do not supply an IPD limited use dataset to a requestor absent evidence of IRB approval by the requestor's IRB

Rule 13: Supply IPD electronic limited use datasets on DVDs, flash drives, or other similar medium
Comment: Supplying as attachments to e-mails increases the probability of files migrating to people not authorized to receive them.

Ninety-nine percent of the effort in preparing IPD de-identified datasets falls to the coordinating center. The goal should be to minimize that effort during the trial and as long as paper writing activities are ongoing.

Broadly, the two options for de-identified IPD data sharing are:

1. Announce the availability of de-identified IPD data and wait to prepare an IPD de-identified dataset until requests for such data are received

or

2. Prepare an IPD de-identified dataset, announce availability in publications and on a public website for the study, and wait for requests.

The advantage of option 1 is that the effort of IPD de-identification is postponed until there is a request for such data, if any. The downside is that if there are such requests the coordinating center must undertake a forced march to prepare the dataset in a timely fashion. The time crunch increases the probability of errors in the de-identification process.

The question, once a de-identified IPD dataset has been prepared is who deals with requests for the dataset. The option is to turn the job over to the sponsoring agency or to a custodian appointed by the agency or handle the requests in the data center. The appeal in "turn over" is that it takes the data center out of having to deal with requests. But the downside is in the detail needed by the custodian for handling requests and in the reality that

the custodian cannot answer questions users are likely to have. That being so, there is not much to be gained by deposit with a custodian.

The effect of the NIH data sharing policy and HIPAA privacy requirements has been to make old-fashioned IPD data sharing impossible. One can plausibly argue that the greatest societal good comes from partnerships where both the data giver and data receiver have something to be gained scientifically. The most useful data for analyses from the perspective of trials are identified (except for stripping of personal identifiers mentioned in Rule 6). The real question today is whether any IRB would approve such use without new consents (impractical) informing subjects of the sharing.