(Friday) 10 December 2010

Memorandum

To: CLM

Fr: Curt Meinert

Re: Options for data sharing

Assuming, as data custodians, we have the right to distribute data outside the study investigatorship, by virtue of having consents indicating intent to do so or that we have been asked to do so in writing by our NIH sponsor and that our IRB approves of such distributions, then the issue is when and how.

As you noted in your memo of 4 December 2010, the problem with the NIH data sharing policy is that it is based on a simplistic view of the publication process in trials, ie, that there is a single "main findings" publication and that the dataset is final when the paper is published. As you note, the reality is that there may be multiple "main findings" publications and that the study dataset may never be "final". That being so, one is left "explaining" to our NIH sponsors as to why data have not been deposited, even though results have been published. The procrastination on deposit means that de-identified IPD (individual patient data) datasets, when finally available, are not of interest to anyone.

The way around the problem is making IPD datasets available in relation to primary study publications and by announcing availability of such datasets in publications. The advantages are several.

First, the approach sidesteps the issue of "when" because deposits are tied to publications; not simply to the "main findings" publication.

Second, it sidesteps the issue of when a dataset is "final" in that every dataset deposited is "final" in relation to the publication to which linked.

Third, data not contained in analysis datasets of published papers remain locked, preserving investigator rights of primacy for data not yet published.

Fourth, the approach means that data made available have been subjected to the editing and cleaning processes inherent in analysis and paper writing. Depositing raw data at the "end" of a study, means that the dataset will contain "clean" and "dirty" data depending on whether or not analyzed. Mixing the two types reduces the utility of the dataset by virtue of the mix.

Fifth, the approach takes the HIH sponsor off our backs, because the data sharing is in direct accord with the NIH policy on data sharing.

The downside is that the de-identification process necessary for deposit has to be done for each dataset deposited, rather than only once at the "end" of the study. The upside is that the counts represented in datasets match those in published papers. The counts represented in a dataset prepared at the end of the study will not match counts in any publication.

The reality, by tying deposit to publication, is that there is no data sharing if a group does not publish. The meta-analysts and systematic reviewers, committed to capturing data from every trial, whether published or not, will object, but data not analyzed and published are of questionable value in any meta-analysis. In any case, one can argue that meta-analysts and systematic reviewers should not be mixing unpublished data, of questionable value and reliability, with published data, to say nothing about their ability to get them if those who collected them do not have the will to analyze and publish them.

If there is any "societal good" to be gained from data sharing, it almost certainly has to be greater when done in relation to publications than when unrelated to publications.

Last, leastwise you think I have had some epiphany and that I am now four square behind mandated data sharing, I am not! I do not believe the gain is worth the pain. Maybe, I could be convinced otherwise if there were publications on what we have learned from deposited datasets from trials to date, but they are nonexistent so far as I can tell.

But, if we are forced to share data by virtue of how we are funded, then we should be doing it in ways that have the greatest relevance to the trials to which the data pertain. A dataset produced at the end of a trial, matching nothing published, is hardly the best way to address that aim.