JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Department of Epidemiology
Johns Hopkins Bloomberg School of Public Health
415 N. Washington Street, 2nd Floor
Baltimore, Maryland 21231

12 February 2020

Memorandum

To: Trialists

Fr: Curtis Meinert

Re: Archiving

So when a trial is finished, money has run out, and investigators have dispersed, what do you archive for safe keeping? Things I would archive include the following:

Finished datasets (identified and deidentified)
Study protocol and revision history
Consent forms and history of changes
Data collection forms and history of revisions
Study curriculum vitae
Study design synopsis
Policy and procedures memoranda
Funding history
Randomization procedure
IRB approval history
Registration history
Posted results to CT.gov
Publications and presentations

There is no Smithsonian Institute for trials. You and the study sponsor are on your own when it comes to archiving.

Archiving is like packing for a trip to outer space without any packing guidelines or notion of what you might need or when.

At a bare minimum you should archive the finished identified datasets before you turn the lights out. You never know when problems arise. For example, investigators in VIGOR (Vioxx Gastrointestinal Outcomes Research) published their results Nov 2000 (NEJM). The NEJM expression of concern regarding counts in VIGOR came five years later.

I have been reminded of archiving twice in the last couple of months by requests for data for the University Group Diabetes Program (UGDP) and the Coronary Drug Project (CDP). The UGDP ran from 1960 to 1978 and the CDP from 1966 to 1985. The CDP request was for an identified dataset suitable for mortality followup of CDP enrollees. Deidentified data for the CDP are available but raw identified data went away, I surmise when the place housing the coordinating center closed years ago.

I have a copy of UGDP data files on my office shelf, but nobody can read the files. Data are on one of those reels you saw in run ups to TV shows with a monster computer with reels of tape spinning.

When we moth-balled the UGDP we deposited a hard and magnetic copy of the dataset at the National Technical Information Service (NTIS) but it would be tedious to harvest information from the print copy and I doubt you can find anyone who can read computer tapes today. The cave people had a more robust system of preserving information than we have in this electronic age.

There are commercial enterprises where deidentified data can be stored, but archiving identified data is different because the files contain personal identifiable information. Basically, the only place where such data can be archived is where they were received and processed. This is possible if the data coordinating center is part of a standing organizational structure. It is not for stand-alone centers with infrastructures that cease to exist when the trial is finished.

Other documents in the list above can be archived electronically in the coordinating center, in the institution housing the coordinating center, or with the study sponsor.

The IRS says that we should keep tax records at least three years after filing. My rule for study data would be to keep deidentified data for at least ten years from completion of a trials and to indicate where those data reside? And how about a field on CT.gov that indicates where deidentified data are stored?