JOHNS HOPKINS
UNIVERSITY

Center for Clinical Trials

Department of Biostatistics
Department of Epidemiology
Department of International Health

Department of Medicine
Department of Ophthalmology
Oncology Center

(Wed) 20 July 2011

Memorandum

To: Trialists

Fr: Curt Meinert

Re: On re-identification of de-identified data

I am attaching a story that I ran across the other day apropos of this topic (*The Fellow Who Did*). The story underscores the risk of re-identification by errant or rogue users. The problem was not with the person receiving de-identified data, but rather with the student to whom passed.

To be sure, the risk of breaches of confidentiality exist with any use of study data, whether within or outside the study research group, but the dynamics of dealing with breaches are different if they occur outside the research group. If the breach had occurred by an errant user affiliated with the Boston clinic, the issue would have been dealt with by the Boston site, and would not have involved the director of the coordinating center, his IRB, and the officers of the study and their IRBs.

Unless the NIH is willing to indemnify coordinating centers against legal action due to breaches, datasets for public use should be heavily redacted to make re-identification unlikely if not impossible. The smaller the dataset and/or the greater the stigma associated with breaches, the greater the redacting and collapsing. Obviously, the more redacting and collapsing, the less the value of deposited data. In some cases, the concern regarding privacy may preclude any deposit, eg, as would likely be the case in a study of sexual perversion.

Assuming datasets in coordinating centers are assembled devoid of study subjects' names, social security numbers, addresses, and other personal identifiers, then the question is what other potentially identifying data should be recoded, redacted, or collapsed for public use?

Clearly, clinic figured heavily in the re-identification of the singer in the attached story. Hence, study Id number, if structured to identify study clinic should be redacted and replaced with a number not identifiable to clinic. Other codes, such as letter codes, used to ensure proper record linkage during data collection, should be redacted, as should be gender and ethnic origin. Age should be redacted or collapsed depending on the size of the dataset. In the case of a study like ADAPT (Alzheimer's, Disease Anti-inflammatory Prevention Trial), that involves a condition with stigma, one likely redacts other variables easily ascertained outside the study, such as marital status, education level, and employment status.

If a dataset involves famous persons, like the singer in the attached story, it is prudent to redact the entire record of such persons.

Obviously, the accompanying documentation, in relation to public use datasets, should indicate what has been redacted and collapsed and should provide raw frequency distributions of variables before redacting or collapsing.

If the consent documents do not include statements indicating that data will be provided to persons outside the research group, study investigators should be required to clear deposits with their IRBs. If an IRB does not approve, data from that clinic should be redacted before deposit.

The NIH data sharing policy and implementation guidance (5 March 2003) recognizes the need for protecting the privacy of information collected on study subjects and for redacting to minimize the risk of "deductive identification".

> *The rights and privacy of human subjects who participate in NIH-sponsored research must be protected at all times. It is the responsibility of the investigators, their Institutional Review Board (IRB), and their institution to protect the rights of subjects and the confidentiality of the data. Prior to sharing, data should be redacted to strip all identifiers, and effective strategies should be adopted to minimize risks of unauthorized disclosure of personal identifiers. Stripping a dataset of items that could identify individual participants is referred to by several different terms, such as "data redaction,"de-identification of data," and anonymizing data. In addition to removing direct Identifiers, e.g., name, address, telephone numbers, and Social Security Numbers, researchers should consider removing indirect identifiers and other information that could lead to "deductive disclosure" of participants' identities. Deductive disclosure of individual subjects becomes more likely when there are unusual characteristics of the joint occurrence of several unusual variables. Samples drawn from small geographic areas, rare populations, and linked data sets can present particular challenges to the protection of subjects' identities.*
>
> *Investigators may use different methods to reduce the risk of subject identification. One possible approach is to withhold some part of the data. Another approach is to statistically alter the data in ways that will not compromise secondary analyses but will protect individual subjects' identities. Alternatively, an Investigator may restrict access to the data at a controlled site, sometimes referred to as a data enclave. Some investigators may employ hybrid methods, such as releasing a highly redacted dataset for general use but providing access to more sensitive data with stricter controls through a data enclave.*

I welcome comments. Absent comment it is hard to know if I have lost my mind on this issue.

The story of the fellow who did

Once upon a time, there was a fellow who wanted to be an engineer but ended up doing clinical trials. Eventually he came to head the coordinating center for the XYZ Alzheimer's Disease Prevention Trial (XYZADPT). The trial was funded by the NIH and started before the NIH policy on data sharing was announced and, hence, used consents not mentioning data sharing.

The trial involved six clinics that, together, enrolled 2,500+ persons into the trial over a four year period, starting in early 2000. Followup ended early 2007.

Subsequently, the funding agency started prodding the fellow to prepare and deposit a de-identified dataset for use outside the study research group. The fellow was reluctant to comply with the request due to concerns regarding risks of violating patients' rights to privacy seeing as consents did not mention data sharing, but eventually relented being careful to follow the yellow brick road of HIPAA de-identification and careful to get approval of his IRB before sending the dataset off to the repository specified by the funding agency for repose of datasets.

About a year later, the fellow was at the check-out counter of his "favorite" supermarket, waiting for the woman ahead of him to sort her coupons. It was during the wait that his gaze landed on the collection of rags masquerading as magazines at the check-out counter. The *National Enquirer* featured a full page cover of a famous black singer with the banner headline DEVASTATED. The singer was one of the fellow's idols years back so he peeked inside to see what the devastation was about. The fellow was saddened to learn that the singer was suffering from AD and flabbergasted to see reference to the XYZADPT and the singer's participation in it.

After furtively glancing about to make certain no one was looking, he slipped the copy onto the check-out counter.

About a week hence the fellow was summoned to his IRB about the article and a few weeks later he learned that he, his institution, and the XYZADPT officers and their respective institutions were named in a breach of confidentiality lawsuit brought by relatives of the singer.

"And now the rest of the story" in the words of Paul Harvey.

About a year before the story ran, a professor at Anything Goes University (AGU) requested the XYZADPT dataset from the repository. The professor used the dataset to address his questions and then "filed" it. Some months later he dug it out and handed it to a graduate student of his just starting work on an AD dissertation.

The student's wife was a journalism student and a close friend of a free lance reporter who, at the time, was working on a story concerning a famous black singer living with AD. The reporter, during background work, had learned that the signer had been enrolled in an NIH AD prevention trial in the early 2000s. Knowing that her girl friend's husband was working on an AD dissertation, the reporter asked if her husband might know of a such a trial.

A few days later, the student's wife passed her friend's question onto him. His curiosity was peaked seeing as, just a few days back, he had been given the XYZADPT dataset by his professor. He asked why the reporter wanted to know, whereupon his spouse told him that she was working on a story about a famous black singer who had been diagnosed with AD while enrolled in a trial to prevent AD.

One thing led to another and before long the student was on his laptop in the XYZADPT dataset.

His wife told him that the reporter knew the singer had been studied in a Boston clinic. The dataset and related data dictionary did identify clinics by location, but study publications did so the student matched published enrollment figures with counts by sorting on clinic. Matching the numbers to those published allowed the student to identify clinics by name. With the Boston clinic identified, the student limited his search to black males with AD in that clinic. There was only one such person.

The student's wife also volunteered that her reporter friend knew that the singer was married and that he had dropped out of school after the 8th grade. The graduate student checked those fields and was satisfied he had found the singer in his dataset.

The next time the graduate student's wife saw her free lance reporter friend, she told her what her husband had found and the rest is history.

When the smoke finally settled and the music stopped, about three years after the *National Enquirer* article was published, the fellow's institution had agreed to pay an undisclosed amount for damages, including court costs incurred by the Boston site because data were deposited without approval of the Boston IRB. The settlement required the fellow's institution to admit negligence in allowing data to be deposited without the express approval of the Boston IRB.

A side bar: While deposit was at the NIH's urging, the NIH remained silent when the shooting started. The government has to agree to be sued. It declined the invitation.

The would-be engineer turned trialist was barred from serving as an investigator on any research involving human beings by his IRB. The legal representatives of the fellow's IRB argued that the fellow had been derelict in his duties in protecting the confidentiality of information in his keeping because he failed to recognize the risk of probabilistic identification.

The fellow retired and faded away without as much as a Timex Watch from the institution as a parting memento.

Moral: Don't be duped into believing de-identification will save you if lawyers come calling.