Data repository Summary

Compiled by Jacqueline B. Persons, Ph.D. and Travis Osborne, Ph.D. of the Behavioral Health Research Collective, who relied on input from Jeffrey Cohen, Ph.D. (cohenj@thehrpconsultinggroup.com) of the HRP Consulting Group, Inc., Kristina Borror (kristina.borror@hhs.gov) of the Office of Human Research Protection, and Chris Apgar (capgar@apgarandassoc.com and 503-384-2538) of Apgar and Associates, a consulting firm specializing in HIPAA issues, and others.

The data repository is a de-identified database of clinical data, and the rationale for creating it is that no IRB review is required for research conducted using de-identified data. However, even though IRB review is not needed for research conducted using the data repository data, it is ideal to get an IRB review of the policies and procedures, consent document, and ongoing activity of the data repository mechanism itself.

A data repository is basically a concept. It sounds like a physical object in a certain place (and in fact was at first, when it was developed for the study of tissue samples), but in the case of behavioral health research, it is mostly a concept (a “virtual repository”). What this means is that there does not have to be an actual separate data database that houses the data for the repository. In fact, data stored in clinical records can be viewed as a data repository. Then someone (ideally someone who is not on the research team) can pull the data from the clinical records into a data repository or research database on an as-needed basis for research.

In fact, this strategy for creating a data repository (keep the data in the clinical file and then move it into a de-identified database when it is time to do the study) is the best way to create a data repository so that a linking key (key that identifies the subjects in the database) is not needed. If the researcher has access to the linking key, then the database is not de-identified.

The 3 components of the data repository that the IRB oversees:

1. Informed consent. Patients have to give consent for their data to be put into the repository and be given some information about how it will be used. The consent should give some information about what types of studies will be done with the data, and may give the potential participants a range of options: participate in all possible studies, or just some, or none. Patients must have an option to opt out. This consent can be given in the treatment agreement that patients review and sign at the onset of psychotherapy. It is ideal to write the consent document at a reading level that makes it accessible to most consumers.

2. Policies and procedures. The investigator or his/her institution must develop policies and procedures for the repository, including what are the security precautions, what are the

policies and procedures to be followed to enter the data (are they blind, are the data de-identified), what policies govern release of the data, and who has access to the data. The institution's policy might typically state that no formal review process is needed to put data in the repository, but that an IRB or some other review is needed to pull data out of the repository. Someone in the organization should oversee the data repository, especially the process of pulling data out of the repository, and the organization's policy should specify who that person is and how that person is selected. Some security consultation may be needed to be sure that procedures provide adequate security.

3. Ongoing oversight. An IRB provides ongoing oversight that the procedures and policies for the data repository are being followed appropriately, particularly when data are extracted. This could involve some kind of yearly report from an organization about how many subjects had data put in a repository, how many subjects had data pulled out and for what purposes, how were the data stored, what consent document the participants signed, what were the security measures, etc.

Regulations to be attended to when developing all these things (consent, procedures, oversight) include: OHRP, HIPAA, and state medical record regulations, and state IRB if relevant.

De-identified database

If you are creating a de-identified database, the data can be de-identified before they go into the repository, or later, before the data are released to the investigator.

The database can be a de-identified database even if a linking code is available so long as the investigator does not hold the linking code.

If the investigator is also one of the clinicians providing data for the repository, s/he may be able to recognize his/her patients in the database even without a linking code. In this situation, the data are not truly de-identified and an IRB review is needed. But if the therapist's patient data are in a larger dataset and it is unlikely that the therapist investigator can identify his/her subject, then no IRB review is needed.

Compliance with HIPAA Privacy Rule requirements

Chris Apgar points out: From a HIPAA perspective, the document we drew up that we view as a consent document, and that OHRP viewed as a consent document, is not a consent document; it is an authorization. In HIPAA, consent is for treatment, payment, or health care operations. This document simply gives permission for use of health care information for secondary purposes, so is best called an authorization from HIPAA's point of view. Chris agreed that the exact name of the document was not crucial, and he was in support of the BHRC plan to call it both a consent and an authorization, because from the OHRP point of view, it is best viewed as a consent document.

Apgar recommends encrypting the computer the database is stored on or at least encrypt the database and encrypt the thumb drive. That provides you greater protection against unauthorized access.

We do need to attend to the method of de-identifying the data. However, once the data are de-identified, they are no longer they are no longer considered PHI.

He recommends using the safe harbor method or a statistically valid methodology to de-identify the data. For more information:

Here's a link to the guidance issued by OCR: http://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html The safe harbor method would be to eliminate all identifiers that make up PHI. They include:

(2)(i) The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:

(A) Names	
(B) All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000	
(C) All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older	
(D) Telephone numbers	(L) Vehicle identifiers and serial numbers, including license plate numbers

(E) Fax numbers	(M) Device identifiers and serial numbers
(F) Email addresses	(N) Web Universal Resource Locators (URLs)
(G) Social security numbers	(O) Internet Protocol (IP) addresses
(H) Medical record numbers	(P) Biometric identifiers, including finger and voice prints
(I) Health plan beneficiary numbers	(Q) Full-face photographs and any comparable images
(J) Account numbers	(R) Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of this section [Paragraph (c) is presented below in the section "Re-identification"]; and
(K) Certificate/license numbers	
(ii) The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.	

The problem with the safe harbor method is it may render the data unusable. A statistically valid method of de-identification would involve using an established method of de-identification of PHI that renders the data such that a reasonable person could not re-identify the data. OCR's guidance states:

(b) *Implementation specifications: requirements for de-identification of protected health information.* A covered entity may determine that health information is not individually identifiable health information only if:
(1) A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:
(i) Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and
(ii) Documents the methods and results of the analysis that justify such determination; or . . .

In the case of the types of research that those of us at the BHRC are doing, we would likely want to include dates of therapy sessions in our database, and yet including these dates seems to pose a low risk of identifying the individual, and so we can ask an expert to certify that for us. Chris Apgar agreed to do that for us.

In the HIPAA literature, there's a concept of granularity, and that refers to the likelihood of any given cell being able to identify any particular individual. For example, if you code age range instead of age in the dataset, then the likelihood of identification of the subject drops.

In these days of computer technology, it is easy to re-identify data. The requirement that the investigator who receives the de-identified database will not re-identify it could be written into the organization policies or the IRB requirements/policies for when data are released.

Research ideas that occur after consent is given

PIs might get research ideas later, after the data are collected, that they did not describe to the subjects and get consent for. Three ways to handle this problem include going back to the research subjects to get their consent, asking the IRB to give a waiver of consent for this use, and writing the consent document in a way that the participant gives permission for future studies that have not been formulated yet.

Contacting patients who are in the de-identified database

If the researcher has findings that they need to get back to the participant (e.g., the research shows that the patient is at risk for some big-ticket problem), if the database is truly de-identified, then it is not possible to contact the patient.

Can participants revoke consent?

There is no standard, but people have to be told ahead of time what the policy will be regarding this issue. It is possible to tell people that once their data have been put in the database they cannot be removed, or that future data will not be included, but past data will not be removed.

Who owns the data?

The issue of who "owns" the data or has access to the data depends on how the data repository is set up. That's described in the policies and procedures that govern the data repository. So who owns the data or has access to the data could vary from center to center depending on the policies that each local institution has developed for that data repository. The primary focus of the IRB when providing review and oversight of these policies is what participants were told at the time of providing consent to include their data in the repository and whether the institution followed these policies when extracting and using these data for research purposes.