

# Effective Health Care Program Research Reports

Number 42

## Options for Developing a Repository of Expired Patient Registries

Richard E. Gliklich, M.D.  
Michelle B. Leavy, M.P.H.  
Laura Khurana  
Daniel Levy, M.S.  
Daniel M. Campion, M.B.A.

Research from the Developing Evidence to Inform Decisions about Effectiveness (DEcIDE) Network



**Agency for Healthcare Research and Quality**  
Advancing Excellence in Health Care • [www.ahrq.gov](http://www.ahrq.gov)

October 2012

The DEcIDE (Developing Evidence to Inform Decisions about Effectiveness) network is part of AHRQ's Effective Health Care Program. It is a collaborative network of research centers that support the rapid development of new scientific information and analytic tools. The DEcIDE network assists health care providers, patients, and policymakers seeking unbiased information about the outcomes, clinical effectiveness, safety, and appropriateness of health care items and services, particularly prescription medications and medical devices.

This report is based on research conducted by the Outcome DEcIDE Center under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2005-0035-I). The AHRQ Task Order Officer for this project was Elise Berliner, Ph.D.

The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or the U.S. Department of Health and Human Services.

None of the authors have a financial interest in any of the products discussed in this report.

Persons using assistive technology may not be able to fully access information in this report. For assistance contact [EffectiveHealthCare@ahrq.hhs.gov](mailto:EffectiveHealthCare@ahrq.hhs.gov).

**Suggested citation:**

Gliklich RE, Leavy MB, Khurana L, Levy D, Campion DM. Options for Developing a Repository of Expired Patient Registries. Effective Health Care Program Research Report No. 42. (Prepared by the Outcome DEcIDE Center, under Contract No. 290-2005-0035-I.) AHRQ Publication No. 12(13)-EHC112-EF. Rockville, MD: Agency for Healthcare Research and Quality. October 2012. Available at: [www.effectivehealthcare.ahrq.gov/reports/final.cfm](http://www.effectivehealthcare.ahrq.gov/reports/final.cfm).

# Options for Developing a Repository of Expired Patient Registries

## Abstract

**Objectives.** Patient registries have received significant attention in recent years and are being used for a variety of purposes. In discussions for a previous Agency for Healthcare Research and Quality (AHRQ) report, stakeholders suggested that a voluntary repository of expired (i.e., closed) patient registries might be a useful tool for ensuring that registry databases remain accessible for future research. The concept of a repository of expired registries raises several questions related to the feasibility, value, and potential cost of such a repository, as well as issues related to research ethics, governance, data access and use, patient privacy, technical requirements, legal considerations, and incentives for donating data to the repository. The purpose of this project is to examine these questions, explore options, and provide actionable information to AHRQ for developing a repository of expired patient registries, should it be determined that such a repository would be both feasible and valuable.

**Data Sources.** Not applicable.

**Methods.** Background research was conducted using literature reviews, Internet searches, and discussions with stakeholders and other relevant organizations. This research focused on identifying relevant, existing repositories of clinical study data and publications that discussed the major issues involved in setting up such a repository. Stakeholder engagement also was a key component of this project. Stakeholder perspectives were gathered through discussions and an in-person meeting. The in-person stakeholder meeting included individuals representing academia and research, government and funding agencies, industry, health care providers and provider organizations, patient/consumer organizations, payers, journal editors, and legal and patient privacy experts. The objective of the meeting was to provide a forum for stakeholders to discuss major issues related to creating a repository of expired registries and to assess the feasibility and potential value of such a repository.

**Results.** Background research and stakeholder feedback clearly demonstrated that there is interest in the research community in a repository of expired registries, and that such a repository is feasible from operational, legal, ethical, and technical standpoints. Stakeholders noted the potential value of a repository of expired registries and identified key objectives of such a system. Despite the interest in this concept, stakeholders did not see a current business case for the repository, making it unlikely that a private entity would develop such a repository on its own and suggesting that the repository would need to be developed and supported with government funding. Stakeholders described a range of possible options for developing and maintaining the repository that were classified into two anchoring models on each end of the spectrum: a basic data archive model and a data archive with research support model.

**Conclusions.** The report concludes that there is a clear interest among stakeholders in a repository of expired registries, but the lack of incentives for registry owners to donate their data is a critical barrier to a successful program.

## **Contents**

Executive Summary .....	ES-1
Introduction.....	1
Rationale.....	2
Project Objectives .....	3
Methods.....	4
Background Research .....	5
NIDDK Central Data Repository .....	5
Overview .....	5
Governance and Oversight .....	6
Patient Privacy .....	6
Research Ethics.....	6
Data Access and Use .....	6
Technical Considerations .....	7
Resources .....	7
National Database for Autism Research .....	8
Overview .....	8
Governance and Oversight .....	8
Patient Privacy .....	8
Research Ethics.....	9
Data Access and Use .....	9
Technical Considerations .....	9
Resources .....	10
NHLBI Biologic Specimen and Data Repository .....	10
Overview .....	10
Governance and Oversight .....	10
Patient Privacy .....	10
Research Ethics.....	11
Data Access and Use .....	11
Technical Considerations .....	11
Resources .....	11
Relevance of Existing Repositories for a Repository of Expired Registries.....	12
Stakeholder Perspectives on Major Issues .....	13
Governance.....	13
Repository Owner.....	13
Governing Body .....	13
Review of Data Requests.....	14
Ownership and Management of Donated Data.....	14
Incentives for Donating Registry Data .....	15
Summary of Governance Issues .....	15
Patient Privacy.....	16
Storage of Patient Identifiers .....	16
Potential Re-Identification of Patients.....	17
Institutional Certification.....	17
Summary of Patient Privacy Issues .....	19

## *Effective Health Care Program Research Report Number 42*

Research Ethics .....	19
Use of Data for Valid Research Projects .....	19
Identifying Ethical Issues With Donated Data .....	20
Summary of Research Ethics Issues .....	20
Data Access and Use .....	20
Conditions for Data Access .....	21
Storage of Informed Consent Information.....	21
Storage of Data Documentation .....	21
Summary of Data Access and Use Issues.....	22
Technical Considerations .....	22
Data Storage and Security .....	22
Metadata .....	22
Technology Standards .....	23
Transfer Procedures .....	23
Summary of Technical Considerations.....	23
Summary of Stakeholder Perspectives .....	23
Legal Issues in Creating a Repository of Expired Registries .....	24
Federal Laws .....	25
U.S. Department of Health and Human Services (HHS) Regulations Governing Uniform Administration of Federal Awards .....	25
HIPAA .....	26
The Common Rule.....	30
Special Considerations Related to the Disclosure of Alcohol or Substance Abuse Registry Data.....	31
Privacy Act of 1974.....	32
Federal Information Security Management Act of 2002 (FISMA) .....	32
Paperwork Reduction Act (PRA) .....	33
Freedom of Information Act (FOIA).....	34
State Law .....	35
Example of a Federally Administered Data Repository Holding Data From Multiple Studies	36
Summary of Legal Considerations .....	37
Proposals for a Repository of Expired Registries .....	39
Should There Be a Repository of Expired Registries?.....	39
What Are the Requirements for a Repository of Expired Registries? .....	40
Basic Requirements .....	41
Additional Requirements To Meet Secondary Objectives .....	42
What Models Might AHRQ Use To Develop and Maintain a Repository of Expired Registries? .....	42
Model 1: Basic Data Archive .....	42
Model 2: Data Archive With Research Support.....	43
What Would Motivate Researchers To Donate Data to the Repository of Expired Registries? .....	45
What Are the Costs of Setting Up and Maintaining a Repository of Expired Registries? .....	46
Conclusions.....	47
References.....	48

## **Tables**

Table 1. Metadata publicly available for NIDDK central data repository (cdr) datasets .....	7
Table 2. Metadata publicly available for NHLBI biologic specimen and data repository datasets .....	11
Table 3. Comparison of repository of expired registry models .....	44
Table 4. Estimated costs by model of setup and maintenance for a repository of expired registries.....	46

## **Figure**

Figure 1. Meeting participants by stakeholder group .....	4
---	---

### **Author affiliations:**

Richard E. Gliklich, M.D.<sup>1</sup>

Michelle B. Leavy, M.P.H.<sup>1</sup>

Laura Khurana<sup>1</sup>

Daniel Levy, M.S.<sup>1</sup>

Daniel M. Campion, M.B.A.<sup>1</sup>

<sup>1</sup>Outcome DEcIDE Center, Cambridge, MA

## **Executive Summary**

Patient registries have received significant attention in recent years and are being used for a variety of purposes, including accruing additional evidence on the effectiveness of new interventions in particular populations, understanding practice patterns and adherence to guidelines, examining patient outcomes, supporting safety surveillance initiatives, and demonstrating value for reimbursement. Given the large amount of funding currently devoted to patient registries and the significant contributions (data and effort) from patients and healthcare providers to these registries, it is important to ensure that the broader societal value that can be derived from registry databases is maximized and ideally not lost when the registry expires. A registry may expire for any number of reasons. For example, the registry may fulfill its purpose, exhaust its funding, or find that its objectives become less scientifically relevant due to changes in treatment patterns (e.g., the product under study is superseded by a new treatment). In discussions for a previous Agency for Healthcare Research and Quality (AHRQ) report, stakeholders suggested that a voluntary repository of expired patient registries might be a useful tool for ensuring that registry databases remain accessible for future research. The concept of a repository of expired registries raises several questions related to the feasibility, value, and potential cost of such a repository, as well as issues related to research ethics, governance, data access and use, patient privacy, technical requirements, legal considerations, and incentives for donating data to the repository.

The purpose of this project is to examine these questions, explore options, and provide actionable information to AHRQ for developing a repository of expired patient registries, should it be determined that such a repository would be both feasible and valuable. Background research was conducted using literature reviews, Internet searches, and discussions with stakeholders and other relevant organizations. This research focused on identifying relevant, existing repositories of clinical study data and publications that discussed the major issues involved in setting up such a repository. Several existing repositories of clinical study data were identified. The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Central Data Repository, the National Database for Autism Research (NDAR), and the National Heart, Lung and Blood Institute (NHLBI) Biologic Specimen and Data Repository are profiled in this paper because they represent models that are particularly relevant to AHRQ in terms of scale, focus, operational approach, and technical structure.

Stakeholder engagement was a key component of this project. Stakeholder perspectives were gathered through discussions and an in-person meeting. The in-person stakeholder meeting included individuals representing academia and research, government and funding agencies, industry, health care providers and provider organizations, patient/consumer organizations, payers, journal editors, and legal and patient privacy experts. The objective of the meeting was to provide a forum for stakeholders to discuss major issues related to creating a repository of expired registries and to assess the feasibility and potential value of such a repository.

Topics covered in the meeting included governance, patient privacy, research ethics, data access and use, technical requirements, and legal considerations. Sound, transparent governance policies and procedures that address data access rules, data ownership, and incentives for donating data were identified as a critical component of a potential repository. Stakeholders also suggested that the governing body include a broad group of stakeholders. To protect patient privacy, the repository could store only de-identified data; however, this would limit the usefulness of the data for linkage projects. Alternately, the repository could store patient identifiers, provided that the informed consents supporting the original data collection allowed

for this use. An institutional certification model could be used to determine whether data could be donated and under what conditions donated data could be used for new research. Regarding data access and use, stakeholders noted that data access requests would need to be accompanied by institutional review board approval for the proposed research project and recommended that the repository store information on the original study and how the data were collected to ensure that researchers can draw valid conclusions from the data. The repository would also need to store some level of metadata to allow for searching of the repository data assets. Various technology models could be used, and the selection of the model and the related technology requirements would be driven by resource considerations.

Regarding legal considerations, stakeholders noted that numerous Federal and state laws potentially are relevant. Additional legal analysis completed for this project found that the critical factors are whether the data will be individually identifiable, how data will be maintained and used, whether the uses of the data will include research, and the sources of the data. This paper discusses Federal laws related to government agency access to and oversight of federally supported projects and the information produced under such projects; the Health Insurance Portability and Accountability Act Privacy, Security, and Breach Notification Rules; the Common Rule; “Part 2” regulations related to alcohol and substance abuse treatment information; the Privacy Act of 1974; the Federal Information Security Management Act; Federal laws related to information collection (e.g., the Paperwork Reduction Act); the Freedom of Information Act; and state confidentiality laws. Examples of ongoing Federal repositories underscore that these potential legal issues can be managed successfully.

Background research and stakeholder feedback clearly demonstrated that there is interest in the research community in a repository of expired registries, and that such a repository is feasible from operational, legal, ethical, and technical standpoints. Stakeholders noted the potential value of a repository of expired registries and identified key objectives of such a system. Despite the interest in this concept, stakeholders did not see a current business case for the repository, making it unlikely that a private entity would develop such a repository on its own and suggesting that the repository would need to be developed and supported with government funding. If AHRQ were to pursue the development of a repository, the Agency could present a clear rationale based on the importance of data sharing for the efficient use of limited health research resources and the Agency’s continuing efforts to encourage the use of high-quality registries for clinical research.

Stakeholders described a range of possible options for developing and maintaining the repository that were classified into two anchoring models on each end of the spectrum: a basic data archive model and a data archive with research support model. The data archive model represents a low-budget, basic approach to data archiving, while the data archive with research support model represents a high-budget, sophisticated approach. Many intermediate models, which would use the basic approach and incorporate one or more features from the sophisticated approach, are possible. This paper examines the strengths and limitations of the models, discusses possible incentives for donating data under each model, and considers the estimated costs for developing and operating each model.

The report concludes that there is a clear interest among stakeholders in a repository of expired registries, but the lack of incentives for registry owners to donate their data is a critical barrier to a successful program. AHRQ may be able to require donation of data from registries funded by the Agency, but few incentives exist to encourage other registry sponsors to contribute their data. Prior to investing resources in developing a sophisticated repository, it may be



necessary to work with stakeholders to identify donation incentives to ensure that the repository has sufficient data to support research projects. Alternately, AHRQ may pursue an incremental approach to developing the repository, in which the initial repository follows the basic data archive model and additional features are added as the repository grows and its use requirements are better understood. Further discussions with both stakeholders and registry owners are highly recommended to help AHRQ further define requirements, priorities, incentive structures, and funding options in planning next steps.

## **Introduction**

Patient registries have received significant attention in recent years. A patient registry is defined as “an organized system that uses observational study methods to collect uniform data (clinical and other) to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure, and that serves one or more predetermined scientific, clinical, or policy purposes.”<sup>1</sup> Common purposes for patient registries include evaluating the safety, effectiveness, or quality of medical treatments, products, and services and studying the natural history of diseases. Some registries are developed and maintained solely to assist in care delivery, coordination, and quality improvement, but many serve broader research purposes.

When properly designed and conducted, patient registries can provide unique insights into real world clinical practice, effectiveness, safety, and quality. As health care stakeholders have increasingly recognized the value of real-world evidence, interest in developing registries has grown. For example, private and public sector payers for health care are potentially interested in registries as a means to accrue additional evidence on the effectiveness of new interventions in particular populations.<sup>2</sup> The Centers for Medicare and Medicaid Services (CMS) Coverage with Evidence Development (CED) is an example of a program that will sometimes require a registry for this purpose.<sup>3</sup> Medical and patient associations are turning to registries as a means of understanding practice patterns, adherence to guidelines, and patient outcomes.<sup>4</sup> Academia, industry, and government groups are using registries for a variety of purposes, including understanding disease and treatments, conducting safety surveillance, and demonstrating value for reimbursement.<sup>1</sup>

In the past few years, registries have also received significant new attention as potential sources of data for comparative effectiveness research (CER). Reports from the Institute of Medicine (IOM) and the Congressional Budget Office (CBO) in 2007 cited the importance of patient registries in developing comparative effectiveness evidence. The IOM identified registries as an important potential data source for CER.<sup>5</sup> The CBO report noted, “Registries collect additional information that is typically not contained in claims records, such as measures of health status or test results. In the United States, a number of registries—established or managed by various entities, including medical specialty societies and product manufacturers—have been used to help determine the clinical effectiveness and cost effectiveness of various products or services.”<sup>6</sup> A 2009 white paper, generated by the Brookings Institution, described multiple categories of methods for CER including systematic reviews of existing research, including meta-analysis; decision modeling with or without cost information; retrospective analysis of existing clinical or administrative data, including natural experiments; prospective non-experimental studies, including registries that observe patterns of care and outcomes but do not assign patients to specific groups; and experimental studies, including randomized clinical trials, in which patients or groups of patients are assigned to alternative treatments, practices, or policies. It further noted that all of these categories are important, that there have been important advances in methods that improve the validity and design and, in particular, that there has been considerable progress in the design and use of clinical registries.<sup>7</sup>

The increased interest in using registries for CER has resulted in new funding opportunities. For example, the National Institutes of Health (NIH) call for Challenge Grants as a first disbursement of appropriated funding under the comparative effectiveness component of the American Recovery and Reinvestment Act of 2009 (ARRA) specifically called for a number of registries. These included such diverse areas of CER as treatments for chronic childhood

arthritis, musculoskeletal and skin disease, fibromyalgia, cancer primary prevention, cancer screening, cleft palate, diabetes, medical implants, atrial fibrillation, autism spectrum disorders, rare diseases, and intervention versus nonintervention in management of asymptomatic vascular abnormalities.<sup>8</sup> Several other grant programs using ARRA funding, such as those managed by the Agency for Healthcare Research and Quality (AHRQ), included patient registries as potential study options, and AHRQ awarded grants to new registry-based projects. A 2009 report from AcademyHealth examined the volume and range of costs of recent CER.<sup>9</sup> Using a combination of interviews with researchers and organizations and reviews of two research listing databases, the authors concluded that prospective cohort studies, registry studies, and database studies comprise the largest block of CER activities (54 percent) and that among observational studies 23 percent are registry studies.

Similarly, the U.S. Food and Drug Administration (FDA) uses registries in assessing the safety and effectiveness of drugs and devices when considering applications from manufacturers, as well as for monitoring products post-approval. For example, since 2005, the FDA Center for Devices and Radiological Health has called for some 120 post-approval studies, many of which use new or existing registries to study the real-world effectiveness of specific devices in community practice.<sup>10</sup>

## **Rationale**

The significant investment in patient registries, and particularly the investment of public funds to create new registries, raises questions about what happens to registry data when a registry expires (i.e., data collection ends and the registry closes). A registry may expire for any number of reasons. For example, the registry may fulfill its purpose, exhaust its funding, or find that its objectives become less scientifically relevant due to changes in treatment patterns (e.g., the product under study is superseded by a new treatment). A registry may collect a large amount of data on a broad patient population, sometimes over several years. The registry investigators may publish findings related to the registry objectives, but the data may also serve other purposes once the registry expires. In discussions for a previous AHRQ report, stakeholders suggested that a voluntary repository of expired patient registries might be a useful tool for ensuring that registry databases remain accessible for future research. For example, the data may be useful for new or supplementary analyses (to avoid duplication of effort), to inform protocol development for future studies, or for comparing, linking, or combining with other data to answer new research, policy, or public health questions. In 2010, AHRQ funded the development and piloting of a new Registry of Patient Registries (RoPR). The RoPR will be a searchable, central listing of patient registries in the United States, with the primary goal enabling interested parties to identify registries in a particular area to promote collaboration, reduce redundancy, and improve transparency. Secondary objectives include encouraging and facilitating the use of common data elements and definitions, providing a central repository of searchable registry findings, and serving as a recruitment tool for researchers and patients interested in participating in patient registries. Once the RoPR is created, there will be an increase in registry visibility and potentially interest from researchers who may seek data from registries that have expired for new uses.

Given the large amount of funding currently used for the development and operation of patient registries and the significant contributions (data and effort) from patients and health care providers to these registries, it is important to ensure that the broader societal value that can be derived from registry databases is maximized and ideally not lost when the registry expires. In

discussions for a previous AHRQ report, stakeholders suggested that a voluntary repository of patient registries might be a useful tool for ensuring that registry data are not lost when a registry ends.<sup>11</sup>

The concept of a repository of expired registries raises several questions. First, should there be a repository of expired registries? Is there value in developing a repository of expired patient registries, and is it feasible to develop such a repository? How would the repository address issues related to research ethics, governance, data access and use, patient privacy, technical requirements, and legal considerations? Assuming that a repository is both feasible and valuable, what are the requirements for the repository? What models might AHRQ use to develop and maintain a repository of expired registries? What would motivate researchers to donate their registries to a repository of expired registries? Lastly, what are the costs of setting up and maintaining a repository of expired registries?

## **Project Objectives**

The objectives of this project are to examine questions related to creating a repository of expired patient registries, assess the overall value and feasibility of such a repository, and explore options for developing the repository. A key component of this project is engagement with stakeholders, including Federal partners, funding agencies, industry sponsors, researchers, health care providers, payers, and patients, to ensure that their views are considered and incorporated. The goal of this paper is to provide actionable information to AHRQ for developing a data repository, should it be determined that such a repository will be both feasible and valuable, that will be relevant to the needs of the Medicare, Medicaid, and other Federal health care programs and will reflect the overall goals of the Effective Health Care program.

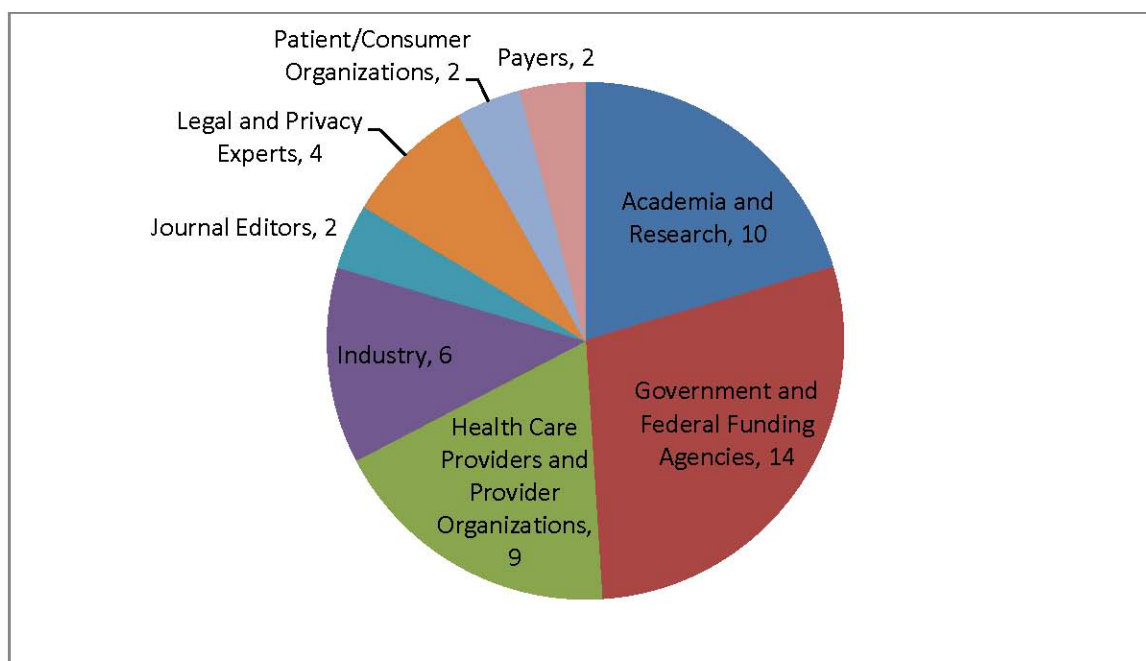
The paper begins by describing existing repositories of clinical study data and discussing how they have addressed the key issues related to governance, research ethics, patient privacy, data access and use, technical requirements, and legal considerations. The paper then discusses stakeholder input related to these issues and summarizes the conclusions related to feasibility and value. Next, the paper examines the potential legal requirements for a repository. The sections on stakeholder perspectives and legal issues cover many of the same topics and often reach the same conclusion. They are presented separately here, though, because they examine the issues from different viewpoints. The stakeholder perspectives section highlights the major concerns and suggestions of the stakeholder representatives, while the legal issues section draws on analysis of existing laws and regulations to draw conclusions as to what is legally feasible. Lastly, the paper makes recommendations for moving forward with a repository of expired registries and explores two models or approaches for such a project. Each proposal includes information on how registries would be identified for inclusion; what incentives would be available for patient registries to donate data; what information would be included; how information would be verified; how information would be provided to researchers; and the cost considerations. The paper assumes that the repository of expired registries would be located within the United States and hold data from studies conducted within the United States. While many registries are international in scope, the legal and ethical issues involved in donating and storing data from those registries in the repository are beyond the scope of this paper.

## Methods

Information for this project was gathered through literature reviews, Internet searches, discussions, and a large in-person stakeholder meeting. The literature reviews and Internet searches focused on identifying relevant, existing repositories of clinical study data and any publications that discussed the major issues involved in setting up such a repository. PubMed was used to identify relevant literature. Search engines, such as Google and Google Scholar, were used to identify existing repositories of clinical study data. In addition, an extensive search of the NIH Web site was conducted to identify relevant repositories and other documents. In some cases, discussions were held with representatives of the NIH repositories cited. The findings from the background research and discussions are summarized in the “Background Research” and “Stakeholder Perspectives on Major Issues” sections of this paper.

The in-person stakeholder meeting was held on January 18, 2011 at the AHRQ Conference Center. Participants included 49 individuals representing academia and research, government and Federal funding agencies, industry, health care providers and provider organizations, patient/consumer organizations, payers, journal editors, and legal and patient privacy experts. The breakdown of participants by stakeholder group is depicted in Figure 1.

**Figure 1. Meeting participants by stakeholder group**



The purpose of the meeting was to provide a forum for stakeholders to discuss the feasibility and potential value of creating a voluntary data repository for archiving expired patient registries. The primary objectives were to (1) discuss issues related to governance, patient privacy, research ethics, data access and use, and technical setup for a repository of expired registries, and (2) assess the feasibility and potential value of such a repository. Outcome DEcIDE Center staff moderated the meeting. The findings from the meeting are incorporated into the “Stakeholder Perspectives on Major Issues” and “Proposals for a Repository of Expired Registries” sections of this paper.

## **Background Research**

Several existing repositories of clinical study data were identified through the literature review, Internet searches, and discussions with stakeholders. Of the identified repositories, three are profiled here as representative models that are particularly relevant to AHRQ because of their scale, focus, operational approach, and technical set-up. The repositories profiled here are the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Central Data Repository, the National Database for Autism Research (NDAR), and the National Heart, Lung and Blood Institute (NHLBI) Biologic Specimen and Data Repository.

Other repositories were identified and are not profiled because they do not represent relevant models for this project. For example, the Virtual International Stroke Trials Archive (VISTA) is an international repository that houses data from completed stroke trials. Investigators from each study in the repository serve on the VISTA governing board, and the original study investigators are active participants in new research using their data. While this model has been successful, it is not highly scalable and depends on sustained investigator interest. The Inter-university Consortium for Political and Social Research (ICPSR) project, based at the University of Michigan, aims to facilitate social sciences research by storing and facilitating access to research datasets. The repository holds a large number of datasets in a range of social science areas and has detailed policies on data donation, curation, and access. While much of the data is not health-related and therefore subject to different legal and ethical concerns, this project may be a useful reference for technical considerations and governance issues, should AHRQ decide to develop a repository of expired patient registries. Biorepositories and archives that focus primarily on storing biologic and genetic specimens are also not profiled, because the legal, ethical, and operational issues involved in those repositories are, in many ways, different than those involved in storing clinical data.

### **NIDDK Central Data Repository**

#### **Overview**

In 2003, the NIDDK established Central Data, Biosample, and Genetic Repositories to house samples and clinical data collected during completed and ongoing clinical studies funded by NIDDK.<sup>12</sup> In creating these repositories, NIDDK aimed to “increase the utility of NIDDK sponsored research by providing access to the samples and data to a wider research community than those research groups involved in the studies.”<sup>13</sup> As of March 2011, datasets from 28 clinical studies are stored in the Central Data Repository (CDR), and most are from large, multi-site, clinical studies funded by the NIDDK. Forty-nine articles have been published in peer-reviewed journals using data from the CDR.<sup>14</sup> The CDR staff receives data access requests about 50 times a year, and 188 external requests for data have been processed.<sup>15</sup>

The CDR serves as the archive for clinical data, maintains the linkages between clinical data and specimens housed by the Biosample and Genetic repositories, and manages requests for samples of those specimens. The CDR’s Web site provides information to the public about available datasets (see Table 1) and facilitates data requests and submissions.

## **Governance and Oversight**

NIDDK contracts out the management of each Repository to separate contractors. The CDR is managed by RTI International. Access to CDR data is regulated by a group of reviewers and an arbiter from the NIDDK whose mission is to review data requests for scientific appropriateness and to ensure that the data requests are consistent with the patient consents given at the time of the original studies. The reviewers and arbiter are not necessarily the same individuals for each request, but are always NIDDK employees selected by the NIDDK project officers.<sup>15</sup> The reviewers offer their opinions based on the data request application and accompanying institutional review board (IRB) approval (or an IRB determination that the research is exempt from such approval), and the arbiter makes the final approval decision and notifies the contractor of that decision.

For data of significant value, the terms of NIDDK clinical research funding require that the researcher donate the data to the Repository at the close of the study; these terms are incorporated into the Notice of Grant Award. Some studies not funded by NIDDK have expressed interest in donating their data and biosamples to the CDR; in these cases, the CDR has executed a Material Transfer Agreement with the donor and amended the agreement as necessary if only data are being donated.<sup>16</sup>

## **Patient Privacy**

The CDR only accepts and houses data that have been de-identified according to the Health Insurance Portability and Accountability Act (HIPAA) requirements for limited datasets (LDS).<sup>17</sup> The CDR has returned data to study managers or data coordinating centers that were not properly de-identified or consented. Datasets accepted by the CDR have study-specific patient IDs associated with them. For samples, the CDR adds a three-digit prefix code specific to each study and to each site within a study. The result is a CDR ID that is unique across all of the samples housed in the Repositories.<sup>15</sup>

## **Research Ethics**

The CDR has established and documented procedures for processing and storing donated data, which include assessing the completeness, utility, and integrity of the data (including recreating analyses of published results); assigning and managing site IDs for the sites that contributed data to the dataset; and extracting metadata about the dataset for display on the public CDR Web site.<sup>18</sup>

Before data are released from the CDR for use in other research, the data request is reviewed to ensure consistency with the informed consent given by the patient at the time of the original study. To facilitate this, the CDR has built a consent database that is available to the NIDDK staff reviewing data requests.<sup>15</sup> A separate bioethics review is available if there are remaining questions after review of the original informed consents; thus far, such a review has not been needed.<sup>16</sup>

## **Data Access and Use**

To ensure that researchers understand the strengths and limitations of the data available in the CDR and can draw appropriate conclusions from their findings, metadata on each available dataset is provided on the CDR Web site (Table 1). CDR staff members are also

available to answer questions and to help identify the most appropriate dataset(s) for a particular research purpose.

To initiate the data request process, researchers complete a request through the Automated Data Request System (ADRS) on the Repository Web site, which requires an executive summary describing the proposed research, IRB approval (or an IRB determination that the research is exempt from such approval), a Data Use Certification (a Material Transfer Agreement), and responses to more detailed questions about the proposed research. These documents are submitted electronically through the ADRS, and original signed copies are mailed to the CDR. The CDR submits this documentation to the project officer at NIDDK, who designates a team of NIDDK reviewers to assess the request for scientific appropriateness and consistency with the consent forms used in the requested datasets. If approval is granted, the CDR is notified and then releases the data to the requestor. Data requests are generally free of charge, although a fee applies if CDR staff provides analytic support for the requestor.<sup>19</sup> Manuscripts using CDR data must acknowledge the NIDDK Repository, and in some cases, cite the specific dataset(s) from which the data were obtained.<sup>20</sup>

**Table 1. Metadata publicly available for NIDDK Central Data Repository (CDR) datasets**

Category	Description
General Description	Narrative description of study that yielded dataset
Metadata	Study design and objectives Participating centers and principal investigators Inclusion and exclusion criteria; enrollment details Outcome(s) of interest Funding organization
Protocol or Manual of Operations	Governing operational document for the study
Forms	Study case report forms (CRFs)
Publications	List of publications using study data
Roadmap	Summary of all documents and files available for the dataset
Dataset Integrity Check	Description of integrity check conducted in an attempt to duplicate published results; includes narrative summary, tables comparing published and calculated results, and SAS code used to replicate results.
Study Samples	List of biologic or genetic samples available from the study

## Technical Considerations

The CDR accepts donated data in multiple formats and converts all datasets into SAS format.<sup>17</sup> Data are sent to approved requestors via a secure FTP site; datasets are not directly accessible through a Web portal because of concerns about dataset size and health data security.<sup>13</sup> Users can browse available metadata for the datasets on the CDR Web site or perform a keyword search of the metadata, including ancillary study documents.

## Resources

NIDDK's contract to the data management vendor had a 2010 annual budget of \$962,000; this included all activities related to the CDR's multiple purposes (including coordination and administrative duties for the Biosample and Genetic Repositories), and not exclusively the curation and distribution of archived databases.<sup>16</sup> NIDDK estimates that five to six contractor employees work on the contract currently, plus two part-time project officers at NIDDK. Overall, considering the other duties of these NIDDK and contractor employees, an estimated 2.0 full-time employees (FTEs) are working exclusively on the CDR, managing



archived data from 28 studies and sample inventory from 37 additional ongoing studies in its current operational phase.<sup>16</sup>

## **National Database for Autism Research**

### **Overview**

NDAR is a biomedical informatics system and research data repository developed and housed at the NIH Center for Information Technology. Its stated purpose is to help accelerate research on autism spectrum disorders by creating an infrastructure that integrates heterogeneous datasets, allowing access to much more quality research data than individual investigators would be able to collect on their own.<sup>21</sup> As of March 2011, NDAR houses data from about 20 studies,<sup>22</sup> and another 55 studies have plans to share their data through NDAR, including many being conducted at centers operating under NIH “Autism Centers of Excellence (ACE)” grants.<sup>23</sup> Several peer-reviewed articles have been published focusing on the technology and informational modeling behind NDAR,<sup>24-25</sup> and NDAR staff report that they have received many requests to access the data, both from individuals associated with ACEs and from other researchers.<sup>22</sup>

### **Governance and Oversight**

The NDAR Governing Committee, which is responsible for the on-going management and stewardship of NDAR policy and procedures, is comprised of the Director of the National Institute of Mental Health (NIMH) and several other NIH Institute and Center Directors or their designees.<sup>26</sup> A Data Access Committee (DAC) is charged with reviewing data access requests; members of the DAC include project officers for grantees<sup>22</sup> and Federal staff with expertise in areas such as the relevant scientific disciplines, research participant protection, and privacy.<sup>26</sup>

All investigators who receive NIH support to conduct autism research are expected to submit descriptive information about their studies (such as the protocol, questionnaires, study manuals, variables measured, and other supporting documentation) for inclusion in NDAR. Investigators may be required by the “Terms and Conditions” of their grant to submit study data.<sup>26</sup>

### **Patient Privacy**

Data submitted to NDAR are required to be “de-identified such that the identities of data subjects cannot be readily ascertained or otherwise associated with the data by the repository staff or secondary data users.” Once received by NDAR, data are assigned Global Unique Identifiers (GUIDs), computer-generated random alphanumeric codes that are unique to each research participant in NDAR. A detailed description of the process for de-identifying data and for assigning GUIDs can be found in the NDAR policy document<sup>26</sup> and the 2010 article by Johnson et al.<sup>25</sup>

NDAR expects that data submission to NDAR will be explicitly discussed in the informed consent process for prospective studies that are designed with the intent to submit data to NDAR. For retrospective studies, since informed consents are less likely to have explicitly addressed data donation, it is the responsibility of the submitting institution and its IRB to determine whether a study is appropriate for submission to NDAR.<sup>26</sup> While not required, NDAR also encourages researchers submitting data to consider whether a Certificate of Confidentiality

is appropriate for their data as an additional safeguard against disclosure of research participant identities.<sup>26</sup>

## **Research Ethics**

Researchers submitting data to NDAR are required to provide a written certification from an official of the submitting institution stating that they approve of the data submission. An additional certification is required that should state that the data submission is consistent with applicable laws and regulations and that donation of the data is ethically permissible.<sup>26</sup> The certification criteria are identical to those in the NIH genome-wide association studies (GWAS) policy<sup>27</sup> and discussed in the “Institutional Certification” section of this paper.

Those wishing to access NDAR data are required to sign a Data Use Certification that confirms that they will use the data only for the approved research.<sup>26</sup> The DAC reviewing these requests determines whether the proposed use of the dataset is scientifically and ethically appropriate.

## **Data Access and Use**

Once a researcher is granted access to NDAR by the DAC, they receive an account to log into the NDAR research portal; the account allows them to simultaneously query data stored in the NDAR repository and data from external sources with which NDAR maintains a federated link. The queries produce datasets that can include clinical data; phenotypic, imaging, or genomic data; or other data such as outcomes data. Queries are run on the entire NDAR data repository and federated data at once, so query results can be returned from multiple studies. The GUID, described previously, ensures that patients present in more than one NDAR data source are not duplicated in query results. Researchers can use their query results to create an “NDAR Study,” a collection of GUIDs that serve as references to the original research data and are able to be shared with other NDAR users.<sup>28</sup> Manuscripts and presentations using data from NDAR are expected to acknowledge the Contributing Investigator who conducted the original study, the funding organization supporting the work, and NDAR.<sup>26</sup>

## **Technical Considerations**

NDAR is the most technologically complex of the examples profiled here. It functions as both a data warehouse and a federation portal (e.g., a single point of entry to access databases located elsewhere). In addition to the data submitted by individual investigators that are stored and maintained by NDAR, researchers have access to a wider net of data through the federation linkages NDAR has (or has planned to) put in place with organizations such as the Autism Genetic Resource Exchange, the Interactive Autism Network, the NIMH Genetics Repository, the NIMH Transcriptional Atlas of Human Brain Development, and the Pediatric MRI Data Repository.<sup>28</sup>

NDAR has developed several standalone software applications to facilitate the collection and exchange of meaningful data. Separate data dictionary tools for genomics, imaging, and clinical assessment data are available to help those submitting data define their data elements in a way that is consistent with data from other studies. The GUID Tool was developed in collaboration with the Simons Foundation Autism Research Initiative to assign GUIDs to data being submitted, which involves checking existing NDAR data to ensure that duplicate patients receive the same GUID. The data validation tool is an application run on the researcher’s local machine that verifies the data conform to the required format and range values, and then

packages and imports the data into NDAR.<sup>29</sup> Because submitting data to NDAR requires effort by the investigator, NDAR provides tools to help investigators estimate the cost of preparing and submitting data. These costs can be incorporated into a project budget in cases where the investigators (e.g., those receiving NIH funding) decide to submit the data to NDAR during the study planning phase or are required to submit the data according to the terms of their funding.

NDAR grants data access to researchers through a Web portal. Researchers can query NDAR data and download the results of their query in XML or CSV format.

## **Resources**

Over the course of developing NDAR, the software engineering team used development, testing, demo, and production environments; this is a more complex system than was originally imagined, but it has proven to be helpful.<sup>22</sup> In its current operational phase, NDAR employs 3 to 4 FTEs with expertise in genomics, imaging technology, management, and training. Eight to ten FTE developers worked for 4 years on developing, testing, and refining the database. The total project funding for 2010, including operation and development, was \$2,400,000.<sup>30</sup>

## **NHLBI Biologic Specimen and Data Repository**

### **Overview**

The NHLBI Biologic Specimen and Data Repository was created by the combination of two formerly separate NHLBI programs: the NHLBI Biologic Specimen Repository and the NHLBI Data Repository. The Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) was established in September 2008 under contract to the NHLBI Division of Blood Diseases and Resources to serve as a central management entity for both biospecimens and datasets in the Repository. The mission of the Repository is to acquire, store, and distribute research biospecimens and datasets to the scientific community. As of March 2011, datasets from 82 separate studies are available for use by approved data requestors,<sup>31</sup> and 189 requests for data access have been received since September 2009.<sup>32</sup> An additional 143 requests for teaching dataset have been received and distributed in that time.

### **Governance and Oversight**

Repository employees review requests for data access for completeness and then forward them to the Division of Cardiovascular Sciences (DCVS) at NHLBI for review and final approval. Many NHLBI-funded studies are required to submit data to the Repository.<sup>33</sup>

### **Patient Privacy**

Datasets submitted to the Repository must be submitted in two parts: a Non-Commercial Purpose Dataset (consisting of participants who requested during the informed consent process that their data not be used for commercial purposes) and a Commercial Purpose Dataset (consisting of participants who consented for their data to be used in this way). Participants who did not consent to their data being shared with researchers other than their principal investigator are not included in either dataset. In addition, obvious patient identifiers must be removed and potentially identifiable variables must be altered, as detailed in the Repository's Policy for Dataset Preparation.<sup>34</sup> The datasets for distribution meet the requirements for de-identified data as defined by HIPAA. The researcher donating the data is responsible for ensuring that the data

meet these standards, but NHLBI and BioLINCC staff review the final datasets and documentation prior making them available for sharing and provide feedback if additional redaction or recoding is needed.

## **Research Ethics**

Data request applications must contain a summary of the proposed research, IRB approval, the investigator's curriculum vitae, and a signed Research Materials Distribution Agreement, which clarifies the legal responsibilities of the data recipient and the Repository.<sup>35</sup> The DCVS reviews the proposed research for significance, appropriateness, design, and "ethical and legal considerations including consistency with the terms of the informed consent and compliance with human subjects and HIPAA regulations."<sup>33</sup>

## **Data Access and Use**

Metadata describing each of the available datasets is posted on the Repository Web site for review by potential researchers (see Table 2). The Repository staff is also available to answer questions regarding a particular dataset's strengths and limitations and appropriateness for a specific research use. Once a data request is approved, the applicant is given access to the data and documentation electronically through the secure BioLINCC site.<sup>33</sup> As specified in the Research Materials Distribution Agreement, any publications or presentations resulting from Repository data are required to acknowledge the NHLBI BioLINCC as the source of that data.<sup>35</sup>

**Table 2. Metadata publicly available for NHLBI Biologic Specimen and Data Repository datasets**

<b>Category</b>	<b>Description</b>
General Study Data	Link to listing on ClinicalTrials.gov Study type (epidemiology study or clinical trial) Date that metadata entry was prepared and last updated Study period Consent (restricted, unrestricted, original researcher only, or proprietary) Commercial use restrictions NHLBI division that funded study
Abstract	Objectives, background, subjects, study design, and results/conclusions
Study Documents	May include protocol, coding manual, data dictionary

## **Technical Considerations**

Study datasets are stored by the Repository and released to approved researchers upon request. The Registry Web site has search functionality with the ability to query dataset metadata for keywords, type of sample available (specimens, dataset, or both), consent type, study type, and presence of commercial use restrictions. In addition, a Google Mini Search is available on the Web site to search for keywords present in metadata or ancillary study documents.<sup>36</sup>

Data are received at the Repository via secure FTP transfer from the NHLBI Program Office, which receives them on CD-ROMs. Data are stored on a secure internal network. Aside from standards applied to the data to ensure patient privacy and appropriate research use (described earlier), no further attempt is made to harmonize data elements or formats.<sup>34</sup>

## **Resources**

The total project funding for 2010 was \$599,990.<sup>30</sup>

## **Relevance of Existing Repositories for a Repository of Expired Registries**

These examples demonstrate that a repository of clinical data is technically, legally, operationally, and ethically feasible. The number of manuscripts published using repository data and the volume of data requests they manage indicate that there is interest within the scientific community in using archived data from completed studies. The existing repositories also suggest that a new data repository for expired registries could be structured in multiple ways, with varying levels of data quality assessment, data aggregation, and research support. Common features across all three examples include a formal data access policy, availability of at least general information about the data and the parent studies that generated them, and a governing body to provide oversight. A major difference between these repositories and the potential repository of expired registries is the scope; these repositories are limited in scope to a particular disease area, while the repository of expired registries would presumably store data from registries in a wide range of disease areas.

## **Stakeholder Perspectives on Major Issues**

The existence of other repositories of clinical study data suggests that a repository of expired registries is feasible. However, the proposed repository presents many challenges. Stakeholders attended an in-person meeting to discuss major issues related to establishing a repository of expired registries, including governance, patient privacy, research ethics, data access and use, and technical considerations. The key points related these issues are summarized here.

### **Governance**

Governance addresses both the policies and procedures for the repository and the oversight body that will ensure that the repository operates according to these policies and procedures. In considering a repository of expired registries, several governance issues arise. For example, the repository will need a clear plan explaining who can access the data and under what conditions. The repository may establish a data use committee that reviews data requests and approves or rejects them. This committee may also monitor ongoing research projects to ensure that they are conducted in accordance with the data use terms. An important aspect of the review may be assessing the potential for inadvertently re-identifying patients when linking data from multiple databases. Depending on available resources and interest, the repository may not be able to accept all donated databases, as each stored database will have intake and maintenance requirements. An oversight committee could be formed to review donated databases and determine which to accept based on pre-specified criteria (e.g., relevance to the priority conditions, quantity of data collected, cost/benefit analysis of storage costs versus likelihood of future use in analyses and publications, etc.). The governance plan will also need to address data ownership, length of time that databases are stored (e.g., indefinitely, 20 years, etc.), and any fees associated with data use. The stakeholder discussions and background research conducted for this project focused on questions related to the repository owner; purpose and composition of the governing body; review of data requests; donations and storage of registry data; and incentives for donating registry data.

### **Repository Owner**

Stakeholders could not identify clear incentives that would drive a private entity to develop such a repository independently. Many reasons were cited ranging from the perceived limited ability of a private entity to recruit data donations to the lack of a business model to leverage the data to support operations without a more defined understanding of risks and potential revenue sources. As such, the consensus view was that such a repository would need to be developed with government funding. This may change if the archive has inherent value and a clear business case emerges. The following discussions assume that the data repository will be created and managed by AHRQ or a contractor of AHRQ (as opposed to a private entity).

### **Governing Body**

Stakeholders agreed that sound, transparent governance procedures are critical for ensuring that the repository is useful for researchers and compliant with legal and ethical requirements. This conclusion is consistent with the findings from discussions with staff at other data repositories. Stakeholders recommended that the governing body include individuals from

across the spectrum of interested parties, including government, industry, providers, patients, researchers, and representatives of registries that contributed data. The purpose of the governing body would be to guide the overall direction and activities of the repository of expired registries. The governing body may be involved in review of data requests and other operational activities of the repository, or it may create committees to oversee these tasks.

## **Review of Data Requests**

The procedures for reviewing data requests are a critical part of the governance plan. Stakeholders recommended using a data access committee to review and approve requests to use repository data. The data requestor would submit information on the proposed research activity, along with IRB approval for the project. This process is similar to that used by existing repositories. Stakeholders also suggested that the governing body could set use fees for accessing the data and commented that use fees could be higher for commercial uses than for non-commercial uses, although both uses should be allowed under appropriate conditions and consistent with data use agreements.

## **Ownership and Management of Donated Data**

The repository of expired registries that stakeholders discussed could include registries that are required to submit their data (e.g., a new registry funded by AHRQ), as well as registries that donate their data voluntarily. Data ownership is an important consideration for registries that donate their data voluntarily. The repository could own the donated data; in this case, the terms and conditions of donating the data to the repository would include a transfer of ownership. In cases where copyright was claimed for the data, the copyright may also need to be licensed or transferred to the repository. Stakeholders noted that the arrangements for data ownership could encourage or discourage donation of data. For example, relinquishing all ownership rights or transferring copyright was seen as a potential disincentive to donation. Stakeholders agreed that the approach to data ownership should be guided by the available resources and the underlying goal of the repository (i.e., to serve as a last resort for data that would otherwise not be maintained or to attract donations of high-quality data and proactively facilitate new research). On balance, stakeholders felt that the ability to be flexible in data ownership, as long as the terms of use were clear and could be readily tracked with the data, would help encourage growth of the repository, but such tracking would come at a higher cost. A related issue is ensuring that the entity donating the data has the necessary rights to do so; stakeholders commented that warrants could be required in the donation agreement.

The concept of voluntary donations of data also raises questions as to whether the repository should have criteria for inclusion. For example, the NIDDK Repository focuses on studies that are considered to be of significant value based on sample size or other factors. Stakeholders suggested that, initially, the only inclusion criteria should relate to contractual issues (e.g., ability to transfer ownership legally) and ethical issues (e.g., appropriate consents or a waiver of consent for the data to be used in future research). Because the proposed repository would cover multiple disease areas and many types of studies, it would be difficult to have appropriate inclusion criteria based on size, number of sites, depth of clinical and/or biological data, length of follow up, or other factors. For example, a rare disease registry that only includes a few hundred patients could be very valuable, as could a single site registry such as the Framingham study. Stakeholders suggested that a value determination would have to be made on

a case-by-case basis. In addition, the repository could begin with an inclusive policy and develop tighter restrictions if a large number of donations become available.

A related question is how long to store registry data. This question applies to both voluntarily donated data and data that are required to be submitted to the repository. The repository could store datasets indefinitely, archive them after a set number of years, or archive them after a period of non-use. Again, the stakeholders commented that having a single policy that would apply to all registries is difficult because of the variation in disease area and registry type. Stakeholders suggested that the data be stored indefinitely, unless prevented by resource constraints, or that the repository follow existing guidelines for data retention (e.g., Medicare or commercial payer guidelines). Because of the cost associated with maintaining each dataset in a usable form, there may also need to be ongoing evaluation of the utility of the datasets being stored in the repository and different processes for those actively maintained as readily retrievable (e.g., regularly updated into current technology formats) versus those simply stored in a durable medium.

### **Incentives for Donating Registry Data**

Stakeholders discussed what factors might motivate or incentivize registry owners to donate their data to the repository and offered some suggestions. As discussed above, in some cases, the NIH repositories have set a precedent by requiring donation of data as a condition of funding; agencies such as AHRQ or other public funders might consider a similar requirement. AHRQ could also collaborate with other government agencies to encourage donations. For example, NIH Institutes without repositories could direct investigators that they fund to donate registry data to AHRQ. The FDA could also recommend donations for post-marketing commitment studies. The donation of data from post-marketing studies raises concerns related to proprietary information that may be contained in these studies; however, the FDA is currently exploring approaches to consolidating and adapting its data sets from clinical trials into research-appropriate data sets.<sup>37-38</sup> These exploratory projects may result in policies or procedures under which data from post-marketing studies may be included in the repository of expired registries.

Second, stakeholders suggested that registry owners may be motivated to donate data if liability resulting from the data is limited or transferred to the repository. Third, because many private registries that would consider donation would likely do so at the end of their existing funding, stakeholders felt that the burden on the registry donor would need to be limited in order to encourage donations from such organizations. Several stakeholders noted that industry-funded registries are unlikely to be donated because of proprietary concerns, and one stakeholder commented that some open-ended registries (e.g., quality improvement registries) may be willing to donate data from completed years, even if the registry were still ongoing. Lastly, some stakeholders suggested that a donation could result in priority access to the data; there was no clear consensus on this point, as some stakeholders commented that it did not “feel right” in the context of a public resource, but no specific barriers were identified.

### **Summary of Governance Issues**

Stakeholders assumed that the data repository would be created and managed by AHRQ or a contractor of AHRQ (as opposed to a private entity). The repository will need sound, transparent governance policies and procedures, and the governing body should include a broad group of stakeholders. A data access committee will be necessary to review and approve requests to use repository data. Various approaches to data ownership are possible, depending on the



available resources and goals of the repository. The repository should be as inclusive as possible when accepting donated databases and should store data indefinitely, resources permitting. Incentives for donating data to the repository may include mandates from funding agencies, liability considerations, or priority access to repository data.

## **Patient Privacy**

Protection of patient privacy is an important consideration for the repository of expired registries. Some registries collect direct patient identifiers, such as medical record numbers or patient names and contact information. Other registries may not collect direct identifiers, but may collect data, such as date of birth or dates of procedures, that could potentially be used to re-identify an individual patient. The HIPAA Privacy Rule protects these types of data. In agreeing to participate in a registry, patients may sign informed consents that allow the registry to collect and store personally identifiable information. However, the permission given in the informed consent may not be sufficient to allow the registry sponsor to donate that identifiable information to a data repository. Privacy concerns related to how the data in registries were gathered and stored may also exist. For example, were appropriate patient consents obtained? Were the data properly de-identified? Were sufficient privacy and security controls used (e.g., “old” registries may not have implemented proper procedures to ensure privacy and security of data)? The stakeholder discussions and background research on patient privacy focused on questions related to determining if informed consents were sufficient for storing identifiers, how the potential for re-identification through data linkage would be assessed, and the role of institutional certification. The legal issues related to patient privacy are discussed in the “Legal Issues in Creating a Repository of Expired Registries” section.

## **Storage of Patient Identifiers**

Stakeholders concluded that identifiers could be stored in the repository if the informed consent allowed for this use. Discussion focused on how to determine if the informed consent was sufficient. Suggestions included storing original consents, storing model consents for studies where multiple consents were used, or tracking major changes to data collection and consents (e.g., specimens were collected from 2004 to 2006 only). These approaches may work for new registries that consider the possibility of donating their data during the design and operational phase. For new registries, the repository could develop suggested language for informed consent documents to facilitate future donations of data. However, stakeholders commented that these options could be difficult for some existing registries, particularly those that have collected data for long periods. The possibility of contacting patients from completed studies to obtain consent was also mentioned, although this approach represents a large burden in terms of costs and resources and may be infeasible for registries that did not maintain contact information. Stakeholders also noted that reviewing consent forms represents a large administrative burden for the repository. A possible solution to this issue is discussed in the “Institutional Certification” section.

In terms of storing identifiers, stakeholders noted that additional security and technical requirements may add cost. Stakeholders also expressed concerns about data in the repository being subject to the Freedom of Information Act. The “Legal Issues in Creating a Repository of Expired Registries” section addresses this concern.

## **Potential Reidentification of Patients**

Even if identifiers are not stored in the repository, data linkage projects using repository data could inadvertently re-identify patients. Stakeholders agreed that researchers may not fully understand the risks of re-identification when considering linkage projects and discussed whether expert review of the research project methods and/or results would be necessary. While this review could reduce the risk of re-identification, it would require additional resources. Other groups that provide data for linkage, such as CMS, do not review methods or results. Stakeholders concluded that a strong data use agreement would be necessary, and, if resources allowed, additional review of the methods and final product should be considered.

## **Institutional Certification**

Stakeholders expressed concerns about whether the repository would be responsible for determining if the informed consent was sufficient for donating data to the repository. One suggestion was to rely on institutions (including the IRB or Privacy Board as applicable) to certify data prior to the data being accepted into the repository. With such “institutional certification,” the entity donating the data must provide certification that the use of the data in the repository is appropriate given the informed consents from individual patients (or the waiver of informed consent) that underlay the data.<sup>39</sup>

A model for the institutional certification process is the NIH GWAS policy, which applies to donations of biospecimens from NIH-funded genome-wide studies. The NIH policy on GWAS Data Submission Certification states:

“The NIH will accept GWAS data into the NIH GWAS data repository after receiving appropriate certification by the responsible Institutional Official(s) of the submitting institution that they approve submission to the NIH GWAS data repository. The certification should assure that:

- The data submission is consistent with all applicable laws and regulations, as well as institutional policies;
- The appropriate research uses of the data and the uses that are specifically excluded by the informed consent documents are delineated;
- The identities of research participants will not be disclosed to the NIH GWAS data repository; and
- An IRB and/or Privacy Board, as applicable, reviewed and verified that:
  - The submission of data to the NIH GWAS data repository and subsequent sharing for research purposes are consistent with the informed consent of study participants from whom the data were obtained;
  - The investigator’s plan for de-identifying datasets is consistent with the standards outlined in the policy;
  - It has considered the risks to individuals, their families, and groups or populations associated with data submitted to the NIH GWAS data repository; and
  - The genotype and phenotype data to be submitted were collected in a manner consistent with 45 C.F.R. Part 46.”<sup>27</sup>

The GWAS policy describes specific points that IRBs should consider when reviewing informed consents to make a determination on certifying data for GWAS submission. For example, the IRB should consider whether the consent forms have any restrictions, such as types

of subsequent research using the data or location of such research, and whether the study involves children or vulnerable populations.

The GWAS policy on certification could be adapted to meet the needs of the repository of expired registries. In general, certification requires the IRB to review the consent forms under which the data were collected over the duration of the study. The three possible outcomes of the certification process are (1) the data are appropriate for widespread research and can be donated to the repository; (2) the data are not appropriate for widespread research use and cannot be donated to the repository; and (3) the data may be used for research with appropriate restrictions. These restrictions could, for example, include a requirement that the data be fully de-identified or only be used in support of certain types of research (e.g., cancer research). The result of the certification process is a letter stating the IRB's findings and detailing any restrictions, which the investigator submits to the repository along with the data.

The same process of certification would apply to registries that have a waiver of informed consent. The IRB would approve an additional waiver of consent to allow the data to be used in the repository. The primary concerns in approving such a waiver would likely relate to the nature of the data. For example, sensitive data such as those related to infertility treatments may not be approved for donation, while less sensitive information such as data on controlling blood pressure in hospitals would most likely be approved. An open question relates to whether a multisite study would need to have institutional certification from each individual participating site; it may be possible for the data coordinating center (or entity holding the data) to seek certification from its IRB for the entire study based on a model consent form.<sup>39</sup>

Because of the broad scope of the GWAS policy, it is likely that many IRBs, particularly at large research institutions, are already familiar with the certification process. In order to use a certification process for the repository of expired registries, the repository would need to develop a certification policy, similar to the GWAS example, that outlines the key points that certification must address. In addition, the policy should describe the broad public health benefits of submitting data to the repository to assure IRBs that the repository is a public health resource and that the data would be used to improve public health. This is an important point for the IRB to consider, as it is charged with balancing individual rights and welfare with broad social goals.

In developing a certification policy, the repository would also need to consider whether to accept data that have restrictions on their use. Restrictions place an administrative burden on the repository, which must then ensure that data use requests comply with the conditions. The repository would need to determine if the data were still valuable and should be included, even with the restrictions. In addition, the repository would need to consider whether the burden of institutional certification is too high for researchers donating data. Registries that are designed with data donation in mind may be able to assemble the necessary information with little effort. However, researchers with older registries may find the process of certification, which requires the collection and submission of a potentially large amount of documentation, to be a barrier to donating data.

On a related note, existing registries may need to review their IRB applications and obtain re-review for those in which data retention plans (e.g., destruction dates or policies) were included, since donation to a repository of expired registries would extend the time to destruction, possibly indefinitely.

## **Summary of Patient Privacy Issues**

Patient identifiers can be stored in the repository if the informed consents supporting the original data collection allowed for this use. An institutional certification process could be used to determine if the informed consents are sufficient for storing patient identifiers and for donating the data generally. The repository would need to develop a certification policy and require researchers donating data to receive institutional certification from their institution prior to donating the data. The certification would state whether identifiers could be donated and for what purposes the donated data could be used. Regardless of whether identifiers are included, some data linkage projects using repository data may raise the possibility of inadvertent re-identification of patients. A strong data use agreement would be necessary for the repository, and, if resources allowed, additional review of the methods and final product of data linkage projects should be considered to prevent accidental re-identification of patients.

## **Research Ethics**

Related to questions about informed consent and patient privacy are questions on research ethics. The repository will need to address several research ethics issues. For example, how will the repository ensure that the data are being used for valid research purposes? How will the repository determine if any ethical issues exist with the registry data (e.g., were appropriate patient consents obtained? Were the data properly de-identified? Were sufficient privacy and security controls used?)? The stakeholder discussions and background research focused on the use of data for valid research projects and identifying ethical issues with donated data.

## **Use of Data for Valid Research Projects**

Stakeholders agreed that requests to use repository data should be accompanied by IRB approval (or an IRB determination that the research is exempt from such approval) for the proposed research project. This approach is consistent with data use procedures for the NIDDK, NHLBI, and NDAR repositories. Stakeholders also commented that additional review on the part of the repository may be necessary to determine if the use of the data represents valid scientific research. IRB approval is primarily focused on protecting human subjects, but it does not address whether the proposed research is scientifically valid or sound, particularly considering the limitations of the proposed data source. A repository steering committee or governing body could be used to review the requests and issue approvals. Should such a committee or governing body be used, stakeholders suggested that the originators of the data collection should not have a role in determining whether access to the data should be granted. Stakeholders concluded that the necessity of having governing body approval depends on the scale and scope of the repository. The repository should have a governing body that reviews data use requests for scientific validity if the repository is providing research support, such as formatting the data, assembling documentation on the data, and assisting researchers with finding appropriate datasets. These additional services add value to the data and, in turn, give the repository more rights to review the request, review papers being published using repository data, and be acknowledged in the publication. If the repository limits flexibility in how data are received and is only charged with archiving and providing data in accordance with data use restrictions, then the additional review for scientific merit may not be necessary.

## **Identifying Ethical Issues With Donated Data**

The institutional certification process discussed above is also a potential approach for identifying ethical issues with donated data. In particular, certification could identify issues with insufficient consents or privacy protections. Stakeholders also recommended that the repository confirm that data that are supposed to be de-identified are, in fact, de-identified, noting that while the concept of de-identification is well understood, the technical requirements are not. Stakeholders suggested that a list of registry publications be included with the registry data, but emphasized that the archive should not give preferential treatment to registries that have published their findings. Some registries might have important negative findings (i.e., finding of no effect of a certain treatment) that may have been difficult to publish. Data requestors should also submit at least some biographical information to ensure that they are part of a legitimate research organization and have not been subject to disciplinary action or named in lawsuits related to their research.

The ethics of extending the research purpose for which data were obtained originally must also be considered. Among some potential applications that may not have been considered when a registry was created are (1) could the data be used for commercial purposes (e.g., to understand the background rates of certain conditions that might be meaningful to guide drug development); and (2) could data be obtained, after appropriate review, to be used in support of litigation (e.g., to evaluate the prevalence of certain conditions when trying to evaluate the causality of a purported event?).

## **Summary of Research Ethics Issues**

Requests for data access should be accompanied, at minimum, by IRB approval (or an IRB determination that the research is exempt from such approval) for the proposed research project, a description of the project, and information on the researcher. Depending on the available resources, the repository may consider using a governing body to review requests for scientific validity and merit before releasing data. The institutional certification process for donated data could be used to identify ethical issues with donated data or determine restrictions on use of the data.

## **Data Access and Use**

Data access and use policies and procedures will be critical for enabling the repository to meet its goal of supporting future research projects. Some data in the repository may be able to support many types of research projects, while other data may be able to support a more limited range of studies; this will likely depend on the conditions under which the data were originally collected. Information on the data will also need to be maintained and made available to researchers to ensure that researchers understand the strengths and limitations of the data and can draw appropriate conclusions from their findings. For example, researchers will need to understand if the data were audited, how the data were collected, how the data elements were defined, and how patients and sites were selected and enrolled. These factors will help researchers to assess the data quality and the potential for selection bias. The stakeholder discussions and background research focused on conditions for data access, storage of informed consent information, and storage of data documentation.

## **Conditions for Data Access**

Stakeholders agreed that a clear, transparent data access process is necessary. The process may begin with a data use request, which describes the proposed use of the repository data. The repository would review the data use request and either approve or deny it. The level of review would depend on the type of services that the repository is providing, as noted above. For approved requests, the requestor would sign a data use agreement and return it to the repository, which in turn would release the data. Stakeholders referenced the model for accessing CMS data as an example and specifically suggested that the CMS data use agreement may be a useful model for the repository.<sup>a</sup> It was also noted that requestors should not assume any liability regarding consent and/or the results of analyses.

## **Storage of Informed Consent Information**

The informed consents (or waiver of informed consent) under which the data were originally collected are important to determining how the data can be used for future research projects. Stakeholders discussed the certification process described above as one approach to determining for which research purposes the repository data can be used. This process would allow the repository to store the IRB's findings on appropriate uses of the data, rather than storing and examining all of the informed consents. However, this model transfers some burden onto the researcher donating the data. In this discussion, stakeholders also returned to the question of including identifiable data and discussed whether registries that could not donate identifiable data could anonymize (i.e., de-identify) the data. This raised questions on the technical requirements for anonymizing data. Stakeholders questioned whether identifiers were necessary at all, but concluded that identifiers could be important for linkage studies and that the repository would need to review data use requests to ensure that identifiers were truly necessary for a project before releasing the identifiers. Whether the registries could anonymize the data prior to donation also would require review of the underlying consent and data use documents to determine if anonymization was an approved use of the data.

## **Storage of Data Documentation**

Stakeholders agreed that information on the original study and how the data were collected are critical to ensuring that researchers use the data appropriately in new research projects. This conclusion is consistent with the approaches used by the NIDDK, NHLBI, and NDAR repositories. In the NIDDK repository, the repository staff replicates tables from the primary paper based on study data to ensure that the data in the repository are the same as what were used to publish the results. Depending on the resources available, stakeholders recommended that detailed data use guides be developed for the datasets and updated periodically to incorporate lessons learned. Stakeholders also discussed whether it is reasonable to store data that were not subject to quality control measures and agreed that these datasets should be included. Information should be available in the repository as to whether the data were subject to quality measures and, if so, what measures.

---

<sup>a</sup>The feedback in this bullet point was sent to the Outcome DEcIDE Center from a stakeholder via email after the meeting.

## **Summary of Data Access and Use Issues**

A clear, transparent, documented data access process is necessary. The process should, at a minimum, include a formal data request, review of the request, and completion of a data use agreement. The repository should store information on the original study and how the data were collected to ensure that researchers can use the data and draw valid conclusions. Depending on the resources available, this information may include case report forms (CRFs) and a protocol or study plan submitted with the donated data, or the repository could develop more comprehensive data use guides. An important point to include in dataset information is what quality control measures, if any, were used during the registry.

## **Technical Considerations**

The repository will need to store donated databases in a manner that allows appropriate access in accordance with the policies and procedures of the repository. Some of the technical challenges to be faced include the implementation of an appropriate process and use of technologies and standards that allow for the likelihood of receiving a diverse range of databases that make use of different database vendor technologies, metadata, and terminologies. In addition, storing registry data requires appropriate levels of security measures based on the types of data contained in the database. The implementation of security technologies will allow adequate access and availability to the data while preventing unauthorized use. The stakeholder discussions and background research in this area focused on data storage and security, metadata, technology standards, and transfer procedures.

## **Data Storage and Security**

Stakeholders clarified that the data storage and security requirements for the repository are different from those imposed upon individual users of the data (data recipients). For example, the repository would need to comply with National Institute of Standards and Technology (NIST) standards, but individual users of the data may not need to comply with those requirements. Stakeholders suggested that individual users of the data be held to a “reasonable” standard, such as the requirements for using CMS data. The data use agreement could include language on this point, such as, “reasonable confidentiality and security measures will be used.” However, it is unlikely that the repository would be able to audit or otherwise police or enforce these requirements.

## **Metadata**

Some level of metadata will need to be available so that users of the repository can find and request access to data. Stakeholders suggested that, in a limited resources model, the only available metadata would be the registry listing in the RoPR system. This could be supplemented with additional documents, such as the CRFs and protocol, provided that the researcher submitted this information. If more resources are available, stakeholders suggested including additional information about the data, either compiled by the researcher or by the repository. These may include the inclusion/exclusion criteria, availability of identifiers, description of how and why the data were collected, bibliography, data format, percent missing for key variables, and summary analytics. Change management information may also be useful to include, particularly if the registry underwent major changes at any point.

## **Technology Standards**

Stakeholders commented that the repository could invest significant resources in formatting data to match technology standards. If resources are limited, the repository could make no effort in this area and preserve the original format of the data. If more resources are available, the repository could modify datasets for consistency in format and standards. For example, the NIDDK repository provides all datasets in SAS format. The NDAR project aggregates data across studies by mapping common data elements. Both of these options require significant resources. Alternately, the repository may make a statement of preferred standards, but accept data in any format.

## **Transfer Procedures**

When a data request is approved, the data will need to be provided to the requestor. The data could be transferred to the requestor (e.g., mailed on a CD-ROM) or maintained through a data enclave. Stakeholders commented that the transfer procedures are less important for de-identified data. For example, these data may be sent on a CD-ROM with a requirement that the requestor destroy the data when the project is complete. Access to identifiable data is more complex. In these cases, the repository may consider transferring the data through other means, such as a secure FTP transfer. To prevent unauthorized use of the data, stakeholders suggested that the ramifications of breaking the data access rules should be in the agreement (e.g., no access to repository in the future, no future grants from AHRQ). A data enclave, such as the model used in NDAR, would provide greater control but would require more significant resources, including developing and maintaining the enclave and preparing data for inclusion

## **Summary of Technical Considerations**

While the repository itself would be held to NIST standards, individual researchers using the data could be held to a “reasonable” standard for data storage and security. Some level of metadata will need to be available so that users of the repository can find and request access to data. At a basic level, metadata could include only the registry listing in the RoPR system. Available resources would also drive technology standards and transfer procedures. If sufficient resources were available, the data could be converted to a standard format; otherwise, the data could be stored in their original formats. Data transfers could be accomplished through shipping durable media or using more sophisticated, secure electronic interface.

## **Summary of Stakeholder Perspectives**

Stakeholders attending the in-person meeting concluded that the creation of a repository of expired registries is feasible from an ethical, operational, and technical perspective. There is interest in the research community in developing a repository of expired registries to facilitate future research, but the value of the repository would depend on the data that were donated.



## **Legal Issues in Creating a Repository of Expired Registries<sup>b</sup>**

The question of whether a data repository consisting of expired registries<sup>c</sup> can lawfully be created and used raises numerous legal questions under both state and Federal law. At the same time, numerous examples of such repositories exist, clarifying the legality of such arrangements.

At its heart, the law is about the rules of conduct that govern relationships. Depending on the relationship to be created, applicable laws can vary. In general, the creation of a data repository will involve relationships among a number of parties: the developer(s)/owner(s) of the repository; the owners of the information whose inclusion is sought; and parties with an ongoing legal interest in how the information is collected, stored, and made available to third parties. It is these relationships that must be accounted for in repository creation and maintenance.

In cases in which the parties consist of a privately supported repository linked to registries created with private funding, ownership of the repository and the information sources may be private, but at the same time other parties (i.e., the individuals whose information is being held and used) will have a legal interest in the enterprise. Thus, the contractual terms that spell out ownership, access, and use rights also will need to account for the legal interests of third parties under state and Federal law, in particular laws that apply to health information privacy and confidentiality. A similar result would pertain when a repository is created and managed under state law; that is, state law would govern the establishment and operation of the repository and would define the interests of participants. Similarly, the state's repository enterprise would have to assure that interests and rights created under Federal laws also are protected.

This analysis focuses on the legal issues that pertain to situations in which the Federal government authorizes and supports the creation of a repository, whether maintained directly by the Federal government or through a private contractor, and where the repository holds the content of registries that in turn have been created as part of one or more federally supported project awards. This analysis uses as an example an AHRQ-sponsored repository (termed here a federally sponsored repository) that holds the contents of registries created through AHRQ-sponsored project awards, although the Federal legal principles that come into play in such an example would arise regardless of which Federal granting agency (e.g., AHRQ, Centers for Disease Control and Prevention [CDC], NIH, CMS) funded the registries that in turn feed a repository.

In creating a federally sponsored repository, two distinct situations can arise. In the first situation, the repository, along with the registry data it holds, has been created with the express assumption (and agreement on the part of individual registry holders) that the contents of individual registries would be linked in the future to the repository. In the second situation, the repository is created after the fact, that is, after numerous individual federally supported registries have been created but not in anticipation of ultimate linkage to a repository. Put another way, in one case the repository is prospective and anticipatory, while in another, the

---

<sup>b</sup>This section was developed by Sara Rosenbaum, Hirsh Professor and Chair, Department of Health Policy, George Washington University School of Public Health and Health Services; Jane Hyatt Thorpe, Associate Research Professor of Health Policy, and Robert Platt, J.D. cand., 2012.

<sup>c</sup>The specific question asked in this project focuses on expired registries. The laws that we identify in our analysis also would be relevant to linking a repository to ongoing registries.

repository is created after registries authorized by Federal project awards have been established and are operating. In truth, of course, both situations may arise, since while repositories may put anticipatory agreements into place going forward, they inevitably will desire to link retrospectively to registries that pre-dated their existence. Thus, any repository project will need legal capabilities to look forward and backward in relation to current and future registries included or to be included therein.

Beyond the timing issues associated with the relationship between a repository and individual registries, other legal issues will arise. For example, legal considerations might change depending on whether a repository seeks to hold identifiable, as opposed to non-identifiable, patient information. Similarly, legal issues will depend on the types of uses that the central repository seeks to permit. Who will have access to repository information and for what purposes? Will access be to identifiable information or non-identifiable data? Who will have the power to set the rules for data access and use, and how will the repository be governed?

The analysis that follows focuses on the following areas of law:

- Federal laws related to agency access to and oversight of federally supported projects and the information produced under such projects
- HIPAA Privacy, Security, and Breach Notification Rules
- The Common Rule
- “Part 2” regulations related to alcohol and substance abuse treatment information
- The Privacy Act of 1974
- Federal Information Security Management Act (FISMA)
- Federal laws related to information collection (e.g., the Paperwork Reduction Act)
- Freedom of Information Act (FOIA)
- State confidentiality laws

## **Federal Laws**

### **U.S. Department of Health and Human Services (HHS) Regulations Governing Uniform Administration of Federal Awards**

Federal regulations establish certain “Uniform Administrative Requirements for Awards and Subawards to Institutions of Higher Education, Hospitals, and Other Nonprofit Organizations and Commercial Organizations.”<sup>40</sup> These regulations, specifically 45 C.F.R. 74.36(c), state that:

The Federal Government has the right to: (1) Obtain, reproduce, publish or otherwise use the data first produced under an award; and (2) Authorize others to receive, reproduce, publish, or otherwise use such data for Federal purposes.

Where Federal funding is awarded with the express assumption that the award will support, among other activities, the creation of a patient registry, the Federal government presumably would retain a right of access to and use of the data and would further have the power to allow others to use the data. However, where the creation of a registry is undertaken as an activity for the recipient’s own use and not one specified or funded under the award, the Federal interest in the data may be less clear, even though the entity is a federally funded recipient. Thus, to the extent that the government desires to have access to and use of patient registry data, this regulation suggests that the relationship between registries created and the award should be made an explicit aspect of the terms and conditions that attach to the award. Put

another way, if creation and maintenance of a registry is a condition of the grant award, then the Federal interest is preserved.

This rule does not obviate the need to assure compliance with other applicable laws. However, it does clarify that the Federal government has a legal interest in registry data (for its own use and that of others) when created as part of federally supported projects whose terms express the Federal interest in advance and make authorized data sharing with the Federal government and legally authorized third parties a condition of the award.

## **HIPAA**

### **The HIPAA Privacy Rule**

The Privacy Rule applies to covered health care entities, which include health plans, health care clearinghouses, and health care providers who conduct certain electronic health care transactions.<sup>41</sup> The Federal government, including AHRQ, is not a HIPAA-covered entity and therefore HIPAA does not apply unless HIPAA requirements are included in the terms of an agreement. However, a private contractor operating under a project award related to a data repository could be covered (or could be the business associate of another covered entity) to the extent that it is authorized to obtain and store data submitted by a Federal award recipient. In this situation the repository contractor might be a “health care clearinghouse” and thus, a covered entity, since its function is to “process[] or facilitate[] the processing of health information received from another entity in a nonstandard format or containing nonstandard data content into standard data elements.”<sup>41</sup>

A private contractor also might be considered a business associate of a covered entity, because its task is to perform or assist in the performance of “a function or activity involving the use or disclosure of individually identifiable health information.”<sup>42</sup> This would be especially true in cases in which, as part of an award, the Federal government has both specified the creation of a registry and has conditioned the award, in advance, on the willingness of the recipient (assuming the recipient is a covered entity) to make its registry data available to a central repository contractor. Even if the recipient of the award or the repository contractor are not covered entities or business associates for purposes of HIPAA, it is likely that as downstream recipients of data (e.g., protected individually identifiable health information) from covered entities (providers or health plans), they will be bound contractually to comply with the HIPAA Privacy Rule requirements.

The Privacy Rule’s purpose is to protect individually identifiable health information (also referred to as protected health information or PHI) used by covered entities and their business associates. For this reason, the Rule requires explicit individual patient authorization or an opportunity for the individual to object to the use or disclosure of PHI unless release is otherwise permitted by the Privacy Rule.<sup>43</sup> Specifically the Privacy Rule requires individual patient authorization, among other cases, in the following instances: (1) for any use or disclosure of psychotherapy (with exception for treatment, payment or health care operations)<sup>44</sup> and (2) for marketing purposes (with some exceptions for face-to-face communication or gift promotions of nominal value).<sup>45</sup> Moreover, the HITECH Act adds a new provision to the Privacy Rule that prohibits, with several exceptions,<sup>46</sup> covered entities from selling protected health information without authorization. A valid authorization must describe how the information will be used or disclosed, the purpose of use or disclosure, the expiration date for the authorization, the individual’s right to revoke consent, and the individual’s actual consent to use the information.<sup>47</sup>

Unless the activity falls within one of the exceptions to this basic requirement, the Privacy Rule would apply to any relationships involving the exchange of PHI, such as identifiable registry data, between covered entities or a covered entity and its business associate, regardless of whether the disclosure is in furtherance a Federal award that explicitly provides for the provision of data to a repository, a subsequent activity not contemplated under the original Federal award, or an activity between two wholly private actors.

Furthermore, because registries may exist over long periods, it is important to note that the Privacy Rule protects a deceased patient's PHI in the same manner and to the same extent as it would a living individual's PHI.<sup>48</sup> Currently, this protection is afforded to the deceased's PHI for as long as the covered entity maintains the information with certain exceptions.<sup>d</sup> However, the Federal government has proposed significant changes to this protection in a July 2010 Notice of Proposed Rulemaking. Under the proposed rule, the definition of PHI would be amended to exclude information of persons who have been deceased for more than 50 years.<sup>49</sup> Thus, the Privacy Rule would not protect the PHI of these deceased persons. It is unclear when and whether the Federal government will finalize this proposed change.

The Privacy Rule does permit a covered entity to disclose protected health information without individual consent for certain specified purposes, the relevant of which are noted here:

- Where the disclosure is for treatment, payment, and health care operations;<sup>e</sup>
- Where the disclosure is "incident to" an otherwise permitted or required disclosure and the disclosure is the minimum necessary to accomplish the disclosure's intended purpose;<sup>f</sup>
- Where an individual has had an opportunity to agree to, or object to, disclosure;<sup>50</sup>
- Where a disclosure is required by law (information disclosed must be limited to what is required);<sup>g</sup>
- Where the disclosure is for research (regardless of the source of funding for research) provided that an IRB or Privacy Board has approved an alteration or waiver of the basic authorization requirement, and provided further, that the purpose of the disclosure is research preparation purposes and, importantly, that none of the information will "be removed from the covered entity;"<sup>h</sup> Where disclosure for research involves a decedent's

---

<sup>d</sup>For instance, a covered entity may disclose PHI to a coroner or medical examiner for purposes of identification or determining cause of death as well as for activities related to organ procurement, and a deceased's personal representative. 45 CFR § 164.512(g)(1), 45 CFR § 164.512(h), 45 CFR § 164.502(g)(1).

<sup>e</sup>45 CFR 164.502 (a)(1)(ii). Whether furnishing data to a repository is considered part of the covered entity's health care operations arises as an issue. For example, if the terms of an award to engage in quality improvement activities related to patients with diabetes require that the registry data be supplied to a central depository, would such terms make the disclosure part of the recipient's ongoing health care operations? Where disclosure is considered part of health care operations, the "minimum necessary" standard normally would apply, since the standard is lifted only in cases in which disclosure is for purposes of treatment. On the other hand, if disclosure is considered part of compliance with the terms of a Federal award, then it may be possible that the minimum necessary rule would not apply because the disclosure is required by law. 45 CFR 164.502 (b)(2) and §164(b)(12)(a)

<sup>f</sup>45 CFR §§164.502 (a)(1)(iii) and (b). A question arises as to whether the transmission of data to a repository as part of the activities contemplated under an award would be considered incident to the main activities of the award, for example, to investigate and improve the quality of health care in order to reduce hospital acquired infections.

<sup>g</sup>45 CFR 164.512 (a)(1). A separate question might be whether a disclosure requirement contained in a Federal grant award would be considered a disclosure required by law under HIPAA.

<sup>h</sup>45 CFR 164.512 (i)(1)(i). The stipulation in the rule that none of the data will leave the covered entity would appear to preclude reliance on this rule to justify access to PHI without individual authorization, since the Federal repository presumably would be considered a separate covered entity. If the repository is considered the business

PHI, a covered entity must obtain assurances that the use or disclosure is solely for research on the PHI of decedents and be provided documentation of the death of the individual about whom information is sought and that the PHI is necessary for the research purposes.<sup>51</sup> Disclosures of decedent's PHI for research purposes do not require prior IRB or privacy board approval, or authorization by a personal representative.

- Where the data release is restricted to LDS that do contain certain specific patient identifiers, release is possible as long as the individuals who will use the data enter into a data use agreement (DUA);<sup>52</sup>
- Where the data are de-identified, disclosure is possible, because the Privacy Rule does not apply to such information.<sup>53</sup>

Furthermore, the Privacy Rule allows a covered entity to share protected health information with business associates that perform certain functions or services for the covered entity. As noted, whether a Federal repository to which disclosure is a required as part of a Federal award is a "business associate" is not clear, but assuming that the terms of the award create such an arrangement, then the Privacy Rule, as modified by the HITECH Act, would require that the business associate relationship, and the conduct of the business associate, conform to certain standards.<sup>54</sup>

### **The HIPAA Security Rule**

Similar to the Privacy Rule, for the type of situation assumed under this analysis in which AHRQ is the repository developer, the HIPAA Security Rule will not apply to AHRQ because AHRQ is not a covered entity nor likely to be a business associate of a covered entity. However, AHRQ will be receiving data from entities that are covered entities (e.g., patient registry data from federally supported projects or Part A and B data from CMS or CMS registries). AHRQ also may receive data from entities that are not covered entities but will be using data received from covered entities to conduct research (e.g., researchers using CMS, health plan, or provider data). If AHRQ utilizes a private contractor, however, that entity may be a covered entity or business associate of a Federal award recipient.

In the event the data shared either directly with AHRQ or through a federally funded researcher are patient-identifiable and transmitted electronically by or from a covered entity, the Security Rule applies to the covered entities and their business associates. In this situation, it is likely that the covered entity will include Security Rule provisions in either a business associate agreement or data use agreement (in the case of a researcher) governing the release of the data. These data sources (e.g., CMS, health plans, and providers) and users (e.g., researchers) will seek assurances that AHRQ will maintain the security of any patient-identifiable data maintained in the registry. In developing policies and procedures and agreements with data sources of patient-identifiable data in the registry, AHRQ will need to consider implementing HIPAA security protections for consistency and to facilitate assurances of security for data transfer and maintenance despite AHRQ's noncovered entity status. Presumably, furthermore, AHRQ's contractor would need to do the same.

---

associate of a covered entity, it may be possible to transfer individually identifiable data for research purposes. But unless the repository can be considered part of the current entity, it would appear that the regulation bars use of the research exception for PHI. Different standards apply if the registry contains decedent data. See also Modifications to the HIPAA Privacy, Security, and Enforcement Rules, 75 Fed. Reg. at 40,919-20 (to be codified at 45 CFR § 164.504)

The HIPAA Security Rule protects a subset of information covered by the Privacy Rule.<sup>55</sup> It protects all individually identifiable health information a covered entity or business associate creates, receives, maintains or transmits in electronic form, and classifies this information as “electronic protected health information” (e-PHI). The Security Rule does not cover PHI that is transmitted or stored on paper or provided orally. Thus, were only non-identifiable information transmitted to a repository contractor, the Security Rule would not apply.

Covered entities and business associates of covered entities are required by the Security Rule to maintain reasonable and appropriate administrative, physical, technical, and organizational safeguards for protecting e-PHI.<sup>56</sup> Specifically, entities must: (1) ensure the confidentiality, integrity, and availability of all e-PHI that the covered entity creates, receives, maintains, or transmits; (2) protect against any reasonably anticipated threats or hazards to the security or integrity of such information; (3) protect against any reasonably anticipated uses or disclosures; and, (4) ensure workforce compliance.

The Security Rule provides entities considerable flexibility in meeting such requirements. Entities may use any security measure that allows them to reasonably and appropriately implement the Rule’s standards and implementation specifications. However, when deciding which security measures to use, an entity must always take into account: its size, complexity, and capabilities, including technical infrastructure, hardware, and software capabilities; the costs of security measures; and the probability and criticality of potential risks to e-PHI.<sup>56</sup> In addition to guidance from HHS regarding HIPAA, covered entities should see the guidance documents issued by NIST to assist entities in properly securing their electronic data in compliance with HIPAA.<sup>57</sup>

### **HIPAA Breach Notification Rule**

As noted above, AHRQ is not a HIPAA covered entity and is unlikely to be a business associate under the assumed situation (AHRQ as repository developer). At the same time, AHRQ may contract with a private entity to develop and manage a data repository. Were AHRQ to directly administer the repository, the new HIPAA breach notification requirements would not apply, but they would in the case of a private contractor if that contractor is considered a covered entity or business associate of the data sources (e.g., Federal award recipient). Furthermore, were AHRQ to receive patient-identifiable information from covered entities or their business associates, or entities operating under data use agreements (e.g., researchers), these entities would be subject to breach notification requirements either directly (as covered entities and business associates) or indirectly (e.g., researchers operating under data use agreements). In developing policies and procedures as well as agreements with data sources that reach patient-identifiable data to be included in a repository, AHRQ therefore will need to consider breach notification provisions, added by HITECH to the HIPAA Privacy Rule. These provisions<sup>58</sup> require covered entities and business associates to notify affected individuals about breaches of “unsecured PHI” (i.e. PHI that has not been rendered “unusable, unreadable, or indecipherable to unauthorized individuals”)<sup>59</sup> that could compromise PHI privacy or security. Encryption and destruction offer safe harbors to the breach notification duty.<sup>60</sup> Where data remain unsecured, the requirements obligate covered entities and business associates to provide notification following a breach that “compromises the security or privacy of the protected health information” such that the use or disclosure “poses a significant risk of financial, reputational, or other harm to the individual.”<sup>61</sup> Whether such a breach of registry data would trigger such notification might be

thought of as a secondary matter; the primary consideration on AHRQ's part presumably would be to assure that data are stored in an encrypted fashion to avoid exposure.<sup>i</sup>

## **The Common Rule**

The Common Rule<sup>62</sup> is applicable to any type of research-related activity involving human subjects that is conducted or supported by HHS, including AHRQ. Certain types of human subject research are exempt from the requirements of the Common Rule, including research involving the collection or study of existing data if the data are publicly available or if information is "recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects."<sup>63</sup> It is conceivable therefore that whereas individual authorization may be required under the Privacy Rule before registry data can be sent to a repository, research ultimately carried out using registry data might be exempt from the Common Rule's consent requirements if the subjects cannot be directly or indirectly identified. Assuming, however, that identification will be possible, the Common Rule presumably would apply to research involving registry data stored in a repository.

It is possible of course, that repository data will not be used for research. But the Common Rule defines research broadly as any "systematic investigation, including research development, test and evaluation designed to develop or contribute to generalizable knowledge."<sup>64</sup> Thus, our working assumption is that in general, repository data maintained by AHRQ would be used for activities that fall within the Common Rule's definition of research; indeed, even were a repository of registries created for purposes other than research, such as health care operations related to quality improvement or performance improvement and efficiency by health care systems, its use for research would trigger the Common Rule to which AHRQ, of course, is subject.

All institutions that utilize an AHRQ repository to obtain data for research purposes would be required to bind themselves in writing to compliance with the rule under a Federalwide Assurance (FWA).<sup>65</sup> As such, research conducted using the repository data would be subject to prospective review and approval by an IRB.<sup>66</sup> IRB approval in turn is conditioned on compliance with the Rule's informed consent requirements.<sup>j</sup>

---

<sup>i</sup>The Privacy Rule imposes a three-step inquiry in relation to security breaches: first, whether use or disclosure would violate the HIPAA Privacy Rule; second, whether the violation or breach would compromise the security or privacy of the PHI; and third, whether exceptions to the breach definition apply. 45 C.F.R. §164.402 Assuming that repository data are personally identifiable, either because authorization has been obtained or because the repository meets a Privacy Rule exception, the failure to store data in an encrypted form would necessitate a risk assessment of the potential implications of a breach. One would assume that AHRQ would proceed with such a repository only if data were to be stored in a secure format.

<sup>j</sup>45 CFR 46.111(4)-(5). An adequate informed consent must contain: (1) A statement that the study involves research, an explanation of the purposes of the research and the expected duration of the subject's participation, a description of the procedures to be followed, and identification of any procedures which are experimental; (2) A description of any reasonably foreseeable risks or discomforts to the subject; (3) A description of any benefits to the subject or to others which may reasonably be expected from the research; (4) A disclosure of appropriate alternative procedures or courses of treatment, if any, that might be advantageous to the subject; (5) A statement describing the extent, if any, to which confidentiality of records identifying the subject will be maintained; (6) For research involving more than minimal risk, an explanation as to whether any compensation and an explanation as to whether any medical treatments are available if injury occurs and, if so, what they consist of, or where further information may be obtained; (7) An explanation of whom to contact for answers to pertinent questions about the research and research subjects' rights, and whom to contact in the event of a research-related injury to the subject; and, (8) A statement that participation is voluntary, refusal to participate will involve no penalty or loss of benefits to which the

The fact that research is governed by the Common Rule is not dispositive in and of itself of the issue of informed consent. The Rule provides for waiver of written informed consent where an IRB finds that the consent document itself represents the only record linking the subject and the research and that the principal risk is the potential harm resulting from a breach of confidentiality. The Rule also permits an IRB to waive informed consent if it finds that the research presents no more than minimal risk of harm to subjects and involves no procedures for which written consent is normally required outside of the research context.<sup>67</sup>

Whether the research-related use of registry data (perhaps created for research and perhaps not) located in a larger repository involves more than minimal risk obviously is a fact-sensitive matter. No blanket statement about waiver is possible; indeed, the most sensible working assumption is that an AHRQ-administered repository (either directly or via a private contractor) would be subject to the Common Rule, and therefore, that institutional compliance with the Rule (including case-by-case IRB determinations) would be required as a condition of using the repository for research.

### **Special Considerations Related to the Disclosure of Alcohol or Substance Abuse Registry Data**

Where individual patient registry data to be disclosed to a repository involves data related to substance abuse, special Federal standards apply.<sup>68</sup> Specifically, 42 C.F.R. Part 2 bars the disclosure of individually identifiable data related to treatment for alcohol or substance abuse, even in circumstances in which disclosure would be permitted under HIPAA (i.e., where disclosure is part of health care operations). The Part 2 standards date to the early 1970s and have remained unrevised under HIPAA. The rules, which are based in Federal statute, reflect Congressional concern over stigma; indeed, the regulations are such a fixture in Federal privacy law that they are referred to simply as “Part 2.”

With certain conditions and exceptions, Part 2 bars disclosure and use of drug and alcohol use records maintained in connection with the performance of any federally assisted alcohol and drug use program. Under Part 2, “disclosure” means “a communication of patient identifying information, the affirmative verification of another person’s communication of patient identifying information, or the communication of any information from the record of a patient who has been identified.”<sup>69</sup> Patient-identifying information is defined broadly and encompasses any “information by which the identity of a patient can be determined with reasonable accuracy and speed either directly or by reference to other publicly available information.”<sup>69</sup> Part 2 broadly defines patients and programs in order to attain maximum reach: it applies to programs and units of general medical facilities that *hold themselves out as providing*, and actually *provide* alcohol or drug abuse diagnosis, treatment, or referral for treatment as well as medical personnel or other staff in a general medical care facility whose *primary function* is the provision of alcohol or drug abuse diagnosis, treatment, or referral for treatment and who are *identified* as such providers. The Part 2 definition of “patient” encompasses “any individual who has *applied for or been given a diagnosis or treatment* for alcohol or drug abuse *at a federally assisted program*.” Unlike HIPAA, Part 2 contains no general “operational” exemption from the individual authorization requirement. A working assumption for a repository of registries would be that the repository would require the express written consent of patients before it could accept

---

subject is otherwise entitled, and the subject may discontinue participation at any time without penalty or loss of benefits to which the subject is otherwise entitled. 45 C.F.R. §46.116(a)



such data, even were the disclosure of other types of individually identifiable data considered as falling within HIPAA's Federal law or operational exemptions.

## **Privacy Act of 1974**

The Privacy Act protects information about individuals, such as patients and practitioners, held by or collected by the Federal government that can be retrieved by personal identifiers such as name, social security number, or other identifying numbers or symbols. The Privacy Act authorizes a Federal agency to release individually identifiable information to identified persons or to their designees with written consent or pursuant to one of twelve exemptions for disclosure.<sup>70</sup> The broadest of the twelve exemptions, the "Routine Use" disclosure, authorizes Federal agencies to release individually identifiable information pursuant to a System of Records (SOR) and Routine Uses.<sup>70</sup> An SOR is a group of any records under the control of a Federal agency from which information is retrieved by the name of the individual or by a particular identifier. When a Federal agency establishes or substantially revises an SOR that contains individually identifiable information, the Privacy Act requires that the agency publish a notice of a system of records (SORN or "notice") in the Federal Register and submit a report about the new or amended system to the Office of Management and Budget (OMB) and to Congress for approval. Federal regulations apply when an agency creates a new or altered SOR.

The Privacy Act also governs how a Federal agency collects, maintains, and uses individually identifiable information. For example, Federal agencies must keep an accurate accounting of all records disclosed without an individual's written consent except for disclosures made to agency employees and disclosures required by the Privacy Act.<sup>71</sup> In addition, agencies may collect only that information needed to accomplish the purpose for its collection.<sup>72</sup> They must also collect information directly from the individual whenever practicable and maintain all records containing such information accurately and completely.<sup>73</sup> Finally, when an agency creates an SOR for collected data, it must publish a notice in the Federal Register that explains its system of records and the acceptable routine uses as described above. Penalties for violations of the Privacy Act include both civil and criminal penalties, including an individual right of action in cases in which the agency violates certain provisions of the Act, including requirements related to maintenance of data, requests for amendment, or other agency conduct that adversely affects the individual.<sup>74</sup>

In the event that an AHRQ repository were to include and/or release individually identifiable information (patient or practitioner or other individual), then the requirements of the Privacy Act of 1974 would be triggered. AHRQ should consider developing a set of routine uses that would govern how AHRQ may release individually identifiable information (along with other applicable laws and regulations) to external researchers and other interested parties.

## **Federal Information Security Management Act of 2002 (FISMA)**

FISMA<sup>75</sup> requires that Federal information systems and information have security protections commensurate with the risk and magnitude of the harm resulting from unauthorized access, use, disclosure, disruption, modification or destruction of that information. FISMA guidance indicates that FISMA applies broadly to the Federal government as well as organizations that possess Federal information, but only if they are using it on behalf of a Federal agency.<sup>76</sup> "On behalf of" has been interpreted to mean that the entity is acting as a "direct extension of the Federal government" and "to accomplish a Federal government function."<sup>77</sup> As

such, it applies to systems that support the operations and assets of the agency, including those provided or managed by another agency, contractor or other source.

FISMA includes requirements for security documentation that encompass matters such as systems security plans, risk assessments and contingency plans.<sup>78</sup> Systems also must be independently tested and formally certified by the business owners as being in compliance with security requirements and reviewed at least every three years.<sup>76</sup> Security controls for these systems, including contingency plans, must be tested annually.<sup>79</sup>

FISMA clearly applies to government systems and contractors. When government contractors use Federal data centers for their analytic activities, FISMA governs their activities. Importantly, FISMA does not apply to entities receiving government data for work that is not performed on behalf of the government, such as external research requests. However, if a researcher requests individually identifiable data from a government agency, it is likely that the agency will require that the entity meet a minimum level of security requirements based on FISMA standards.<sup>80</sup>

To the extent that an AHRQ-administered repository contains individually identifiable information or even limited datasets, AHRQ and/or its private contractor will need to conduct a risk assessment and develop policies and procedures to protect the security of the data held in the registry. To the extent that AHRQ releases identifiable data from the repository, the receiving entity also must comply with FISMA. FISMA does not apply to the release of identifiable data to entities that are not working for or on behalf of the releaser (e.g., an independent researcher).

## **Paperwork Reduction Act (PRA)**

The PRA applies to the collection of information from the public by the Federal government. It is designed to ensure the maximum utility of any information collected while minimizing the burden imposed on the public.<sup>81</sup> Specifically, the PRA applies to collections of information involving ten or more persons, other than employees of the United States, or to answers to questions posed to employees of the United States that are to be used for general statistical purposes.<sup>82</sup> More specifically, it applies to collections of information using identical questions posed to ten or more person in the course of any 12-month period or to identical reporting requirements imposed on ten or more persons.<sup>83</sup> OMB regulations have defined “information” as “any statement or estimate of fact or opinion, regardless of form or format, whether in numerical, graphic, or narrative form, and whether oral or maintained on paper, electronic or other media.”<sup>84</sup> The PRA applies to both mandatory and voluntary collections of information, including collections required to obtain a Federal benefit such as a job or grant.

Despite its broad reach, the PRA exempts certain types of information. These include, for example, affidavits, receipts, consents, tests of aptitude, and facts that are submitted in response to general solicitations, such as for rulemaking, or alternatively, obtained at public hearings or during clinical trials.<sup>85</sup> The only stipulation is that these collections entail no burden other than that necessary to identify respondent, the date, the respondent’s address and the nature of the instrument.<sup>86</sup> Importantly, in addition to the exceptions to “information” listed above, OMB has authority to identify other “like items” that are not considered to be information subject to the PRA. For example, OMB does not consider data collected in the creation of user accounts or profiles for agency Web sites to be information.<sup>87</sup> However, if the agency Web site collects any additional information during registration, such as age, sex, or ethnicity, the collection of data will be beyond what is necessary for self-identification during account registration and thus will be subject to the PRA.<sup>87</sup> Guidance also notes that if the agency uses online accounts to collect

any information for programmatic purposes, for example to determine eligibility for a program, the PRA will apply.<sup>87</sup> It is unlikely based on this analysis that these exemptions would reach patient registry data.

In general, if AHRQ collects information from the public (e.g., obtains registry-relevant information from non-Federal sources), then the Agency will need to comply with the PRA.<sup>88</sup> For example, when AHRQ collected information from interested communities to participate in the Chartered Value Exchange (CVE) program, AHRQ received a PRA number that was included on the CVE application. If an AHRQ-administered repository receives registry-relevant information from other Federal agencies (e.g., CMS) or federally funded researchers, it is unlikely that AHRQ will trigger PRA requirements.

## **Freedom of Information Act (FOIA)**

FOIA was enacted by Congress in 1966 and later amended in 1996 to cover electronic records.<sup>89</sup> The goal of FOIA is to ensure that there is an informed citizenry capable of holding the government accountable for its actions.<sup>90</sup> Accordingly, FOIA provides that any person has a right, enforceable in court, to obtain access to Federal agency records in a timely fashion. FOIA applies to all agencies within the Executive Branch of the Federal government, including independent regulatory agencies and some components within the Executive Office of the President.<sup>91</sup> Therefore, FOIA effectively establishes a statutory right of public access to Executive Branch records and information in the Federal government upon an individual's request.

There are, however, limitations on the access to Executive Branch information. FOIA provides nine exceptions to its otherwise extensive disclosure requirements.<sup>92</sup> The nine exemptions articulated in FOIA are: (1) information containing classified materials of national defense or foreign policy;<sup>93</sup> (2) information related solely to the internal personnel rules and practices of an agency;<sup>94</sup> (3) information specifically exempted by other Federal statutes;<sup>95</sup> (4) trade secrets and commercial or financial information obtained from a person and privileged or confidential;<sup>96</sup> (5) inter-agency or intra-agency memorandums or letters which would not be available by law to a party other than an agency in litigation with the agency;<sup>97</sup> (6) personnel and medical and similar files the disclosure of which would constitute a clearly unwarranted invasion of personal privacy;<sup>98</sup> (7) investigatory records compiled for law enforcement purposes;<sup>99</sup> (8) information contained in or related to examination, operating, or condition reports prepared by, on behalf of, or for the use of an agency responsible for the regulation or supervision of financial institutions;<sup>100</sup> and (9) geological and geophysical information and data.<sup>101</sup>

The exemption most relevant to the creation of registries and an AHRQ-maintained repository is exemption 6, which states that information about individuals in "personnel and medical files and similar files" can be withheld from disclosure by Federal agencies when the disclosure of such information "would constitute a clearly unwarranted invasion of privacy."<sup>98</sup> In interpreting this exemption, the Supreme Court held that Congress intended the three above-mentioned categories (personnel, medical, and similar files) to be broadly interpreted and to protect information that applies to a particular individual.<sup>102</sup> Once it has been determined that the information fits into one of those categories, the Court then evaluates whether the disclosure would amount to an unwarranted invasion of privacy. If the Court finds that there is a protectable privacy interest, it then balances the public's right to disclosure of information against the individual's right to privacy.

The Court's decision in *United States Dep't of Justice v. Reporters Comm. for Freedom of the Press*<sup>103</sup> elucidates how privacy interests are determined and weighed against public interests under FOIA exemption 6. The Court held that the inquiry of the request must turn on the nature of the requested document and its relationship to the underlying purpose of FOIA—to promote transparency of agency action to allow for public scrutiny—not on the particular purpose for which the document is being requested.<sup>104</sup> With this interpretation, the Court narrowed the scope of the public interest prong to information that will help identify the agency's performance of its statutory duties.<sup>105</sup>

It is important to note that exemption 6 cannot be invoked to withhold information from a requester pertaining only to himself.<sup>106</sup> With regard to the strength of exemption 6, the Court of Appeals for the District of Columbia Circuit has declared that “under Exemption 6, the presumption in favor of disclosure is as strong as can be found anywhere in the Act.”<sup>107</sup> However, exemption 6 has been commonly used to withhold individually identifiable information requested under FOIA.<sup>108</sup>

Clearly, the scope of FOIA could pertain to information held in an AHRQ repository. However, there are two relevant points that limit the application of FOIA to an AHRQ repository. First, in the event the AHRQ repository holds individually identifiable information, exemption 6 could be applied to protect individual patient and practitioner data. Second, release of repository-held registry data (identified or de-identified) is unlikely to shed light on AHRQ's operation and the performance of its statutory duties. Therefore, given the Supreme Court's interpretation of FOIA, any FOIA request is likely to be interpreted as beyond the underlying purpose of FOIA.

## **State Law**

It is possible that the laws of a state in which the entity that either discloses data to a repository or that uses repository data may apply separate and more stringent standards than those found in Federal law.<sup>109</sup> For example, certain states may completely bar the disclosure of individually identifiable data related to mental illness or HIV/AIDS without express individual consent, regardless of Federal exemptions that otherwise might apply. In such a situation, the HIPAA Privacy Rule would not preempt such state laws.<sup>110</sup>

Generally speaking, privacy refers to an individual's desire to limit or protect access to personal information. As such, privacy involves an individual's interest in avoiding disclosure of personal matters.<sup>111</sup> Accordingly, in the context of the HIPAA privacy refers to an individual's interest in limiting access to his or her own health information. Confidentiality on the other hand refers to an agreement between an individual and a covered entity regarding how the entity will protect the privacy of the individual's health information.

Although the HIPAA Privacy Rule preempts conflicting state laws, the Rule, as noted, allows states to create more stringent confidentiality laws.<sup>112</sup> Accordingly, some states have created health information laws that have a broader scope than HIPAA. Many state laws extend the notice, access, amendment, and safeguard requirements to a broader range of entities—e.g., pharmaceutical companies.<sup>113</sup> For example, California has created the California Confidentiality of Medical Information Act (CMIA), which requires that pharmaceutical companies preserve the confidentiality of the medical records they possess; the law also requires that covered entities receive written authorization to disclose those records.<sup>114</sup> Additionally, California's law also contains a provision requiring employers that receive medical information from their employees to make certain efforts to maintain the confidentiality of the information and prevent its

unauthorized disclosure.<sup>115</sup> Likewise, Texas state law extends the breadth of what is considered a covered entity to include any person who “comes into possession of protected health information” or “obtains or stores protected health information.”<sup>116</sup> Accordingly, state confidentiality laws must be considered in the formation of repositories holding registry data.

## **Example of a Federally Administered Data Repository Holding Data From Multiple Studies**

The NIDDK Central Data Repository provides an important example of a federally supported repository of multiple studies. In 2003, NIH set up the NIDDK Repository to increase the impact of current and future NIDDK-funded studies. Funded and operated by NIDDK, the repository archives and distributes the results of studies provided by contributing investigators in order to further scientific progress and make data readily available to all investigators in the research community. All relevant NIH-funded awards greater than \$500,000 and with sufficient sample size as determined by NIDDK are required to submit their data to the Repository or submit a statement explaining why that is not possible.

Requirements for data contributors are set forth clearly in the Institute’s Requests for Applications (RFAs), which include reference and incorporation of the NIH Grants Policy Statement and Terms and Conditions of NIH Grant Awards. In keeping with Federal laws allowing a Federal interest in information created with Federal award support, the RFA stipulates that NIDDK-funded research will be made available to the Repository, provided that research subjects consent to such inclusion.

The NIDDK Repository does not receive or contain individually identifiable data or codes linking such information to any of the datasets or studies included in the Repository with the exception of dates of service with some submissions. Where dates are included, the data are considered an LDS for HIPAA purposes. Provided that DUAs are entered into with users, HIPAA requirements would be met. Numerous examples of other Federal clinical data repositories follow a similar pattern, that is, the repositories do not contain individually identifiable data (e.g., NDAR;<sup>26</sup> NHLBI BioLINCC;<sup>34</sup> and the National Institute of Neurological Disorders and Stroke [NINDS] repository<sup>117</sup>). This eliminates any HIPAA Privacy or Security Rule implications as well as any Privacy Act of 1974 or FISMA concerns, since the entities that are disclosing data are disclosing it in an LDS (provided a proper DUA is executed as described below) or de-identified format only. It should be noted, however, that even were these Federal repositories to hold PHI, they would most likely not be directly bound by HIPAA, since they do not fall into the definition of either a covered entity (unless they are acting as a clearinghouse as defined by HIPAA) or a business associate of a covered entity. However, the registries included in the Repository are likely compilations created by covered entities (e.g., health plans or providers) or business associates of these covered entities. All of these actors would be bound by HIPAA requirements. The registries might also be compilations created by researchers who are neither covered entities nor business associates of a covered entity, but in this case, researchers would be bound by the terms of a DUA with a covered entity that stipulates the extension of HIPAA protections. As a result, while HIPAA does not bind the repositories directly, the repositories operate with a full expectation that individual privacy principles will govern, and have designed their operations to include only LDS or de-identified data. This step undoubtedly helps with not only the challenge of compliance with Federal privacy laws but also the need to secure individual consent. In the event that a repository were to hold individually identifiable

data, both Federal privacy laws and the Common Rule would be implicated and IRB decisions would need to be made regarding which exemptions might apply.

Approved users of data in the NIDDK Repository must enter into a DUA with NIH to access information in the Repository, and users may use the contents only for approved research. Review of proposed projects is conducted internally by the NIDDK Repository Specialists Office. Users must provide a high-level abstract of their proposed work that includes the study objectives, background, importance and design of their research. The application process is handled electronically. The DUA also includes specific data security and non-transferability requirements as well as the following important requirements:

- Approved users recognize any restrictions on data use delineated within the original informed consent agreements of contributing studies, as identified by the submitting institutions and stated on database Web sites.
- Approved users acknowledge that IRB approval or waiver is required for use of datasets in the repository.
- Approved users agree not to use the requested dataset(s), either alone or in concert with any other information, to identify or contact individual participants from whom data were collected.

## **Summary of Legal Considerations**

This analysis shows that numerous Federal and state laws potentially come into play where the subject is creation of federally administered repositories holding multiple registry data, particularly data created with Federal funding support. Whether the registries linked to a repository are expired or ongoing does not appear to be the issue. Instead, the critical factors appear to be whether the information will be individually identifiable, how data will be maintained and used, whether the uses of the data will include research, and the sources from which the data will be gathered. The inclusion of data from deceased individuals raises additional questions that must be addressed prior to data use. State law also may affect the creation and management of repositories containing registry data, as well as access to and use of data.

Examples of ongoing Federal repositories, such as the NIDDK Repository, underscore that these potential legal issues can be managed successfully, thereby permitting research to proceed. Important issues in Federal agency repository administration include interagency agreements that allow federally supported registries created under one program to be linked to repositories of multiple agencies, as well as a Federal governance structure that assures compliance with Federal and state laws related to privacy, security, confidentiality, and human subject research, among other matters. Notably, however, referenced repositories either do not accept or hold individually identifiable data or in the case of the NIDDK Repository include dates of service (constituting an LDS). This de-identified or limited nature of the data reduces legal privacy and security concerns, although expectations of privacy and security may still govern and DUA rules still apply. To navigate these issues and create a governance arrangement, AHRQ may wish to establish an expert advisory panel consisting of individuals with expertise in law, ethics, research methods, security, and other relevant matters.

Despite the fact that other repository examples do not include PHI, there are advantages to repositories that hold individually identifiable registry data (e.g., broader research). If AHRQ determines it would like its repository to hold PHI, it will be critically important for AHRQ to explicitly establish a Federal interest in registries created as part of a project award and by

condition awards involving the creation of registry data on the donation of such data to a larger Federal repository. In this way, AHRQ positions itself to assure that appropriate individual consents for disclosure and research are obtained at the time that the registry is created. In terms of securing repository data consisting of registries that were not so linked to the repository from the outset, it may be that the data collected in this fashion will need to be de-identified unless one of the HIPAA Privacy Rule's exemptions is determined to apply. For example, were provision of registry data to a repository to be considered part of a covered entity's health care operations, an exemption might be possible; similarly, were the AHRQ repository to enter into a business associate arrangement with the registry, then transfer presumably would be possible. Research using identifiable data might then be possible assuming that the activity qualifies under the Common Rule's minimum risk exemption.

## **Proposals for a Repository of Expired Registries**

Using the information gained from the background research and discussions, the questions raised earlier in this paper can now be addressed:

1. Should there be a repository of expired registries?
2. What are the requirements for a repository of expired registries?
3. What models might AHRQ use to develop and maintain a repository of expired registries?
4. What would motivate researchers to donate their registries to a repository of expired registries?
5. What are the costs of setting up and maintaining a repository of expired registries?

## **Should There Be a Repository of Expired Registries?**

The clear finding from background research and stakeholder discussions is that there is interest in the research community in developing a repository of expired registries to facilitate future research. Several repositories exist for clinical study data in specific disease areas, and the data in these repositories are used to support new research projects, including linkage projects and analyses to inform future studies. No such repository currently exists for registry data. A new repository of expired registries could fill that gap and ensure that registry data are not lost when a registry closes. In addition, the repository could hold archived data from long-term, ongoing registries (e.g., quality improvement registries with no defined endpoint). Despite the interest in this concept, stakeholders did not see a current business case for the repository, making it unlikely that a private entity would develop such a repository on its own. Stakeholders suggested that the repository would need to be developed and supported with government funding, such as by AHRQ or an AHRQ contractor.

If AHRQ were to pursue the development of a repository of expired registries, the Agency could present a clear rationale for the repository based on two factors. First, AHRQ could draw on the precedent of archiving and storing clinical study data for future use established by NIH. NIH strongly supports continued access to and sharing of clinical study data. The NIH policy on data sharing states:

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.<sup>118</sup>

AHRQ could develop a similar data sharing policy and apply it to registries funded by the Agency, requiring that these registries donate their data to the repository of expired registries at the conclusion of the project. This could be particularly important for ensuring that resources



invested in developing data sources for comparative effectiveness research remain available for future use.

Second, AHRQ has invested significant resources in developing good practices for registries and improving the usefulness and quality of registry data. AHRQ is also developing the RoPR system to promote collaboration, reduce redundancy, and improve transparency in registry research. The RoPR system represents a large step forward in increasing the visibility of registries as tools for clinical research. Taken together, these registry-related activities have established AHRQ as a major supporter of the use of registries to conduct practical, high-quality clinical research. The development of a repository of expired registries could be viewed as a continuation of AHRQ's support for patient registries and as an investment in preserving existing resources, particularly those created with AHRQ funding, and promoting efficiency in health care research.

A repository of expired registries would both allow AHRQ to promote data sharing and continue AHRQ's support for the efficient use of high-quality patient registries. Alternately, AHRQ could promote data sharing, at least pertaining to studies that it funds, by requiring data sharing plans as a condition of funding. Some NIH Institutes have created data repositories to facilitate data sharing, while others rely on investigators to create their own plans for making data available after the close of the study. AHRQ could use the latter model, which would incur no additional costs to the Agency.

The development of a repository of expired registries is feasible from operational, legal, ethical, and technical standpoints, as described in the "Stakeholder Perspectives on Major Issues" and "Legal Issues in Creating a Repository of Expired Registries" sections. Background research and stakeholder discussions identified several potential objectives for the proposed repository of expired registries. The repository's primary objective would be to promote the efficient use of limited health care research funding by ensuring that previously collected data are not lost when a registry expires. Second, the repository may encourage and facilitate the use of registries for clinical research. Finally, the availability of registry data and supporting documentation (CRFs, protocols or study plans) may encourage collaboration and promote the use of standardized data elements or outcome measures.

While a repository of expired registries could be developed to meet these objectives, the value of the repository will depend on the quality and quantity of the data in the repository. AHRQ could require that registry data developed with AHRQ funding be donated to the repository, but motivating other researchers to donate their data to the repository will be a challenge. The existing repositories discussed here primarily rely on the NIH requirement for data sharing to obtain data. While there may be some value in storing even a small number of databases, the value of that effort would need to be weighed against the costs of the different approaches for developing and maintaining the repository.

## **What Are the Requirements for a Repository of Expired Registries?**

Based on stakeholder discussions and discussions with other relevant organizations, requirements were developed for the repository of expired registries. To fulfill its primary objective of ensuring that previously collected data are not lost when a registry expires, the repository would need to meet the basic requirements described below. Meeting the secondary objectives would result in additional requirements.

## **Basic Requirements**

The repository will need sound, transparent governance policies and procedures, and the governing body should include a broad group of stakeholders, including Federal, industry, provider, researcher, and patient representatives. The governing body will need to develop a transparent data access process. The process should include a formal data request, review of the request, and completion of a data use agreement between the requestor and the repository. The data request should include IRB approval (or an IRB determination that the research is exempt from such approval) for the proposed research project and information about the investigator. A data access committee or other review mechanism will be necessary to review and approve requests to use repository data. The data use agreement should cover requirements for data storage and security and, for data linkage projects, it should outline precautions that should be taken to prevent accidental re-identification of patients. The repository will also need to store information on the original study and how the data were collected to ensure that researchers can use the data to draw valid conclusions. The data access committee should also have clear guidelines on acceptable applications for use and requirements for users. These guidelines should be developed by the governing body, with input from stakeholders. Questions to consider include:

- Should all studies be required to be submitted for publication or presentation, or otherwise made available in the public domain? What is the role of the registry donator in the publication of additional analyses using the registry's data?
- Should multiple permissions for data access be granted to researchers pursuing nearly identical lines of research? If not, should researchers who have donated expired registries be given any preferential access to other data?
- Would commercial uses, such as understanding incidence or prevalence of certain conditions, be permitted for the goal of promoting product development?
- Should data be made available for non-research purposes, such as for use in litigation support to understand incidence and prevalence of certain conditions, or to support or challenge research on causal inference?
- What fees will be set for accessing the data? Will fees be the same for commercial uses and noncommercial uses?
- Should registry donators retain any control over how their registry data are used for future research projects?
- Should all research requests be publicly disclosed?

The repository will also need policies and procedures for reviewing and accepting donated data. These policies should include a requirement for institutional certification of donated data. Institutional certification will be used to determine if the data can be donated, if identifiers can be included, and for what research purposes the data may be used. The repository will need to develop a certification policy. In addition, the repository will need to either have a procedure in place for comparing proposed research projects' uses against appropriate data uses for each dataset or only accept datasets without restrictions on use. The repository will also need a policy on data ownership. Again, the governing body should develop these policies, with input from stakeholders.

At the most basic level, the technical requirements for the repository include a Web site to provide information on the repository and a secure location to store datasets. Methods for

encrypting the data and appropriate security and access controls would be necessary if the repository is storing identifiable information.

### **Additional Requirements To Meet Secondary Objectives**

The repository could include a large number of additional features to meet secondary objectives, which would result in additional requirements. For example, additional requirements may include policies and procedures for data quality audits, assembly of guidelines for data use, review of proposed research projects for scientific validity, review of papers and abstracts generated from repository data, and efforts to transform data into standard formats.

From a technology standpoint, a more sophisticated repository would require a Web interface with search tools to identify relevant databases and workflow management tools to request and transfer datasets. Resources would be necessary to convert donated databases to a common database technology and to create a summary level data model that is searchable. Encryption tools and appropriate security and access controls would also be necessary for the storage of identifiable data and would be more complex or costly than those required for the basic model due to the manipulation of the data.

## **What Models Might AHRQ Use To Develop and Maintain a Repository of Expired Registries?**

Stakeholders described two possible models for developing and maintaining a repository of expired registries: a basic data archive model and a data archive with research support model. The basic data archive model represents a low-budget, basic approach to data archiving, while the data archive with research support model represents a high-budget, sophisticated approach. These models are described below and compared in Table 3. It is important to note that many intermediate models are possible; these models would use the basic approach and incorporate one or more features from the sophisticated approach. The intermediate models are not described here because stakeholders did not prioritize the additional features that could be included in such a model. Additional stakeholder discussions and public comment would be necessary to define the intermediate models.

### **Model 1. Basic Data Archive**

This data archive model would meet the basic requirements outlined above and would fulfill the primary objective identified by stakeholders of ensuring that previously collected data are not lost when a registry expires. The repository would serve as an archive for registries that were created with AHRQ funding and for registries that were voluntarily donated by their owners. In this model, the repository would function as an archive whose primary purpose is to store datasets and transfer them to approved researchers. The archive would only accept data with no restrictions on future use, and data access requests from qualified investigators with IRB approval would be approved with no further review. The repository would provide any available information on the donated datasets to researchers, but no efforts would be made to develop or obtain additional documentation. No additional quality assurance activities would be conducted. A limited Web site would be developed to provide information on the available datasets, primarily by referring users to the RoPR system, and to provide information on the data access process. Search capabilities would not be available. The data would be stored in their original format in a secure system with regular back-up procedures, and data that contain patient

identifiers would be encrypted and stored with appropriate access controls. Researchers would receive data transfers on a CD-ROM sent by mail or through a secure server.

While this model requires relatively few resources, there are several limitations to this approach. First, the repository in this model essentially acts as a data warehouse, receiving donated datasets and distributing them for new research projects. The repository does not add any value to the donated data. Researchers must be proactive to identify datasets and obtain any necessary information on the data to inform their findings. Because of the lack of research support, some researchers may not be interested in using the repository data. Other researchers may draw erroneous conclusions from data because they are not fully informed of the conditions under which the data were originally collected. Lastly, registry sponsors may be less willing to donate their data to such a repository if they do not perceive the repository as adding value.

## **Model 2. Data Archive With Research Support**

A second approach is to develop a data archive with some level of research support. This model would meet the primary objective identified by stakeholders of ensuring that previously collected data are not lost when a registry expires. In addition, this model would meet the secondary objectives of encouraging and facilitating the use of registries for clinical research and promoting collaboration and the use of standardized data elements or outcome measures. As with Model 1, the repository would serve as an archive for registries that were created with AHRQ funding and for registries that were voluntarily donated by their owners. In Model 2, the repository would both store datasets and provide support for researchers interested in using repository data. The repository would accept all data, even those with restrictions on use, and procedures would be in place to cross-check research requests with restrictions on use. Patient identifiers would be included in the repository, and data access would be granted at varying levels depending on whether the dataset contains identifiers. Data requests would be reviewed for scientific validity and merit, and results of the research projects would also be reviewed. The repository staff would assemble data use packages for each registry dataset. The data use packages may include documents provided by the registry owner, such as the CRFs, data dictionary, and protocol. The repository staff could assemble additional information, such as summaries of quality assessment activities or lists of publications using registry data, for inclusion in the data use packages.

Donated databases would be converted to a common database technology, which would allow the repository staff to manage and navigate the databases using common database tools and facilitate updates to the databases as technology changes over the duration of the repository. Data would also be analyzed to extract common summary level data elements, such as demographics, vital signs, and diagnoses. This information would be used to create a summary level data model that could be searched by users. The repository Web site would include a search interface to allow users to query the repository to identify, for example, all datasets that include males between the ages of 60 and 80 with a particular diagnosis. The user could then complete a data request to access those datasets.

In addition to search tools, the repository Web site would include profiles of each registry, the data use documents, information on restrictions for use and presence of identifiers, and quality assurance findings. The Web site would also include a workflow management tool to allow users to initiate data requests online, track the status of their requests, and, depending on the size of the dataset to be downloaded, access the archived dataset electronically.

The major limitation of this approach is the significant amount of resources that it requires. This model represents a comprehensive approach to data sharing, but the value of the repository will still depend on the data that are donated. With this model, the repository would need to actively seek new data donations to grow the available resources and find ways to publicize the availability of the data to encourage use. Because of the additional services that are provided, the repository could charge fees for data access or for related research support services. However, this may deter use of the repository.

**Table 3. Comparison of repository of expired registry models**

	<b>Basic Data Archive</b>	<b>Data Archive With Research Support</b>
<b>Governance</b>	Governing body, including Federal, industry, researcher, provider, and patient representatives, provides oversight.	Governing body, including Federal, industry, researcher, provider, and patient representatives, provides oversight.
<b>Research Ethics</b>	Institutional certification is required for donated data. Only data with no use restrictions are accepted.	Institutional certification is required for donated data. All data, even those with restrictions on use, are accepted. Restrictions on use are documented and cross-checked against research requests to ensure appropriateness of use.
<b>Patient Privacy</b>	Identifiable data are stored, when permissible according to institutional certification. Requestors must have IRB approval to receive identifiers (e.g., data cannot be de-identified for some projects).	Identifiable data are stored, when permissible according to institutional certification. Data requests are reviewed to determine if identifiable information is necessary for the request. If identifiers are necessary, they are released; otherwise, de-identified data are released.
<b>Data Access</b>	Data access policy requires a data use request describing the data to be used and showing IRB approval (or waiver) for the proposed project. Researchers must sign a data use agreement to receive the data. Archive does not review methods or final products of analyses using archive data.	Data access policy requires a data use request describing the data to be used and proposed research project and showing IRB approval (or waiver) for the proposed project. Governing body (or a subcommittee) reviews requests for scientific validity and merit. Approved researchers must sign a data use agreement to receive the data. For linkage projects, governing body (or subcommittee) reviews methods and final product to avoid accidental re-identification of patients. For all projects, governing body (or subcommittee) reviews publications based on repository data.

**Table 3. Comparison of repository of expired registry models (continued)**

	Basic Data Archive	Data Archive With Research Support
<b>Supporting Documentation</b>	Repository provides any information that is donated with the registry data to researchers. This information is provided upon request or with approved data requests.	Repository assembles detailed data use documents for each registry dataset, including CRFs, data dictionary, protocol, and other information. Data use documents are posted online and provided with approved data request.
<b>Identifying Datasets</b>	The repository Web page lists all available datasets, grouped by disease area, and links each dataset to the registry's RoPR listing. Search capabilities are not available.	Common summary level data (e.g., demographics, disease areas) are extracted for each study and are searchable. Additional search tools are available to identify studies based on data use restrictions, sample size, etc. Profiles of each dataset are posted on the repository Web site. Profiles include the basic information contained in RoPR, plus the supporting documentation for the registry. Repository staff may assist researchers in identifying relevant datasets.
<b>Data Quality Assurance</b>	Repository maintains records of what quality assurance was done in the registry and provides that information with data requests.	Repository maintains records of what quality assurance was done in the registry and provides that information with data requests. Repository conducts quality assessments, such as describing percent missing for key variables or lost to follow-up rates for key time intervals (e.g., percent lost at one year, five years, etc.).
<b>Data Format and Transfer</b>	Data are stored and transferred in their original format. Data are mailed to researchers on a CD-ROM.	Data are converted to a common database technology. Common summary level data are extracted for each dataset and made available through a search interface. Researchers receive data via a secure online interface or via mail, depending on their preference.

## What Would Motivate Researchers To Donate Data to the Repository of Expired Registries?

In addition to selecting a model for the repository, AHRQ must consider how to incentivize registry sponsors to donate their data to the repository of expired registries. As discussed above, AHRQ could require donation of data from studies that are funded by the Agency. AHRQ could collaborate with other government funding agencies, such as NIH and CDC, to encourage or require donation of studies funded by these groups. In addition to government-funded studies, voluntary donations of registry data would be the other major source of data for the repository. Registry sponsors may donate data because the registry has closed, and they no longer have resources to maintain the data. Sponsors may also donate data if they perceive the repository as providing continuing value beyond that which they are able to fund. Stakeholders suggested that registry owners might be motivated to donate data if liability resulting from the data is limited or transferred to the repository, or if donations resulted in a tax advantage.

In order to develop a robust repository that has value to researchers, AHRQ would likely need to invest resources in identifying and targeting registries for donation and in identifying and creating other incentives to encourage donation. Donation of registries is particularly important if

AHRQ pursues Model 2, as the benefits of developing that repository would only be realized if researchers use the data contained in the repository. The issue of encouraging donations is less critical for Model 1, since the set-up and maintenance costs are lower for that model.

## **What Are the Costs of Setting Up and Maintaining a Repository of Expired Registries?**

Without a full evaluation of each requirement with respect to cost paid by AHRQ or the NIH repositories for similar work, cost information can only be generally estimated. However, it is clear that the two models have very different cost implications. Model 1, building a basic data archive with limited administrative management, is the least expensive option. Model 2, building a data archive with extensive curation and research support, would be significantly more expensive, depending on the number and type of additional features that are included.

Using the available data and making additional estimates for technological investments in software and hardware, estimated costs can be generated for Model 1 (patterned after the NHLBI repository) and Model 2 (patterned after NDAR). These setup costs are estimated to range from approximately \$1,000,000 for a limited archive to greater than \$6,000,000 for an archive that includes more significant research support and is likely to include a data enclave or similarly advanced technological investments. In terms of ongoing operating costs, the available data on these and other NIH repositories suggests costs in the range of \$7,500 to \$120,000 (from a limited repository such as NHLBI to sophisticated repository such as NDAR) per database per year in the repository. It is prudent to assume that a minimal base level of infrastructure is required in each setting, and, therefore, the costs for limited research support models such as NHLBI should serve as a floor. Current programs range from approximately 20 to 80 studies archived. While efficiencies of scale could be imagined for these programs, it should not be assumed for cost estimation purposes that there will necessarily be efficiencies of scale at the same level of base infrastructure investment. It is simply unknown. Therefore, assuming a repository with 50 registries, with a similar level of use activity to those described in this paper (few requests to up to 10 requests per month), we estimate annual operating expenses of the repository in the range (from Model 1 to Model 2) of \$500,000 to \$6,000,000 depending on the level of research and technological support provided. As described, these estimates are entirely dependent on the number of procedures or technological features incorporated from Table 3; the number of new datasets being archived by the repository each year; the number of existing datasets being maintained in the repository; and the frequency and complexity of external data use requests. Estimated cost ranges from this example are also shown in Table 4.

In addition to the direct funding by AHRQ, there are other potential financing options. These include seeking financial support from public and private entities that might benefit from the repository of expired registries (e.g., CMS, FDA, medical professional societies, manufacturers). External support will be easier to obtain once success has been demonstrated in collecting registry datasets.

**Table 4. Estimated costs by model of setup and maintenance for a repository of expired registries**

	Data Archive Model	Data Archive and Research Support Model
<b>Setup</b>	\$\$\$	\$\$\$\$
<b>Annual Maintenance*</b>	\$	\$

\$ = under \$500,000; \$ = \$500,000 to \$1,000,000; \$\$\$ = \$1,000,000 to \$6,000,000; \$\$\$\$ = over \$6,000,000

\*Costs are highly dependent on procedures/features from Table 3; the number of new datasets being archived in the repository; the number of existing datasets being maintained; and the frequency and complexity of external data use requests.

## **Conclusions**

Background research and discussions with stakeholders clearly demonstrated that there is interest in a repository of expired registries. Existing repositories are focused on specific disease areas and largely include data from clinical trials. No repository exists for registry data currently. Government agencies, medical associations, foundations, and health services researchers all noted the potential value of a repository of expired registries and identified key objectives of such a system. Despite the interest in this concept, stakeholders did not see a current business case for the repository, making it unlikely that a private entity would develop such a repository on its own. Stakeholders suggested that the repository would need to be developed and supported with government funding, such as by AHRQ or an AHRQ contractor. If AHRQ were to pursue the development of a repository of expired registries, the Agency could present a clear rationale for the repository based on the importance of data sharing for efficient use of limited health research resources and the Agency's continuing efforts to encourage the use of high-quality registries for clinical research.

Despite the broad interest in a repository of expired registries, a key issue is the lack of current incentives for registry owners to donate their data. AHRQ may be able to require donation of data from registries funded by the Agency, but few incentives exist to encourage other registry sponsors to contribute their data. Donation of data will be critical to establish the value of the repository. Prior to investing resources in developing a sophisticated repository, such as Model 2, it may be necessary to work with stakeholders to identify donation incentives to ensure that the repository has sufficient data to support research projects. Alternately, AHRQ may pursue an incremental approach to developing the repository, in which the initial repository follows Model 1 and additional features are added as the repository data archive grows and its use requirements are better understood.

Additional discussions with both stakeholders and registry owners are highly recommended to help AHRQ further define requirements, priorities, incentive structures, and funding options in planning next steps.



## References

1. Gliklich RE, Dreyer NA, eds. Registries for Evaluating Patient Outcomes: A User's Guide. (Prepared by Outcome DEcIDE Center [Outcome Sciences, Inc. dba Outcome] under Contract No. 290-2005-0035-I.) AHRQ Publication No. 07-EHC001-1. Rockville, MD: Agency for Healthcare Research and Quality. April 2007.
2. Dhruva SS, Phurrough SE, Salive ME, et al: CMS's landmark decision on CT colonography—examining the relevant data. *N Engl J Med* 2009;360(26):2699-2701.
3. Centers for Medicare & Medicaid Services. Guidance for the Public, Industry, and CMS Staff: National Coverage Determinations with Data Collection as a Condition of Coverage: Coverage with Evidence Development. July 12, 2006.
4. Kennedy L, Craig AM. Global registries for measuring pharmacoeconomic and quality-of-life outcomes: focus on design and data collection, analysis and interpretation. *Pharmacoeconomics* 2004;22(9):551-68.
5. Institute of Medicine. Learning What Works Best: The Nation's Need for Evidence on Comparative Effectiveness in Health Care. 2007:1-80.
6. Congressional Budget Office. Research on the Comparative Effectiveness of Medical Treatments: Issues and Options for an Expanded Federal Role. 2007. Available at: [www.cbo.gov/ftpdocs/88xx/doc8891/12-18-ComparativeEffectiveness.pdf](http://www.cbo.gov/ftpdocs/88xx/doc8891/12-18-ComparativeEffectiveness.pdf). Accessed March 23, 2011.
7. McClellan M, Benner J, Garber AM, Meltzer DO, Tunis SR, Pearson S, The Brookings Institution. Implementing Comparative Effectiveness Research: Priorities, Methods, and Impact. Available at: [www.brookings.edu/~media/Files/events/2009/0609\\_health\\_care\\_cer/0609\\_health\\_care\\_cer.pdf](http://www.brookings.edu/~media/Files/events/2009/0609_health_care_cer/0609_health_care_cer.pdf). Accessed March 23, 2011.
8. National Institutes of Health. NIH Challenge Grants in Health and Science Research. Available at: [grants.nih.gov/grants/funding/challenge\\_award/](http://grants.nih.gov/grants/funding/challenge_award/). Accessed March 23, 2011.
9. Holve, E. and P. Pittman, A First Look at the Volume and Cost of Comparative Effectiveness Research in the United States. AcademyHealth. June 2009.
10. U.S. Food and Drug Administration. Postmarket Requirements (Medical Devices): Post-Approval Studies. Available at: [www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMA/pma\\_pas.cfm](http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMA/pma_pas.cfm). Accessed March 23, 2011.
11. Gliklich RE, Leavy MB, Campion DM. Developing a Registry of Patient Registries: Options for the Agency for Healthcare Research and Quality. Effective Health Care Program Research Report No. 34. (Prepared by Outcome DEcIDE Center [Outcome Sciences, Inc. dba Outcome] under No. 290-2005-0035-I.) AHRQ Publication No. 10(12)-EHC036-EF. Rockville, MD: Agency for Healthcare Research and Quality. February 2012.
12. RTI International. "About the Central Data Repository (CDR)." Available at: <https://www.niddkrepository.org/niddk/jsp/public/about.jsp>. Accessed February 24, 2011.
13. Cuticchia AJ, Cooley PC, Hall RD, Qin Y. NIDDK data repository: a central collection of clinical trial data. *BMC Med Inform Decis Mak.* 2006;6:19.
14. RTI International. "Publications Based on Repository Data." Available at: <https://www.niddkrepository.org/niddk/preCitationsByStudyName.do>. Accessed March 1, 2011.
15. Philip Cooley; RTI International. Personal correspondence, February 7, 2011.
16. Rebekah Rasooly, PhD; National Institute of Diabetes and Digestive and Kidney Diseases. Personal correspondence, February 15, 2011.
17. RTI International. "Requirements for Providing Study Data Archives to the NIDDK Central Data Repository (CDR)." Available at: [https://www.niddkrepository.org/niddkdocs/resources/Update\\_ArchiveReqs\\_v3.9.pdf](https://www.niddkrepository.org/niddkdocs/resources/Update_ArchiveReqs_v3.9.pdf). Accessed March 2, 2011.

## Effective Health Care Program Research Report Number 42

18. RTI International. "Quality Control Procedures for the NIDDK Data Repository." Available at: [https://www.niddkrepository.org/niddkdocs/integrity/RTIPProcessDescription\\_4.pdf](https://www.niddkrepository.org/niddkdocs/integrity/RTIPProcessDescription_4.pdf). Accessed February 24, 2011.
19. RTI International. "Cost for Data and Samples." Available at: <https://www.niddkrepository.org/niddk/jsp/public/cost.jsp>. Accessed February 24, 2011.
20. RTI International. "Acknowledgements." Available at: <https://www.niddkrepository.org/niddk/jsp/public/acknowledgements.jsp>. Accessed March 2, 2011.
21. National Database for Autism Research. "Frequently Asked Questions." Available at: <http://ndar.nih.gov/ndarpublicweb/faq.go#FAQ01>. Accessed March 1, 2011.
22. Matthew McAuliffe, PhD; NIH Center for Information Technology (CIT). Personal correspondence, February 11, 2011.
23. National Database for Autism Research. "NDAR Contributors." Available at: <http://ndar.nih.gov/ndarpublicweb/aboutNDAR.go#contributors>. Accessed March 2, 2011.
24. Tu SW, Tennakoon L, O'Connor M, Shankar R, Das A. Using an integrated ontology and information model for querying and reasoning about phenotypes: The case of autism. *AMIA Annu Symp Proc* 2008:727-31.
25. Johnson SB, Whitney G, McAuliffe M, Wang H, McCreedy E, Rozenblit L, et al. Using global unique identifiers to link autism collections. *J Am Med Inform Assoc*. 2010 Nov 1;17(6):689-95.
26. National Database for Autism Research. "Policy for the National Database for Autism Research." Available at: [http://ndar.nih.gov/ndarpublicweb/Documents/NDAR\\_Policy.pdf](http://ndar.nih.gov/ndarpublicweb/Documents/NDAR_Policy.pdf). Accessed March 2, 2011.
27. National Institutes of Health. Genome-Wide Association Studies. Available at: [http://grants.nih.gov/grants/gwas/gwas\\_ptc.pdf](http://grants.nih.gov/grants/gwas/gwas_ptc.pdf). Accessed March 31, 2011.
28. National Database for Autism Research. "Data Sharing." Available at: <http://ndar.nih.gov/ndarpublicweb/sharing.go#research>. Accessed March 3, 2011.
29. National Database for Autism Research. "Tools." Available at: <http://ndar.nih.gov/ndarpublicweb/tools.go>. Accessed March 3, 2011.
30. NIH Research Portfolio Online Reporting Tools (RePORT). Project Information. Available at: [http://projectreporter.nih.gov/project\\_info\\_details.cfm?aid=8164199&icde=7323428](http://projectreporter.nih.gov/project_info_details.cfm?aid=8164199&icde=7323428). Accessed March 7, 2011.
31. National Heart Lung and Blood Institute. "About BioLINCC." Available at: <https://biolincc.nhlbi.nih.gov/about/>. Accessed March 3, 2011.
32. Carol Giffen, DVM; Director, Clinical Research, IMS. Personal correspondence, 16 March 2011.
33. National Heart Lung and Blood Institute. "Biologic Specimen and Data Repository Operational Guidelines." [https://biolincc.nhlbi.nih.gov/static/guidelines/guidelines.html#\\_Toc241381324](https://biolincc.nhlbi.nih.gov/static/guidelines/guidelines.html#_Toc241381324). Accessed March 3, 2011.
34. National Heart Lung and Blood Institute. "Policy for Dataset Preparation." Available at: [https://biolincc.nhlbi.nih.gov/new\\_data\\_set\\_policy/](https://biolincc.nhlbi.nih.gov/new_data_set_policy/). Accessed March 3, 2011.
35. National Heart Lung and Blood Institute. "NHLBI Research Materials Distribution Agreement (RMDA)." Available at: <https://biolincc.nhlbi.nih.gov/static/RMDA.pdf>. Accessed March 3, 2011.
36. National Heart Lung and Blood Institute. "Advanced Study Search." Available at: <https://biolincc.nhlbi.nih.gov/search/?formonly>. Accessed March 3, 2011.

## Effective Health Care Program Research Report Number 42

37. RECOVERY American Recovery and Reinvestment Act Office of the Secretary (ARRAOS): RFP ST10-1125 Clinical Trials Repository Design and Development for the Food and Drug Administration (FDA) Comparative Effectiveness Research (CER) Initiative. Solicitation Number: ST10-1125. FedBizOpps.gov. Available at: <https://www.fbo.gov/spg/HHS/NIH/FCRF/ST10-1125/listing.html>. Accessed May 21, 2011.
38. RECOVERY ARROS: ST10-1067 For an initiative called "Janus Infrastructure Development and Implementation for FDA CER" Solicitation Number: ST10-1067. FedBizOpps.gov. Available at: <https://www.fbo.gov/spg/HHS/NIH/FCRF/ST10-1067/listing.html>. Accessed May 21, 2011.
39. Susan Adams, JD; Director, Committee for the Protection of Human Subjects, Dartmouth College. Personal correspondence, 9 February 2011.
40. 45 C.F.R. Part 74.1 et. seq.
41. 45 C.F.R. § 160.103 (2010).
42. 45 C.F.R. § 160.103(1)(i)(A) (2010).
43. 45 C.F.R. § 164.508(a)(1).
44. 45 C.F.R. § 164.508 (a)(2).
45. 45 C.F.R. § 164.508 (a)(3).
46. Pub. L. No 11-5, Div. A, Title XIII, §13405(d)(2), 123 Stat. 264-68.
47. 45 C.F.R. § 164.508 (a)(3)(ii), (c)(1)-(2).
48. 45 C.F.R. § 164.502(f) (2010).
49. 75 F.R. 40894 (July 14, 2010).
50. 45 C.F.R. §§ 164.502 (a)(1)(vi) and 164.510.
51. 45 C.F.R. § 164.512(i).
52. 45 C.F.R. § 164.514(e).
53. 45 C.F.R. § 164.514(a).
54. Modifications to the HIPAA Privacy, Security, and Enforcement Rules, 75 Fed. Reg. at 40,919-20 (to be codified at 45 CFR § 164.504).
55. 45 C.F.R. § 164.304.
56. 45 C.F.R. § 164.306.
57. *See, e.g.*, Nat'l Inst. Of Standards & Tech., U.S. Dep't of Commerce, NIST Special Publication 800-66, An Introductory Resource Guide for Implementing the HIPAA Security Rule (2008), available at <http://csrc.nist.gov/publications/nistpubs/800-66-Rev1/SP-800-66-Revision1.pdf>.
58. Breach Notification for Unsecured Protected Health Information, 74 Fed. Reg. 42,740 (August 24, 2009) (to be codified at 45 C.F.R. pt. 160 and 164). Available at: <http://edocket.access.gpo.gov/2009/pdf/E9-20169.pdf>.
59. 45 C.F.R. § 164.402.
60. Guidance Specifying the Technologies and Methodologies That Render Protected Health Information Unusable, Unreadable, or Indecipherable and Request for Information, 74 Fed. Reg. 19,006 (Apr. 27, 2009). Available at: [www.hhs.gov/ocr/privacy/hipaa/understanding/coveridentities/federalregisterbreachrfr.pdf](http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveridentities/federalregisterbreachrfr.pdf).
61. 45 C.F.R. § 164.402.
62. 45 C.F.R. § 46.101.
63. 46 C.F.R. § 46.101(b)(4).
64. 45 C.F.R. § 46.102 (d).
65. 45 C.F.R. § 46.103(a).
66. 45 C.F.R. § 46.111.
67. 45 C.F.R. § 46.117(c)(2).
68. J. Zoe Beckerman, Joy Pritts, Eric Goplerud, Jacqueline Leifer, Phyllis Borzi, and Sara Rosenbaum, Health Information Privacy, Patient Safety, and Health Care Quality, Issues and Challenges in the Context of Treatment for Mental Illness and Substance Abuse, BNA Health Policy Report 16:2 (Jan. 14, 2008)
69. 42 C.F.R. § 2.11
70. Privacy Act of 1974, Pub. L. No. 93-579, § 3, 88 Stat. 1896, 1896 (codified as amended at 5 U.S.C. § 552a (2006)).
71. 5 U.S.C. § 522a(c).
72. 5 U.S.C. § 522a(e)(1).
73. 5 U.S.C. § 522a(e)(2), (5).
74. 5 U.S.C. § 522a(g)

## Effective Health Care Program Research Report Number 42

75. The Federal Information Security Management Act of 2002 (FISMA), Pub. L. No. 107-347, 44 U.S.C. § 3541, et seq. (2009).
76. Policy for Privacy Impact Assessments, Office of the Chief Information Officer, Office of the Assistant Secretary for Resources and Technology, Department of Health and Human Services, Document Number: HHS-OCIO-2009-0002. Available at: [www.dhhs.gov/ocio/policy/policydocs/2009-0002.001.doc](http://www.dhhs.gov/ocio/policy/policydocs/2009-0002.001.doc). Accessed March 31, 2011.
77. House Report, Commerce Security Act of 1987, H.R. Rep. 100-153, pt. 1 (predecessor to FISMA); See also HHS Memo on FISMA and Grants, Memorandum from Jaren Doherty, Chief Information Security Officer, October 29, 2007. Available at: <http://oamp.od.nih.gov/NewsEvents/Symposium/Symposium08/Slides/HHSFISMAandGrantsMemo.ppt#257,1>, HHS Memo On FISMA and Grants. Accessed March 31, 2011.
78. The Federal Information Security Management Act of 2002 (FISMA), 44 U.S.C. § 3544(b).
79. The Federal Information Security Management Act of 2002 (FISMA), 44 U.S.C. § 3545.
80. See CMS standard Data Use Agreement, [http://www.resdac.umn.edu/docs/CMS-R-0235\\_06\\_2008.pdf](http://www.resdac.umn.edu/docs/CMS-R-0235_06_2008.pdf).
81. 44 U.S.C. § 3501 (2006).
82. 44 U.S.C. § 3502(3)(A).
83. 44 U.S.C. § 3502(3)(A)(i).
84. 5 C.F.R. § 1320.3(c).
85. 5 C.F.R. § 1320.3(h).
86. 5 C.F.R. § 1320.3 (h)(1).
87. Office of Mgmt. & Budget, Executive Office of the President, Social Media, Web-Based Interactive Technologies, and the Paperwork Reduction Act (Apr. 7, 2010). [www.whitehouse.gov/sites/default/files/omb/assets/inforeg/SocialMediaGuidance\\_04072010.pdf](http://www.whitehouse.gov/sites/default/files/omb/assets/inforeg/SocialMediaGuidance_04072010.pdf). Accessed March 31, 2011.
88. Paperwork Reduction Act of 1980, Pub. L. No. 96-511, 94 Stat 2813 (1980) (codified as amended at 44 U.S.C. §§ 3501-3521).
89. Freedom of Information Act, 5 U.S.C. § 552 (2006), amended by OPEN Government Act of 2007, Pub. L. No. 110-175, 121 Stat. 2524.
90. U.S. Dep't of Justice, Freedom of Information Act Guide (2009). [www.justice.gov/oip/foia\\_guide09.htm](http://www.justice.gov/oip/foia_guide09.htm). Accessed March 31, 2011.
91. U.S. Dep't of Justice, Freedom of Information Act Guide (2009). [www.justice.gov/oip/foia\\_guide09.htm](http://www.justice.gov/oip/foia_guide09.htm). Accessed March 31, 2011. Citing 5 U.S.C. § 552(f)(1) (2006), *amended by* OPEN Government Act of 2007, Pub. L. No. 110175, 121 Stat. 2524.
92. 5 U.S.C. § 552(b)(1)-(9) (2006).
93. 5 U.S.C. § 552(b)(1) (2006).
94. 5 U.S.C. § 552(b)(2) (2006).
95. 5 U.S.C. § 552(b)(3) (2006).
96. 5 U.S.C. § 552(b)(4) (2006).
97. 5 U.S.C. § 552(b)(5) (2006).
98. 5 U.S.C. § 552(b)(6) (2006).
99. 5 U.S.C. § 552(b)(7) (2006).
100. 5 U.S.C. § 552(b)(8) (2006).
101. 5 U.S.C. § 552(b)(9) (2006).
102. *See U.S. Dep't of State v. Washington Post.*, 456 U.S. 595 (1982).
103. 489 U.S. 749 (1989).
104. 489 U.S. 772 (1989).
105. 489 U.S. 773 (1989).
106. 489 U.S. 749, 771 (1989), citing *United States Department of Justice v. Reporters Committee for Freedom of the Press.*
107. U.S. Department of Justice, Freedom of Information Act Guide: Exemption 6 (2009). [www.justice.gov/oip/foia\\_guide09.htm](http://www.justice.gov/oip/foia_guide09.htm). Accessed March 31 2011. Citing *Multi Ag Media LLC v. USDA*, 515 F.3d 1224, 1227 (D.C. Cir. 2008).

## Effective Health Care Program Research Report Number 42

108. *Florida Medical Association, Inc. v. Department of Health Education and Welfare*, 479 F. Supp. 1291 (M.D. Fla. 1979).; *Alley v. U.S. Department of Health and Human Services*, 590 F.3d 1195 (11th Cir. 2009); *Consumers' Checkbook, Center for the Study of Services, v. U.S. Department of Health and Human Services*, 554 F.3d 1046 (D.C. Cir. 2009).
109. See discussion in J. Zoe Beckerman, Joy Pritts, Eric Goplerud, Jacqueline Leifer, Phyllis Borzi, and Sara Rosenbaum, Health Information Privacy, Patient Safety, and Health Care Quality, Issues and Challenges in the Context of Treatment for Mental Illness and Substance Abuse, BNA Health Policy Report 16:2 (Jan. 14, 2008)
110. 45 C.F.R. § 106.203
111. William Roach, Medical Records and the Law, 4th ed. American Health Information Management Association. 2006. p.5. Citing *Whalen v. Roe*, 429 U.S. 589 (1977).
112. 45 C.F.R. § 160.203.
113. William Roach, Medical Records and the Law, 4th ed. American Health Information Management Association. 2006. p.5. Citing *Whalen v. Roe*, 429 U.S. 531 (1977).
114. William Roach, Medical Records and the Law, 4<sup>th</sup> ed. American Health Information Management Association. 2006. p.5. Citing Cal. Civ. Code § 56.05(c).
115. William Roach, Medical Records and the Law, 4<sup>th</sup> ed. American Health Information Management Association. 2006. p.5. Citing Cal. Civ. Code § 56.20.
116. William Roach, Medical Records and the Law, 4th ed. American Health Information Management Association. 2006. p.5. Citing Tex. Health & Safety Code ch.181, § 181.001(b)(2).
117. Discussion with Joanne Odenkirchen, Clinical Research Project Manager at NINDS (Tuesday, February 8, 2011).
118. National Institutes of Health. "Final NIH Statement on Sharing Research Data." <http://grants.nih.gov/grants/guide/notice-files/not-od-03-032.html>. Accessed March 31, 2011.