Published in final edited form as:
Clin Biochem. 2014 March ; 47(0): 252–257. doi:10.1016/j.clinbiochem.2013.11.014.

Sample and Data Sharing: Observations from a Central Data Repository

Mary-Anne Ardini[a], **Huaqin Pan**[a], **Ying Qin**[a], and **Philip C. Cooley**[a,*]
[a]RTI International, PO Box 12194, Research Triangle Park, NC 27709

Abstract

Objectives—From 2003–2013, RTI International served as the data repository for the National Institute of Diabetes, Digestive and Kidney Diseases (NIDDK). RTI worked closely with two sample repository partners to build and maintain the Central Repository (CR) that made data and samples available to approved requestors. In this paper, we recap aspects of establishing the mechanism; detail the challenges and limitations of data and sample sharing, and explore the future of resource sharing in light of the evolving environment of research funding.

Development and Implementation—Effective maintenance required the system to be flexible and dynamic while at the same time compliant with established data standards.

Challenges and Solutions—Our years serving as the CR for NIDDK have yielded a number of observations about the difficulties of running a repository, an operation that is by definition dependent on many outside parties whose degree of expertise and efficiency have a direct impact on repository functioning.

The Path Ahead—The bio-banking industry will likely continue to become more globally centralized for studying specific genetic diseases and monitoring the health of our environment. The dynamic relationship between emerging technologies and the infrastructure will be needed to support future research that requires the ability of organizations providing support to remain flexible even while following established standards.

Keywords

Clinical data repository; biobank; NIDDK

1. Background and Objectives

The NIH Data Sharing Policy, initially released in 2003, requires all investigator-initiated applications with direct costs greater than $500,000 in any single year to incorporate data sharing features in the application. This approach recognizes that the research may have impact beyond the original intent when data can be used by other researchers without undergoing the expense of data collection. In response to this policy, individual institutes of the NIH developed data and biological archives (repositories) to house materials generated from funded studies as a stable, reliable, and cost-effective means for distributing data and materials. Data repositories ensure safe, secure archiving of data and meta-data, enabling

*Corresponding author: Tel: Phone: 1 919 541 6509; Fax: +1 919 316 3539; pcc@rti.org.

continued use in academic and other research environments. The National Library of Medicine now lists 45 NIH Data Sharing Repositories [1]. In addition to clinical research study data, there are resources that aggregate information about mechanistic and genetic data and information sharing systems.

In 2003 when the data sharing policy was initiated, the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) decided to establish a repository to house data and samples from studies they funded. Three separate repositories, collectively known as the 'NIDDK Central Repository' (CR), now enable scientists not involved in an initial study to test new hypotheses without conducting data or bio-specimen collections. The CR stores samples and data from > 70 major multi-site clinical research efforts (125 protocols) in diabetes, digestive, kidney, liver and urologic diseases. In addition, 11 GWAS datasets from these studies are available for request (in collaboration with dbGaP) and DNA samples are available from 24 studies. Table 1 shows a breakdown of studies by disease type.

The CR also provides the opportunity to pool data across several studies to increase the power of statistical analyses. In addition, most NIDDK-funded studies generate genetic material for testing and some carry out high-throughput genotyping, making it possible for other scientists to use repository resources to perform informative genetic analyses using well-curated phenotypic data.

2. Development and Implementation

2.1. System design versus reality

Development began in 2003 with an analysis that considered the requirements of both NIDDK and the scientific community. A complex system composed of primary databases in a private domain and a support database in a public domain was envisaged. Creating databases in both domains was deemed necessary for security and accessibility for authorized users. The primary databases in the private domain would include 1) a project management database with tables and views (stored queries) to help manage project functions, track and manage study databases, and provide information for reports; and 2) a studies database with tables that contained study data, code books, stored samples and information to track researcher requests and provide data in response to researcher queries. The support database in the public domain was intended as the foundation for the public website, storing information about available studies and supporting access to private pages, a hosted user forum, and researcher requests for data based on available fields.

Ultimately, due to time and cost constraints, our system design was modified so that study data were not stored in databases but rather data files were stored in a secure archive/warehouse that was not searchable by external researchers. Instead, researchers had to develop a proposal describing how they would utilize the data and upon approval of their request, the data and supporting documentation were provided to the researcher. This design involved less IT development and offered greater security for the data but required a greater degree of personal assistance from repository staff to help the researcher determine if required data and samples were in fact available. Although this may be viewed as a limitation, it is widely accepted that a high level of personal assistance from a repository is preferential and beneficial [2], not only to the researcher, but also for the long term sustainability of the repository. Repositories providing personalized assistance are sought by investigators, particularly less experienced researchers, and become well known as reliable partners. This level of support is appreciated by the researcher and is a potential source of revenue since additional support can be incrementally billed as part of the data/sample request.

The support database implemented on the public side resembled the original design -- presenting materials that clearly described each NIDDK study included in the archive [or identified for future inclusion]. Materials presented to a public user included (1) a general description of the study, (2) manuals of operations and protocols, (3) data forms used to collect clinical data, (4) descriptions of available data, and (5) listings of study publications.

A web portal served as the interface for electronic information exchange for the NIDDK CR. All of portal's data sharing features were accessible to the public, but only registered users accessed and provided information to the private section of the portal which was governed by role based restrictions. As mentioned above, the clinical study data were archived on a private network accessible only to CR project staff.

The NIDDK CR portal ran on RTI's Oracle Application Server 10g (v. 10.1.2.0.2) server farm and used Oracle technology to manage the information within the repository. The software tools that supported the CR are summarized in Table 2 and the hardware supporting the web portal including the sample database is presented in Table 3.

Study data from the Data Coordinating Centers (DCC) were submitted in SAS and retained in SAS format when archived stored. Requestors seeking alternative formats were provided with alternative formats using the dbCOPY tool [3]. All study documentation except some electronic data capture forms was stored in PDF. Some older data capture forms were delivered as image files. In all cases these were readable via Adobe Acrobat.

Study data were not shared until a request was authorized. At that time the entire contents of the archive of the requested study was sent by a secure FTP process to the requestor site. This meant that the data could be uploaded to the requesters system regardless of the target operating system.

Over time, we enhanced the public system to provide some of the functionality of the original design that was lost during initial implementation. To help users explore the vast amount of data and samples stored in the repository, we developed a set of Public Query Tools (PQT) that allowed public users to explore data elements in both structured and unstructured ways [Figure 1]. The structured searches used parameters to identify studies with resources that could support a new research hypothesis (e.g., types of stored samples, intervention method, and primary outcomes). PQT opened a window to the data for users and was an important enhancement of public data sharing for the repository. Researchers and the lay public were able to learn specific results about the research funded by NIDDK, and in this way PQT served as a valuable public education tool. However, this value came at a high labor cost since study data was stored only in archived data sets [4]. To fuel PQT, select data elements were curated by repository staff and uploaded to a database that supported the PQT functionality. This level of curation required clinical expertise available only through senior repository staff. Thus, maintaining PQT was a costly effort that required significant investment which ultimately has to be weighed against the benefit. There were cost advantages. With researchers able to personally explore the availability of stored samples and link them to specific data elements, the amount of expert labor required for sample request processing was reduced.

Offering mechanisms that make data more available for public inquiry is surely an important function of a data repository. In fact, with the increase in genomic research, sharing of actual study results with participants is increasingly critical [5]. The question becomes one of cost and technological innovation and associated development costs so that data sharing can take place without a high level of content review by CR curation staff. Use of data standards and common data elements during collection should allow for a more automated presentation of results. Such consistency must begin at the design stage and requires that the data repository

be a partner right from the start to streamline processes and reduce the cost of post study data sharing.

2.2. Model growth and adaptation to changing research landscape

Over the course of the development and implementation of the CR, planning and conduct of clinical trials changed with respect to the application of information technology. Clinical trial management systems (CTMS) have become common tools used by data coordinating centers to manage the operational features of clinical trials including designing and annotating the Case Report Form (CRF) and supporting database, data-entry which is frequently web-based, data validation, and medical coding. The use of laboratory measures to track patient outcomes and drug reactions is standard operating procedure and with expanded use of genetic analyses, there is an ever greater reliance on biological samples which means tracking processes and annotations at a level of greater detail. Recently the FDA approved a Phase 1 trial that used remote patient monitoring to collect biometric data in patients' homes and transmit it electronically to the trial database. This technology will provide more data points and reduce the occurrence of missing data [6]. While these changes are on-going and will have impacts on the nature of the data entering the repository in the future, they had minimal impact to the CR to date. In part, this is due to the inherent lag in submission of data to a repository which routinely does not take place until analysis for publication of results. Submission can occur years after actual collection, a fact that challenges resolutions if data discrepancies are found during the submission process.

One of the most significant changes in the conduct of clinical trial studies is in its infancy, incorporation of "omics" data into the fabric of clinical research. Omics aims at the characterization and quantification of pools of biological molecules that translate into the structure, function, and dynamics of an organism or organisms [7]. Fields include genomics, proteomics or metabolomics. As the knowledge related to personalized medicine (tailoring medical decisions, practices, and/or products to the individual patient) increases, treatments based on a patient's biological assays will become a routine feature of clinical studies. Omics signatures that predict disease progression will become a standard part of clinical studies and treatment assignments will be wedded to both the genetic makeup of patients and disease progression. These changes have had an impact on CR functions and will continue to do so in coming years. A number of NIDDK studies became part of the database of genotypes and phenotypes (dbGaP), a platform that was developed to archive and distribute the results of studies that have investigated the interactions of genotype and phenotypes. A number of the CR studies were represented in the dbGaP and notification procedures between the CR and dbGaP were developed to direct requestors from the CR to dbGaP if genotype information was sought and from dbGaP to the CR if the most recent phenotype information was sought. This interaction required development of additional standardized procedures and expanded communication for staff at the CR, NIDDK and sometimes, the DCC.

In other examples, omics data appeared as part of ancillary studies that extended investigation of the disease studied by the parent study investigators. Such studies incorporated immunology, virology, biochemical and genetic biomarkers. As mentioned earlier, reliance on samples has grown in the past decade due to the wider availability of testing. Use of omics data in clinical trials requires carefully annotated sample/tissue archives that reflect sample storage practices and can be linked to outcome information (phenotypes). Annotation of processing and storage methods is of interest because methods that are adequate for DNA preservation may be inadequate for metabolite preservation. Thus sample data is as important to the data archive as clinical data collected during the study. Storage and presentation of omics data presents challenges in staffing, quality control and storage that we were only beginning to address as these data began to appear for inclusion.

2.3. Working with researchers

A central activity of a data repository is effective interaction with researchers, a function that is served by an understanding of the research process and an appreciation of the challenges faced by investigators conducting high quality research. Our operation required working with a variety of researchers: 1) those conducting NIDDK projects that would submit to the CR; 2) those seeking samples and data from the CR, and 3) those at NIDDK involved in specific specialty areas. Researchers reflected experience levels from expert to novice and brought with them differing attitudes towards sending their data to a central repository (discussed below). To address this degree of complexity, we used an array of staff that brought different strengths to the processes. A dedicated communications group maintained ongoing relationships with the DCC responsible for preparation and submission of data, primarily interacting with study managers to keep informed about progress and answer questions about the submission process. Experienced clinical researchers were among repository staff able to communicate with sites about details of study protocols, outcomes and issues relating to shipped samples. Often our statistical experts that performed data assessments communicated with investigators on interpretation of analytic results. Repository biologists worked with requestors to identify appropriate biospecimens to conduct their planned research. Genetic analysts worked with the genetic data repository dbGaP to insure that requestors of NIDDK data from that site were provided with the best phenotype data available for a study, since often, the CR had richer clinical data sets than those available through dbGaP. We found that repository staffing, in many ways, mirrors that of a data coordinating center, providing a central counterpart that can effectively communicate on topic areas (“speak the same language”). A combination of proactive and reactive communication styles are essential to insure that data are submitted to the CR according to timetables provided by NIDDK funded researchers. The large group of researchers involved in the operation of the repository can often overburden the resources of the most responsive and flexible staff.

2.4. Promoting continued use

The essential goal of a repository is to enable additional research using resources already available. This re-use increases the value of the initial investment and acknowledges the importance of the research subject’s willingness to contribute to disease prevention or cure. If we establish a smooth functioning comprehensive archive of research data and samples but no one ever uses them again, what’s the point? What is the role of the repository in promoting the resource and how is the aim best realized? This important aim can only be effectively undertaken when the enterprise is strongly established and thus, during our tenure as the CR, promotion gained priority only in the later years. The repository required time to develop archives and processes to facilitate requests and approvals. NIDDK, as sponsor, played the dominant role in connecting interested researchers with the available resources. CR staff raised the visibility of the CR within the scientific community by attending conferences, although this required support that was not always a budget priority in comparison to other development needs. The primary means of gaining recognition was through the public website that contained a portal to which researchers could subscribe to explore CR contents with the assistance of repository staff. Search tags ensured that the site was a top listing for those searching on an array of key words related to NIDDK research. In the last year of operation, increased effort was devoted to developing links with other bio-banks and organizations in the forefront of health information. In 2013, the CR became one of the Genetic Alliance “Neighborhood”, a position that affords the CR an opportunity to be periodically highlighted in GA electronic mailings. We feel this type of relationship building across the research community is vital to increasing repository use. Clear and detailed information about the archive contents must be accessible for researchers to determine the usefulness for their investigations.

It is also important to connect with the patient and caregiver community, making them aware of the work being done that may lead to disease control or cure. In addition to providing financial support, disease advocacy organizations participate directly in multiple aspects of research, ranging from study design and patient recruitment to data collection and analysis. These organizations believe that scientists and research sponsors should engage them as partners in the conduct of clinical research [8]. Information sharing of this type can lead to greater awareness of the important work supported by government funding and may serve to generate new approaches to health problems. The data repository as an entity in possession of resources, technology and an extensive network of researcher connections is in a prime position to take the lead in promotion, if there are funds available to support this aspect of the operation.

3. Challenges and Solutions

Our years serving as the CR for NIDDK have yielded a number of observations about the difficulties of running a repository, an operation that is by definition dependent on many outside parties whose degree of expertise and efficiency have a direct impact on repository functioning. In this section, we detail five major areas that can be primary obstacles to effective data and sample sharing.

3.1. Repository challenge: lack of provenance

The NIDDK CR houses individual clinical trials as well as networks with dozens of protocols, some research carried out as early as 1982. Study data and biospecimens were collected in various forms, formats and methods. Requirements were developed to standardize the study datasets archived in the Central Repository. For active research, we were able to work with DCCs to conform to these requirements, but data from early studies was problematic. Often the research infrastructure had been disbanded so there were few, if any, persons who could answer questions about the data. Likewise in more current studies, problems arise with data and biospecimens that are submitted without complete documentation. Usual problems that arise include receipt of data from unconsented individuals; clinical data files that lack a subject ID; incomplete submission of data (not all subjects or not all data points included).

Annotation of biospecimens is a recurring and more widespread problem. Often a standard data structure is not determined at the start of the study so that information about biologic samples is missing, limited or collected in an ad hoc manner. This is particularly problematic since biospecimens are shipped in advance of study results and thus, annotations are not scrutinized until the study data is submitted sometimes many years later. Details about sample processing and preservation that are critical to downstream testing are often overlooked in the collection/storage stage since it may not be relevant to the protocol being conducted. Consent information that guides how biospecimens are allowed to be used is usually not incorporated into the annotation. To comply with confidentiality, biospecimen data is often submitted with no way of linking the sample to the clinical data (discussed below). All of these omissions constitute a lack of provenance for a sample which can create uncertainty about its quality and appropriateness for the planned research effort.

3.2. Repository challenge: lack of linkage

Clinical studies typically use one set of subject IDs for internal study purposes, and -- as a privacy precaution -- create “masked” IDs when depositing data with the Repository. Use of a masked ID helps insure that someone who discovers a participant’s subject ID during the study could not easily identify that individual’s shared data. While DCCs maintain “linkage files” identifying which study biospecimen IDs belong to which study subject IDs, the

shared data needs an additional linkage file that allows these biospecimen IDs to be linked to the "masked" IDs employed when the data are deposited with the Repository.

Early in the operation of the repository we discovered that some DCCs did not include such linkage files with the study documentation when they submitted biospecimens and data to the archive. Requestors would be unable to link their biospecimens with the phenotype (clinical) data from the study. The Repository PI and staff undertook a campaign to remind extant and new biospecimen depositors of the crucial need for accurate and well maintained linkage files to be deposited along with their biospecimens. This problem highlights the importance of involving repository staff early in the process so that they can work toward linking biospecimens deposited in the repository to subject data. Often this was a matter of allaying concerns about a breach of subject confidentiality.

3.3. Repository challenge: data ownership

NIH rules on data sharing are well defined and publically documented. Although timelines by institute may vary slightly, data is generally expected to be ready for sharing within two years of study completion or no later than the acceptance for publication of the main findings from the final dataset. If data from large epidemiologic or longitudinal studies are collected over several discrete time periods or waves, data should be released in waves as data become available or main findings from waves of the data are published [9]. Data sharing plans are now a standard required component of proposal submissions. However, since the time period is often after the conclusion of the funding period, it becomes challenging to insure that an investigator complies with the proposed plan. For this reason, acquiring study data for sharing sometimes requires a proactive approach by the data repository. To obtain the data in a timely manner in the proper format requires a collaborative relationship of mutual respect between the repository, the DCC and principal investigators. The primary concern of any investigator is data ownership and maintaining control of the intellectual property of research findings. Working through this sensitive ground requires an understanding of this perspective and recognition that, although data sharing is required, provision of data to a repository is voluntary. The NIDDK CR sometimes encountered protracted delays in receiving data or obtaining the linkages needed to share the data with requestors. Particularly with large longitudinal studies that conduct enrollment over many years, data are typically stored at the DCC on an open-ended basis since analysis and publication is ongoing. In fact, some long standing research enterprises prefer to act as their own repository, spending added effort to coordinate requests from outside researchers. This compliance with data sharing is perfectly responsive to the NIH mandate but tends to dilute the power of a central repository.

3.4. Repository challenge: consent

Study consent documents are typically generated by methods that make them awkward to automate. They may vary by study, clinical site, study subpopulation, and time interval, and different restrictions may apply to different uses of the data or biospecimens (e.g., only for use in diabetes research). These consent parameters are nonetheless crucial to Repository operations since they govern the conditions under which data and biospecimens from a study may be released. The critical nature of consent information raises the question of why a standard approach has not been promoted more forcefully across the research community, as has been done for other data elements that are common across all research. Perhaps privacy concerns that have influenced research over the past decade have curtailed efforts at standardization. Consent is usually looked upon as a very individualized process, one of the only domains where the subject retains ultimate control. This perspective should not negate an effort to compile consent data in a standardized format for more efficient sample and data sharing. In fact, standardized consent data would better ensure handling of resources in

compliance with individual subject preferences. However, until recently, consent forms have been isolated from other data collection instrumentation used in research where efforts towards standardization have gained increased momentum. The Clinical Data Interchange Standards Consortium (CDISC) was formed to establish standards to support the acquisition, exchange, submission and archive of clinical research data and metadata. The CDISC mission is to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas [10] of healthcare; however, to date there lacks effort to catalogue consent parameters that represent all aspects of subject preferences with regard to sharing data and biologic samples for continued research.

3.5. Repository challenge: sustainability

Does the importance of data sharing necessitate a centralized data archive and if so, must this archive be maintained by the institute that initially provides funding for the research? We will discuss the evolving nature of repositories in the next section but for now will look at the challenge of sustainability for a central repository such as the CR, rather than a biorepository designed to operate commercially. There are few ways to generate revenue to maintain repository operations. If the repository is directly aligned with an agency, the expectation is that the agency bears a major responsibility for maintenance through direct support grants. These costs can be supplemented by charging requestors for services performed by the repository, costs that may in turn be built into grant requests submitted by the researcher to fund the planned project. The most straightforward way of raising revenue is through charges for biologic samples and corresponding data. In fact, high quality clinical data adds substantial value and having a detailed picture of how samples were handled since collection is equally important, as discussed in earlier sections. The emphasis on strict quality control is a key characteristic of an effective repository operation. Usually there are differences in the rates charged for commercial and non-profit requestors. These fixed rates encompass the labor required to prepare the requested materials and, to a small degree, maintain the repository infrastructure. However, the financial gains from these charges in no way offset the cost of running a repository and thus, other services such as statistical expertise would need to be required. As the contractor for the CR, we did not attempt to achieve sustainability but we were able to raise revenue on a modest scale.

4. The Path Ahead

The bio-banking industry will likely continue to become more globally centralized for studying specific genetic diseases and monitoring the health of our environment. This effort will gain strength from the more recent push by patients to form networks that support research into specific diseases. A prominent example of this trend is the Rare Disease Clinical Research Network. The RDCRN Contact Registry operates in Australia and seeks to enroll subjects with rare diseases (they can register themselves) in order to be contacted in the future about clinical research opportunities and updates on the progress of the research projects. The contact registry is anonymous and free of charge. The RDCRN is designed to enable multiple-level access by a range of user groups within a region or across regional/country borders in a secure and private manner. This system was developed for patients with clinical and genetic data in geographical disparate locations. The system addresses issues of multiple-level access by key stakeholders, security and privacy. Networks such as the RDCRN cut across borders and funders to join governments and private industry in the research effort.

The role of the biobank (defined as sample or data repositories) may expand in coming years as the issue of return of research results to individual subjects gains greater traction in the advent of omics testing. Results based on downstream use of samples may produce results of

consequence to the health of the subject. A two year project funded by NIH resulted in a recommendation that biobanks assume significant responsibility for considering the return of research results whose samples and data were used in primary or secondary genomic research if the findings have potential health, reproductive or personal importance to the contributor [11].

The trend toward personalized medicine will continue to impact. To date, these impacts include higher standards for bio-repositories and the wider development of molecular data repositories ('atlas'). Translational research, the 'bench-to-bedside' approach to life-science research where what is discovered in the laboratory is translated into practical applications, will increase the occurrence of personalized medicine and, to a certain degree, personalized research. In a clinical trial, omics data could be used to guide assignment to the most appropriate treatment group. Genomic technologies are a key force involved in translational research and have a vital role in determining how patients are treated. Individuals will be prescribed drugs based on their genetic profile in combination with traditional clinical signs and symptoms. Oncology has already made strides in this area by using samples from an individual's tumor to be tested for biomarkers (indicators of a biological state) associated with a particular cancer treatment option which in many cases depends on the type of cancerous cell. The Cancer Genome Atlas (TCGA) program with its supporting repository had a goal to create an atlas of genetic changes that manifested as a cell became affected by cancer. TGCA characterized tumor types which required a massive collection of cancer samples. Dozens of bio-repositories in the US assured the effort that at least 500 samples of each required cancer type could be easily provided. Early in the project, it became clear that many specimens were unfit for analysis due to the lack of sample storage standards. To achieve their mission of mapping the genetic changes in cancer, the investigators had to develop standards for storing, and annotating samples which led to the Office of Biorepositories and Biospecimen Research (OBBR) publishing its first guidelines for the industry in 2006. Biobanking facilities will need to meet these newer standards and this may result in fewer facilities that are larger in size with greater resources which should reduce sample storage failure rates. Such facilities employ modern sample processing technologies that use robotic liquid handling systems and utilize LIMSs that track samples and maintain chain of custody from collection to storage to extraction [12]. That said, we do acknowledge that genetic testing methods are becoming more sophisticated and are able to make better use of degraded samples that have heretofore been of little value (e.g., Molecular Inversion Probe). There is a dynamic relationship between emerging technologies and the infrastructure needed to support future research that requires the ability of organizations providing support to remain flexible even while following established standards.

Another continuing research trend involves national and multinational efforts to launch population-based biobanks and longitudinal cohort studies for use in large-scale genetic research. These biobanks provide the materials to support genetic and epidemiological studies that examine health determinants. Including clinical data about the subject's health status (phenotypes), an understanding of the complex interactions between genes and environment is possible [13]. However, conduct of health research can be influenced greatly by the structure of the health care delivery system in which the research is conducted. A health care system can either facilitate or hinder the research. Countries with government run health care and a national database of members can provide a ready source of potential research subjects for large-scale genetic research. The UK Biobank is a research effort whose strategy relies on centralized conduct of most aspects of a trial but handles recruiting through multiple temporary assessment centers. Between 2007 and 2010, the UK Biobank completed recruitment and examination of 503,000 participants. Embedding recruitment in a structure that facilitates outcome determination, utilizes comprehensive information technology, automated biospecimen processing, ensuring broad consent and establishing

essentially autonomous leadership with appropriate oversight were all critical to its success [14]. In the US, major health care providers are beginning to establish similar models (e.g., Kaiser Research Program on Genes, Environment and Health, Vanderbilt University BioVU, etc.) that enable recruitment of eligible subjects from an established patient registry and extraction of relevant clinical measures from the electronic medical record (EMR). This model, which reduces costs, could radically change the method for conducting medical research and will likely become an important model in the future. Similarly, the National Institutes of Health, supported by the Common Fund, is investigating the feasibility of conducting "simplified" clinical trials within a similar model where research takes place in the context of health care delivery. The Health Care Systems Research Collaboratory (HCSRC) is intended to improve the way clinical trials are conducted by creating a new infrastructure for collaborative research. The ultimate goal is to ensure that healthcare providers and patients can make decisions based on the best available clinical evidence [15].

It is uncertain where the role of a repository fits into these newer research models. Will there be a greater profusion of repositories or an effort to create a national database that would become a resource containing research data across all types of funders?

In closing, we wish to note that RTI's tenure as the NIDDK Central Repository ended in August 2013 after an unsuccessful re-competition bid. However, the CR lives on and NIDDK will continue to offer high quality data and biologic samples to the wider research community [16]. RTI is pleased to have been a founding partner of the repository with the rich experience of managing its growth for many years. Over time, data and biorepositories have evolved from a simple location-management focus to one that supports a much wider and often institution-wide imperative, such as translational research. For now, the repository remains an essential component in medical research [17].

Acknowledgments

This work was supported by National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), Bethesda, MD 29892, USA, contract number HHSN267200800016C

References

1. National Institutes of Health, Trans-NIH BioMedical Informatics Coordinating Committee (BMIC). [Accessed September 18, 2013.] NIH Data Sharing Repositories. http://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html
2. Horn, L.; Bledsoe, M.; Sexton, K. Marketing Your Biobank Collection Effectively. Oct 23. 2012 Retrieved from http://www.youtube.com/watch?v=5Et4IIUqBlY
3. Heinemann, Florian; Sven Kolber, GbR. [Accessed November 13, 2013.] dbCOPY Database Tool. http://www.dbcopy.com
4. Pan, H.; Ardini, MA., et al. [Accessed September 18, 2013] "What's in the NIDDK CDR?"-- public query tools for the NIDDK central data repository. Database (Oxford). 2013 Feb. 2013:bas058 Print 2013. http://www.ncbi.nlm.nih.gov/pubmed/23396299
5. Wolf SM JD, Crock BN JD, et al. Managing Incidental Findings and Research Results in Genomic Research Involving Biobanks and Archived Data Sets. Genet Med. 14:361–384. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3597341/. 10.1038/gim.2012.23 [PubMed: 22436882]
6. Baum, Stephanie. [Accessed September 18, 2013.] FDA gives go ahead for Phase 2 trial using remote monitoring and crowd sourcing. Dec 18. 2012 http://medcitynews.com/2012/12/fda-gives-go-ahead-for-phase-2-trial-using-remote-monitoring-for-multiple-sclerosis-drug/#ixzz2fAgQhTfh
7. [Accessed September 18, 2013.] "Omics," from Wikipedia. http://en.wikipedia.org/wiki/Omics
8. Landy DC, Brinich MA, Colten ME, Horn EJ, Terry SF, Sharp RR. How disease advocacy organizations participate in clinical research: a survey of genetic organizations. Genet Med. 2012; 14(2):223–8. Epub 2012/01/21. 10.1038/gim.0b013e3182310ba0 [PubMed: 22261756]

9. [Accessed September 18, 2013.] NIH Data Sharing Policy and Implementation Guidance, Updated. Mar 5. 2003 http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm
10. Mission & Principles. Clinical Data Interchange Standards Consortium; http://www.cdisc.org/mission-and-principles [Accessed September 18, 2013]
11. Wolf SM JD, Crock BN JD, Van Ness B PhD, et al. Managing incidental findings and research results in genomic research involving biobanks and archived data sets. Genet Med. 2012 Apr; 14(4):361–84. http://www.nature.com/gim/journal/v14/n4/full/gim201223a.html. 10.1038/gim.2012.23 [PubMed: 22436882]
12. Frey, M.; Summers, A.; Napier, M. [Accessed September 18, 2013.] The Future of Biobanking. http://www.laboratoryfocus.ca/the-future-of-biobanking/
13. Cami, J.; Bertranpetit, J. [Accessed September 18, 2013.] The Promising Future of Biobanks: Building a Global Perspective. http://www.jcami.eu/PDF/biobanks.pdf
14. Manolio TA, Weis BK, Cowie CC, et al. New Models for Large Prospective Studies: Is there are Better Way? Am J Epidedemiol. 2012; 175(9):859–866. http://aje.oxfordjournals.org/content/early/2012/03/11/aje.kwr453.full.
15. NIH Collaboratory. Health Care Systems Research Collaboratory; https://www.nihcollaboratory.org/about-us/Pages/default.aspx [Accessed September 18, 2013]
16. NIDDK Central Repository. the National Institute of Diabetes and Digestive and Kidney Diseases; https://www.niddkrepository.org/home/ [Accessed September 18, 2013]
17. Blackman, G. [Accessed September 18, 2013.] Biobanking: Saving for the Future. http://www.scientific-computing.com/features/feature.php?feature_id=232

Highlights

- Effective operation and maintenance of the NIDDK Central Data Repository required the system to be flexible and dynamic while at the same time compliant with established data standards.
- We describe some difficulties of managing a large repository, an operation that is by definition dependent on many outside parties whose degree of expertise and efficiency have a direct impact on repository functioning.
- The bio-banking industry will likely continue to become more globally centralized for studying specific genetic diseases and monitoring the health of our environment.
- The dynamic relationship between emerging technologies and the existing infrastructure will be needed to support future research that requires supporting organizations to remain flexible even while following established standards.

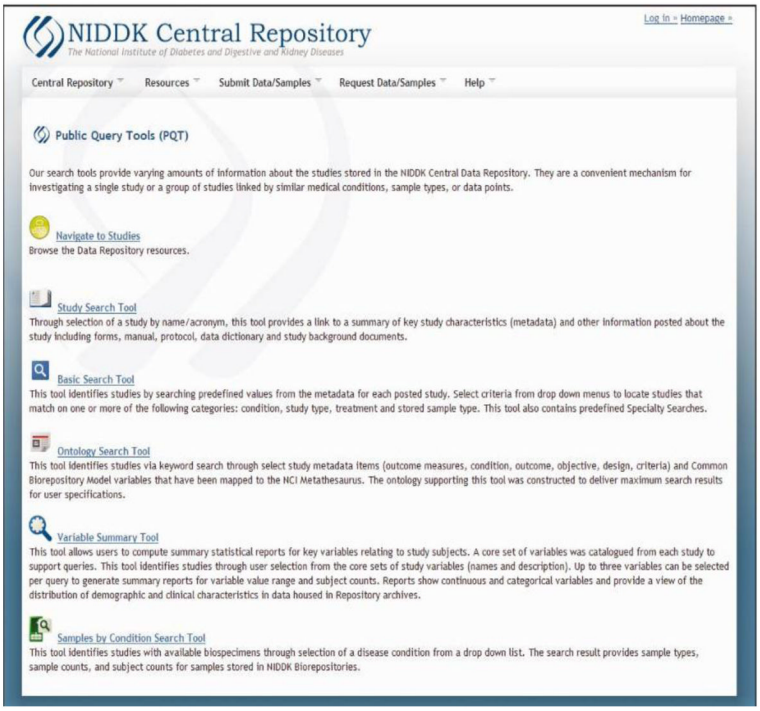

Figure 1.
Public Query Tools allowed public users to explore data elements in various ways.

Table 1

Primary Diseases Represented in CR

Kidney Disease	17
Liver Disease	6
Diabetes (adult)	4
Diabetes (juvenile)	8
Urologic Disease	10
Interstitial Cystitis/Prostatitis	8

Table 2

Software platforms supporting the design of the CR

Software	Function
SAS	Store clinical data files
PDF (sometimes Word)	Store all documentation
Oracle 10g Application server, version number 10.1.2.0.2	Hosts the Web portal infrastructure
Oracle 11g database Enterprise Edition, version number is 10.2.0.3.0, Microsoft Windows	Hosts relational components of CR
J2EE within a Struts Framework	Provides functionality to the CR Infrastructure

Table 3

Hardware Description Required to Support the CR

Hardware	Operating System	Description
Dell PowerEdge 2850 processor	Red Hat Enterprise Linux ES Release 4	3 server farms are used to host the Oracle 11g application server
Dell PowerEdge R610	Red Hat Enterprise Linux ES Release 5.9	Hosts the Oracle 11g database