Clinical Biochemistry 47 (2014) 252–257

Contents lists available at ScienceDirect

Clinical Biochemistry

journal homepage: www.elsevier.com/locate/clinbiochem

Sample and data sharing: Observations from a central data repository

Mary-Anne Ardini, Huaqin Pan, Ying Qin, Philip C. Cooley *

RTI International, PO Box 12194, Research Triangle Park, NC 27709, USA

ARTICLE INFO

Article history:
Received 20 September 2013
Received in revised form 13 November 2013
Accepted 15 November 2013
Available online 26 November 2013

Keywords:
Clinical data repository
Biobank
NIDDK

ABSTRACT

Objectives: From 2003 to 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.

Design and methods: Effective maintenance required the system to be flexible and dynamic while at the same time compliant with established data standards.

Results: 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.

Conclusion: 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.

Introduction

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 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.

Development and implementation

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 study

* Corresponding author. Fax: +1 919 316 3539.
E-mail address: pcc@rti.org (P.C. Cooley).

0009-9120/$ – see front matter © 2013 The Canadian Society of Clinical Chemists. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.clinbiochem.2013.11.014

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 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

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 (DCCs) 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.

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

Study data were not shared until a request was authorized. At that time the entire content 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 requester's 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 [Fig. 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 datasets [4]. To fuel PQT, selected 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.

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