The Agony and the Ecstasy: A Tale of Repository Data Analysts

Norma Pugh, Sylvia Tan, Charles Turner, Sue Rogers
RTI International
RTP, NC

Background

In 2003, The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) at the National Institutes of Health (NIH), initiated the 'NIDDK Central Repository' (see www.niddkrepository.org). Consisting of three branches, Data, Biospecimen and Genetic, the NIDDK Central Repository houses the phenotypic data and biological specimens from NIDDK-funded research. The repository aims to facilitate the testing of new hypotheses by making these data and specimens available to the broader scientific community. For example, researchers have the opportunity to conduct informative genetic analyses using well-curated phenotypic data and to pool data across several studies in order to increase statistical power, all without the collection of new data or biospecimens. The management team of the repository is multidisciplinary, and includes consent specialists, data curators, database programmers, and web designers. The repository also includes a small team of data analysts, who face a unique set of challenges: performing statistical analyses on submitted clinical datasets to ensure dataset integrity; speaking for study data coordinating centers when responding to user inquiries about the data; and harmonizing datasets across protocols to facilitate larger-scale studies. Here, we discuss these unique challenges and the tremendous value we strive to provide as repository data analysts.

Challenges and Agonies

Introduction

The repository tracks data and samples from more than a hundred NIDDK-funded studies and houses more than 4 million biospecimens. Each study may have multiple protocols. In 2011, there were 75 data requests and 163 sample requests. Data analysts worked extensively with the institutions that house DNA and other samples, as a total of 1,728,816 samples were loaded into the repository database.

Data Curation

Using a checklist that the repository provides to guide study submission, each study coordinator submits a standard set of study materials to the repository. This includes, at a minimum: raw data, data forms, the protocol, manual of operations and data dictionaries. A repository analyst will then use these materials to "curate" the data, that is to say, thoroughly check the data for completeness, HIPAA compliance and consent for the repository to receive the data. The challenge is that each study data coordinating center (DCC) manages its data slightly differently. For example, some DCCs are able to provide analysis datasets to the repository, while others are uncomfortable providing that level of detail. We choose a middle ground in terms of working with existing DCC standards while requesting the necessary materials for future data use. HIPAA compliance and documentation of proper consents are essential.

Dataset Integrity Checks

The repository data analyst team is a core component of the NIDDK Central Repository. Without ensuring the integrity of the data housed at the repository, and assisting researchers with their use of the data, the repository can't fulfill its role of providing data that are complete, accurate and comply with HIPAA regulations. Our initial challenge, ensuring the integrity of the data, is accomplished by performing a rigorous review of the data, which we call the Dataset Integrity Check (DSIC). The intent

of the DSIC is to provide confidence that the data distributed by the NIDDK repository are a true copy of the study data, ***not*** to assess the integrity of the statistical analyses reported by study investigators.

The process begins with a review of study publications. The analyst chooses an article for the purpose of replicating the results using the repository's copy of the study data. Preference is given to articles that report more central findings of the research, report a wider array of study data, and are published in more prestigious scientific journals. Occasionally, it will be necessary to replicate more than one article. Some studies, for example, may publish one paper on baseline results and another paper based on final results. In other instances, a study may have multiple primary outcomes and may choose to publish a paper for each outcome.

After choosing a publication, the analyst will identify statistical analyses to be replicated. Selection of analyses for replication will be guided by the importance of the results and the availability of analysis data sets and/or necessary variables in the repository data set. As with all statistical analyses of complex datasets, complete replication of a set of statistical results should not be expected on a first (or second) exercise in secondary analysis. This occurs for a number of reasons including differences in the handling of missing data, restrictions on cases included in samples for a particular analysis, software coding used to define complex variables, etc. Our experience suggests that most discrepancies can ordinarily be resolved by consultation with the study DCC, however this process is labor-intensive for both DCC and repository staff. Thus, it is not our policy to resolve every discrepancy that is observed in an integrity check. Specifically, we do not attempt to resolve minor or inconsequential discrepancies with published results or discrepancies that involve complex analyses, *unless NIDDK repository staff suspect that the observed discrepancy suggests that the dataset may have been corrupted in storage, transmission, or processing*. We do, however, document in footnotes to the integrity check those instances in which our secondary analyses produced results that were not fully consistent with those reported in the target publication.

If the replication is deemed successful, the analyst will prepare a narrative report, tables and/or figures, programming code and output documenting the equivalence of results. If the replication is unsuccessful, the analyst will consult with other analysts in order to confirm that appropriate methods were used for the replication. If there is confirmation that the published results cannot be replicated, the analyst will consult with the principal investigator to determine an appropriate course of action. This action may include additional analytic work to determine the cause of the non-replication, and/or requests for clarifications or assistance from the study DCC.

The final DSIC is distributed to the DCC and the NIDDK Project Officer for review and approval. If approved, the study data are made available for use by approved researchers. If the DSIC is not approved, discussions are held among interested parties, new/additional specifications are generated and appropriate action taken. The DCC may need to send new and/or updated data to the repository, or a new DSIC may need to be developed. Study data may NOT be made available on the NIDDK data repository website until discrepancies identified in the DSIC have been resolved with the DCC and to the satisfaction of the NIDDK Project Officer.

The value of the DSIC is evident by issues we commonly uncover. Data Coordinating Centers may send too much data (data we were not consented to receive) or too little data (a clinic may be inadvertently left out). Additionally, the DCC documentation may be inconsistent or unclear. Using the documentation to replicate results identifies this information. For example, the mislabeling of gender in one published article, led to the mistaken conclusion that atrial fibrillation is more common in women than in men with chronic renal insufficiency. The reverse is actually true. This issue was not uncovered until the DSIC was performed by a repository data analyst, well after publication of the results.

User Inquiries

An analyst's work is not done once study data have been released for interested researchers. The challenge of responding to user inquiries consumes much of an analyst's time. Each request for data and/or samples must include a high-level abstract of proposed work, including study objectives, background, importance and design of research. Of course, the requestor must also obtain clearance from their Institutional Review Board or Human Subjects Protection Panel before data and/or samples will be released [2].

The NIDDK Central Repository website contains a wealth of resources: a thoroughly detailed description of each study, data forms, study protocols and manual of operations, publications, dataset integrity checks, and other pertinent information. In addition, public query tools (PQT) have recently been added to the website. These tools provide a set of search mechanisms, allowing users to find study information, without prior knowledge of each individual study. PQT are of major value to the repository and are discussed in more detail in the Public Query Tools section, below.

These varied tools, however, do not completely eliminate questions for the analyst. As the person who has used the study data to conduct complex statistical analyses, the analyst is the obvious choice to answer questions related to using the data. Some researchers are more comfortable interacting with the analyst directly, in order to refine and finalize their proposed statement of work. The data analyst is charged with reviewing a requestor's needs to determine what study data will be the most useful. This includes determining if variables of interest were collected and if there are enough observations to adequately power the researcher's study design.

User inquiries regarding biological samples are also common. Researchers need to know if there are adequate numbers of the right types of samples. The data analyst can determine the number of available samples that meet particular subject profiles and verify the numbers needed. As most biological samples are non-renewable, these questions take on particular importance. The repository must ensure that samples are released fairly and prudently.

As an example of a user inquiry, a researcher wanted to secure DNA samples in order to determine the effect of specific diabetes-related genotypes on the ability of oral insulin to induce remission. The importance of the research lies in the potential ability to use genotype to predict response to therapy. The researcher, working with a member of the data analyst team, was able to identify a cohort of patients in the repository who had a high titre of insulin autoantibodies and, as it turned out, a significant treatment effect from oral insulin in preventing development of diabetes.

Harmonization of Datasets

The linking, or harmonization, of datasets across studies is, obviously, a most useful practice. The NIH has, in fact, mandated the sharing of research data across all institutes. Persons particularly interested in data harmonization may include researchers interested in genome-wide association studies. The public query tools that now exist are essential to dataset harmonization. However, these tools require the identification and extraction of the data elements that support possible searches. This is a labor intensive exercise, as the data elements are extracted manually and uploaded to the database. In order to make the search tools useful, there must be accurate, up-to-date alignment to the data that feed the search results.

The PQT were designed with three types of end users in mind; casual users, study-specific users, and disease domain-specific users. While casual and study-specific user needs are straightforward, disease domain-specific user needs are trickier, as they need to harmonize variables from multiple studies in order to examine the underlying and undiscovered properties related to a disease and its treatment.

Members of the data analyst team have intimate knowledge of the data housed in the repository and are skilled at determining which studies collect similar types of data and how these studies are best linked. The combination of the public query tools and the support of a data analyst, is often the best and most efficient way to facilitate the harmonization of datasets.

Examples of Challenges

An example of a challenge we faced, concerned an article published about chronic kidney disease and atrial fibrillation. The label for men and women was reversed. This seemingly simple error led to potential clinical consequences when the authors concluded that atrial fibrillation is more common in women than in men with chronic renal insufficiency. The reverse is actually true. This error wasn't discovered until a repository data analyst performed the dataset integrity check, well after publication.

Another common issue is the receipt of data we were not intended to receive. In other words, data for which the study participant did not give consent for outside researchers to receive.

Rewards and Ecstasies

Value of the Repository

Despite these complex challenges, the outcomes of our efforts make the job satisfying. The value of the repository cannot be overstated. Data collection is both time-consuming and expensive. These are primary reasons for the data-sharing mandate and why repositories are becoming more and more prevalent. Clinical data linked to biological samples, extend the usefulness and value of a study far beyond the original hypothesis of interest. The repository allows researchers to look at diseases, their treatments and outcomes from new and varied perspectives.

As the stature of the repository has grown over time, there are an increased number of requests for study data and/or samples. As this trend continues, more research is conducted and understanding of diabetes and digestive and kidney related-disease processes grows. Thus, the value of the repository grows.

Publications using repository data continue to increase. To date, there have been more than 70 publications using repository data and/or samples. Additionally, pilot results can both motivate new studies and identify unpromising lines of research. Study results often generate new scientific debates which can be quite productive and is yet another positive benefit of the repository. Finally, study results can provide supporting evidence for previous study conclusions.

The value of the repository also extends to the formulation and testing of new statistical methods. By harmonizing datasets and increasing statistical power, researchers are able to use repository data to develop novel statistical tests and methods.

Public Query Tools (PQT)

PQT are funded by the NIH American Recovery and Reinvestment Act of 2009. In the days before PQT, researchers had to submit a proposed investigation statement and gain NIDDK approval *before* accessing any data housed at the repository. This process sometimes led to frustration for researchers who needed a more thorough understanding of the repository offerings, in order to refine and finalize their proposed statement of work. PQT allow researchers to be more certain that the repository can provide the necessary types of data and samples, before completing the detailed paperwork needed to gain NIDDK approval to access data.

There are 5 tools in the PQT suite, defined as follows. The Study Search Tool allows users to select a study by name. The Basic Search Tool allows users to select studies that match predefined criteria, such as disease, study type, or the availability of genetic data. The Ontology Search Tool makes use of the U.S. National Cancer Institute (NCI) Metathesaurus, to which repository study variables have been mapped. The NCI-Metathesaurus is a web-based terminology browser created through a collaborative effort of the NCI Center for Bioinformatics and the NCI Office of Communications. For example, using this tool for a search of "renal disease" will return NIDDK-funded studies of "kidney disease", "disordered renal", "disordered kidney" and the like. The Variable Summary Tool allows users to compute summary statistics for study variables. Summaries are currently limited to ranges (continuous variables) and frequencies (categorical variables). However, cross tabulations can be generated if the user wishes to look at more than one variable. The user may also look at a variable across studies. Finally, the Sample by Condition Search Tool allows the user to search based on a condition (e.g., Cirrhosis) selected from a drop down list. This tool displays both studies and types of biological samples available.

Repository team members have found several unexpected benefits of the PQT. The tools provide additional support for the monitoring of the registry process and for conducting quality control spot checks on the registry database. Further, we are able to conduct queries in order to gain a better understanding of the nature of the research data being funded by NIDDK. This type of information is of interest to both researchers and NIDDK staff, which can compare these trends against their strategic plans. Finally, we are easily able to check any number of study details, for example, sample tallies for types of biospecimens and subject totals. Thus, the PQT have proved immensely useful to us, at the repository.

Examples of Rewards

The range of the research is wide; including biochemical, clinical, statistical and genetic. Select articles include, The Role of Blood Pressure Variability in the Development of Nephropathy in Type 1 Diabetes, Genome-wide Association Data Identifies Novel Loci for Type 1 Diabetes, and Evolution of Causes and Risk Factors for Mortality Post-liver Transplant. Among the significant findings, it was found that mean blood glucose is a better predictor of cardiovascular risk than glycated hemoglobin (Hba1c). Also, in addition to Hba1c, mean blood glucose and within-day blood glucose variability are associated with risk of hypoglycemia.

Users of repository data have suggested alternative statistical measures rather than what has been published, sparking lively scientific debates and/or confirming previous study conclusions. One user developed and tested a novel statistical method that uses prior information to improve power in genome-wide association analyses.

Additional rewards have come, unexpectedly, from the PQT suite. These tools provide additional support for the monitoring of the study registry process and for conducting quality control spot checks on the registry database. We are able to conduct queries to monitor trends in the types of research being funded by NIDDK. This information is of interest to both researchers and the NIDDK, which can compare these trends against their strategic plans. Finally, we are able to easily check study details (e.g., sample tallies for types of biospecimens, subject totals).

Conclusion

Within the grand scheme of the NIDDK Central Repository, data analysts are vital team members. We ensure that the repository datasets are accurate & complete copies of study datasets. We are a point of contact for outside researchers and we contribute the data elements used in the all important public query tools. The repository goal to increase the impact of NIDDK-funded studies is fulfilled, in great part, due to the work of the data analyst team.

For more information, please visit the repository website at http://www.niddkrepository.org. The website contains a wealth of information, including all NIDDK-funded studies with available data and/or samples. The public query search tools may also be accessed from the website.

References

1. The NIDDK Central Repository at 8 years--ambition, revision, use and impact. Charles F. Turner, Huaqin Pan, Gregg Silk, Mary-Anne Ardini, Vesselina Bakalov, Stephanie Bryant, Susanna Cantor, Kungyen Chang, Michael DeLatte, Paul Eggers, Laxminarayana Ganapathi, Sujatha Lakshmikanthan, Joshua Levy, Sheping Li, Joseph Pratt, Norma Pugh, Ying Qin, Rebekah Rasooly, Helen Ray, Amanda Flynn Riley, Susan M. Rogers, Charlotte Scheper, Sylvia Tan, Stacie White, Philip C. Cooley; Database (Oxford). 2011 Sep 29;2011:bar043. Print 2011.
2. National Institutes of Health. Final NIH statement on sharing research data. Notice: NOT-OD-03-032, released date: February 26, 2003. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html (2 March 2010, date last accessed).
3. http://www.niddkrepository.org (NIDDK Central Repository)
4. PQT paper
5. http://en.wikipedia.org/wiki/NCI_Metathesaurus (Wikipedia, NCI Metathesaurus)
6. Kilpatrick, E.S., Rigby, A.S. and Atkin, S.L. (2008) Mean blood glucose compared with HbA1c in the prediction of cardiovascular disease in patients with type 1 diabetes. *Diabetologia,* **51**, 365-371.
7. Kilpatrick, E.S., Rigby, A.S., Goode, K. *et al.* (2007) Relating mean blood glucose and glucose variability to the risk of multiple episodes of hypoglycaemia in type 1 diabetes. *Diabetologia,* **50**, 2553-2561.
8. The Role of Blood Pressure Variability in the Development of Nephropathy in Type 1 Diabetes. ES Kilpatrick, AS Rigby, SL Atkin. Diabetes Care 33:2442 - 2447, 2010
9. Follow-up analysis of genome-wide association data identifies novel loci for type 1 diabetes. Grant SF, Qu HQ, Bradfield JP, Marchand L, Kim CE, Glessner JT, Grabs R, Taback SP, Frackelton EC, Eckert AW, Annaiah K, Lawson ML, Otieno FG, Santa E, Shaner JL, Smith RM, Skraban R, Imielinski M, Chiavacci RM, Grundmeier RW, Stanley CA, Kirsch SE, Waggott D, Paterson AD, Monos DS; DCCT/EDIC Research Group, Polychronakos C, Hakonarson H. Diabetes. 2009 Jan;58(1):290-5. Epub 2008 Oct 7.
10. Evolution of causes and risk factors for mortality post-liver transplant: results of the NIDDK long-term follow-up study. Watt KD, Pedersen RA, Kremers WK, Heimbach JK, Charlton MR. Am J Transplant. 2010 June; 10(6):1420-7.