Published in final edited form as:

Prog Neurobiol. 2017 May ; 152: 200–212. doi:10.1016/j.pneurobio.2016.03.008.

Can Data Repositories Help Find Effective Treatments for Complex Diseases?

Gregory K. Farber

Office of Technology Development and Coordination, National Institute of Mental Health, National Institutes of Health, 6001 Executive Boulevard, Room 7162, Rockville, MD 20892-9640, USA

Abstract

There are many challenges to developing treatments for complex diseases. This review explores the question of whether it is possible to imagine a data repository that would increase the pace of understanding complex diseases sufficiently well to facilitate the development of effective treatments. First, consideration is given to the amount of data that might be needed for such a data repository and whether the existing data storage infrastructure is enough. Several successful data repositories are then examined to see if they have common characteristics. An area of science where unsuccessful attempts to develop a data infrastructure is then described to see what lessons could be learned for a data repository devoted to complex disease. Then, a variety of issues related to sharing data are discussed. In some of these areas, it is reasonably clear how to move forward. In other areas, there are significant open questions that need to be addressed by all data repositories. Using that baseline information, the question of whether data archives can be effective in understanding a complex disease is explored. The major goal of such a data archive is likely to be identifying biomarkers that define sub-populations of the disease.

Introduction

Over the past few years, big data has been touted as a way to advance many areas of biomedical science (Margolis, et al., 2014). This review article will focus on trying to understand whether data repositories can help in the search for treatments for complex diseases. Complex diseases are defined as disorders that do not have a single deeply penetrant genetic cause or a single infectious agent. Generally, these diseases are thought to have multiple genetic and/or environmental contributions. Almost all of those diagnosed with a mental illness have a complex disease.

It is very common for complex diseases to be composed of multiple subpopulations. Each of these subgroups is defined by a unique set of underlying biological causes. However, the subgroups often share common symptoms. The symptoms allow a diagnosis but do not reflect the underlying biological causes and so do not allow us to understand which subpopulation a patient belongs to. Type 1 and Type 2 diabetes are a good example. In type 1

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

diabetes, the body does not produce insulin. In type 2 diabetes, there is some problem with the way the body uses insulin. The biological causes of these two sub-categories of diabetes are quite different, yet those with either type of diabetes share the symptom of elevated blood sugar. The useful treatment options for those with type 1 diabetes are quite different from the treatment options for type 2 diabetes. In this case, testing for the presence of the C-peptide of insulin could provide an effective biomarker to differentiate those individuals who are producing insulin from those who are not (VanBuecken and Greenbaum, 2014). Useful biomarkers help differentiate the subgroups of a disease so that effective treatments can be discovered for each subpopulation.

Finding useful biomarkers for complex diseases is very difficult because of our limited understanding of the number of subpopulations for the disease as well as our limited understanding of the underlying genomic and environmental factors that have caused the disease. The purpose of this paper is to explore whether a data repository can contribute to the discovery of a biomarker that would be useful for identifying distinct subtypes of disorders. Data aggregation also raises questions about the best way to combine data from multiple laboratories and the amount of data that might be needed to uncover subpopulations.

Specialized Infrastructures?

A data repository for complex diseases will need to deal with large amounts of heterogeneous data. This raises the question of whether specialized infrastructures will be needed to store the data. A number of recent reviews explore various aspects of what is big data (DeMauro, et al., 2014; Jadadish, et al., 2014). There is no doubt that the amount of data being collected in biomedical research laboratories is rapidly increasing. However, the scale of data collected by some physics experiments, by retailers, by social media providers, and by the government is often much larger in size or in the speed of acquisition than most current biomedical experiments (Schadt, et al., 2010), and we already have informatics infrastructures that allow both the storage and analysis of those data. Large biomedical data sets have terabytes to petabytes of data while data sets in those other domains have petabytes or even exabytes (Leung, 2014). The data generated by a biomedical research laboratory can certainly tax the data storage resources in that lab or in the department or even the university, but the same data could be easily stored using the solutions that have been created for big data in other areas at relatively low cost.

Will biomedical data become big data? The change in pace of biomedical data acquisition is hard to measure, but it is likely that genomic data will be the driver for increased storage and computational needs in biomedicine. A recent perspective (Stephens, et al., 2015) argues that the amount of sequencing data produced is doubling every seven months. This is roughly consistent with the growth of the sequence read archive (Kodama, et al., 2013) seems to be increasing by an order of magnitude roughly every 31 months since January 2009.

The question of the growth of genomic data is important since if biomedical data is growing more quickly than the growth in data storage capacity, infrastructure investments will have

to be made specifically to accommodate biomedical data. This genomic data will be only one sort of data that is needed for data mining to understand complex diseases. All of the relevant data will need to be made available in ways that are easily usable by the biomedical research community.

It has been estimated that the unit cost of storage capacity decreases by roughly an order of magnitude every 48 months (Komorowski, 2014). The increase in biomedical data storage is currently a little faster than, but in line with, the performance improvements for storage capacity. However, as Stephens et al. (2015) argue the increase in biomedical data may still be accelerating. For the near future it seems unlikely that biomedical researchers will need to worry about creating special data storage infrastructures or technologies beyond what has been created to deal with existing big data. However, if genomics experiments really begin to outpace the increases in generic storage capacity, specialized infrastructures will be necessary if all of the data are to be preserved. Stephens et al. (2015) correctly argue that the exascale data and computing centers that are in use today are the result of long range planning. The biomedical research community will need to assess this data growth carefully over the next few years and will need to find consensus to build appropriate data sharing and computational infrastructures to be used by the whole community for data mining and analysis of big data. The good news is that it is not necessary today to build a specialized informatics infrastructure to analyze data relevant to complex diseases.

While biomedical researchers might not have the largest datasets today, collectively, they probably have created the most diverse data sets. The heterogeneity in biomedical data arises from both the individual variability of subjects and samples as well as the diversity of experimental protocols utilized among labs at any given time and over time. One way to uncover true biological heterogeneity is by aggregating data from multiple laboratories. An example of a data archive that has begun to aggregate data from human subjects in mental health research will be discussed below. Just the early stages of that data aggregation show that research laboratories prefer to use idiosyncratic data collection instruments. The use of common data elements (CDEs; measures collected in the same way in multiple laboratories) can help minimize the variability due to different experimental protocols. The timing for the creation of CDEs will be discussed below. Finding ways to effectively aggregate heterogeneous data is necessary to have a data infrastructure relevant to complex diseases.

Successful Biomedical Data Repositories

There are several biological databases that have been established for a long periods of time and have evolved to be very useful for the research community. It is worth looking at the history of three of them to see what they tell us about effective ways to aggregate data. This history is not exactly the same as understanding whether a data repository can help understand complex disease, but it is helpful to understand what biomedical data repositories have been capable of and to see if there are common features for successful data repositories.

Protein Data Bank

The Protein Data Bank (PDB) began in earnest as a grass roots effort in 1971 (Berman, 2008). As the number of protein and DNA structures increased, the PDB steadily evolved. The early PDB provided a way to distribute the x,y,z coordinates of macromolecules to the research community. The repository performed some quality control (checking atom labels, checking bond lengths and angles), but in the early days of the PDB, those performing the experiments did not usually provide the raw data that would allow a full reanalysis of the structure that the laboratory submitted. The PDB assigns unique identifiers to each structure. The assignment of persistent identifiers seems to be a characteristic of useful data infrastructures.

Over time, much of the structural biology community agreed that it was important to share both the coordinates and the raw data. The objections to sharing data from some in the research community were overcome by data sharing requirements from both funding agencies and the journals (Berman, 2008). From the earliest days a subset of the structural biology community were strongly committed to data sharing, although not all members of the community were enthusiastic about making their data available.

Some of the enthusiasm for data sharing might be traced to an existing very successful data repository for crystal structures of small molecules (Groom and Allen, 2014) that preceded the creation of the PDB. Business models for databases will be discussed below, so it is worth noting that this small molecule database has evolved into a self-supporting non-profit organization that recovers operating costs from academia and industry while still providing structures free of charge to users. It is also worth noting that initial reluctance to share data seems to be the rule in most biomedical research communities.

As the number of macromolecular structures in the PDB increased, the need to automate the validation process during data submission became much more important. During validation, things like atom names, bond length, and bond angles are checked to see if they are within agreed upon ranges. The PDB pioneered the definition of the macromolecular crystallographic information file (Fitzgerald, et al., 2005) which describes all aspects of both the experiment and the resulting structure.

Finding the ranges of acceptable experimental results was greatly facilitated by the existence of a centralized repository along with an experimental technique that was mature enough to permit standardized descriptions of the experiment. The aggregated data in the repository made it straightforward to define standards for things like hydrogen bond lengths and facilitated the conversations that allowed the community to arrive at consensus about the expected range of variation for a particular structural parameter. Emerging standards and a data repository with many different structures naturally resulted in other discussions and inventions such as a single metric that describes the quality of a crystallographic experiment (Brunger, 1992). The fact that the companies providing hardware for the x-ray experiment did not attempt to use proprietary data reduction algorithms or proprietary data formats was a great advantage to understanding the strengths and weaknesses of various data collection methods and made it possible to compare data collected from the same sample on different devices.

The PDB is now at the point where virtually every macromolecular structure is deposited very quickly. Unlike the early days of structure determination, the PDB has made it possible to be a structural biologist without being a laboratory experimentalist. Today there are many academics interested in molecular evolution, classes of protein structures, drug design, and the simulated dynamics of protein motion who use the PDB as the source of data for their careers. These scientists contribute greatly to the field, but their work would not be possible without a data repository that has become part of the intellectual infrastructure for a research community.

The history of the PDB suggests that a very successful data infrastructure has several phases. The infrastructure starts by accumulating data and trying to make the data easy to use both for experts in the field as well as for other researchers. This ease of use generally starts with defining ways to make queries across data from different laboratories possible. At some point, these queries are refined by the accumulating data. This is when a discussion of standards related to experimental results is ready to happen.

The next step for the PDB was to try to develop consensus in the community around the requirements for a complete experimental description and for experimental results. The existence of a data infrastructure was a great aid in resolving issues at this point. Imposing standards before a technique is mature might stifle innovation or, more likely, will result in the community ignoring the proposed standard. Comparing a proposed standard with aggregated data is essential. If a standard proposes that a hydrogen bond should be linear, just how many violations to that rule can be found in the database? What is the cause of the violations? Does the preliminary standard need to be adjusted? Standards developed in the absence of a database may not be as robust as the standard developed by the PDB. After the standards have been defined, an automated validation tool that checks data before it is deposited into the database can be really useful. Such tools minimize the data repository staff time needed to validate each dataset and, more importantly, allows researchers to see potential problems or interesting features that they may not have noticed about their structure or their experimental data.

Once useful standards have been defined for both the experiment and the structure that is derived from that experiment, researchers from outside the laboratory and perhaps even outside the structural biology domain have a much easier time using the data. Rather than mastering a variety of different versions of an experiment and trying to relate them to each other, the outside user can just launch queries that make use of the defined standards. For the PDB, data collected using different data collection devices, in different locations, and even with dramatically different types of x-rays can be described using the established standard that describes the experiment. The standardized formats for experimental descriptions and for experimental output are also of great benefit for those writing data analysis software. With the emergence of standards and of powerful desktop or laptop computers, the structural data in the PDB became accessible to any researcher (Berman 2008), and it has become indispensable to many.

GenBank

GenBank (Benson et al., 2015) is a second very successful data repository that has had an evolution similar to the Protein Data Bank. GenBank is an annotated collection of all publically available DNA sequences. A brief history of GenBank is available (Strasser, 2008) and suggests many similarities to the evolution of the PDB.

The idea behind GenBank started with a gathering of molecular biologists and computer scientists at Rockefeller University in 1979 (Smith, 1990) although many of the ideas around data aggregation discussed at that gathering built on the early work of Margaret Dayhoff (Strasser, 2010).

As was true for the PDB, parts of the sequencing community recognized the value of a centralized repository and obtained funding for a data infrastructure (Smith, 1990). It appears that a significant difference between the early PDB and the infrastructures that turned into GenBank was the desire to correlate the sequence information with as much other biological information as possible (Smith, 1990). Ultimately, the home for GenBank became the National Library of Medicine at NIH which is well suited to that correlation task. The Entrez data retrieval system at NLM which provides access to GenBank and many other NLM data repositories allows the research community access to tools that make it very straightforward to correlate different types of data (Geer, et al., 2003). The community expectation for sharing genomic data ultimately resulted in the Bermuda Principles that governed the sharing of data from the human genome project (Green et al., 2015). Like the PDB, persistent identifiers are also assigned to all entries by GenBank.

The data stored in GenBank also helped spur conversations about standards. As the goals for the Human Genome Project evolved, the expectations for data quality were discussed and were adjusted as the experiment continued to develop (Collins, et al., 1998). GenBank has developed ways to upload data from different data collection devices in an automated fashion (Benson et al., 2015). Standards to describe file formats, the results from experiments, as well as standards related to combining genomic and clinical data all resulted from the compilation of data in GenBank (Husser, et al., 2006).

If there is a significant difference between the two databases, it might be in the evolution of the data collection technology. In the past 10 years, the experimental details of the sequencing experiment and related technologies continue to change in ways that are much more significant than the changes seen in collecting data related to macromolecular structures. As a result, metrics to judge the quality of the sequencing experiment (like a crystallographic R free) are still evolving (Uttukar et al., 2014).

Finally, there is no doubt that the data in GenBank informs much of the research in biomedicine today. It isn't possible to imagine the average biomedical research laboratory functioning without the data in that repository. The development of GenBank shares resulted in a number of features that the PDB also has.

Gene Expression Omnibus

Although not as old as either the PDB or GenBank, the Gene Expression Omnibus (GEO) is a third example of a successful data repository. GEO holds high-throughput gene expression data (Edgar et al., 2002). GEO allows queries either the experimental platform or the molecules of interest, and it allows data to be grouped together in meaningful ways.

As was true for the PDB and for GenBank, GEO originated with a desire from some in the research community to have a repository for high-throughput gene expression data (Edgar et al., 2002). The accumulation of data resulted in conversations about standards and the implementation of a useful set of standards (Edgar et al., 2006). These standards set up a set of data elements that should be reported by all experiments (Barrett, et al., 2006).

As was true for GenBank and the PDB, GEO assigns persistent identifiers for submitted data. It appears that the deposited datasets are being reused at an increasing rate (Barrett, et al., 2013). GEO differs a little bit from the PDB and GenBank in the ability to group raw data together into a “series” that is related to a biological question. This seems similar to the way of aggregating data into a “study” in the NIMH Data Archive that will be discussed below.

GEO is younger than either the PDB or GenBank, but it does seem to share the basic developmental history for a successful data repository:

- A call from parts of the community for a data repository.
- Accumulation of sufficient data to allow the community to develop standards to describe the experiment and the results from the experiment.
- Community expectations (often mediated by policies from the journals or the funding agencies) that the data be shared.
- The development of a query system that allows both experts and non-experts to use the data effectively.
- The development of a cadre of investigators who are focused on reanalyzing data made available in the data repository.
- The development of ways to link the data in the repository to other sorts of data.
- The creation of persistent identifiers that initially were only for the benefit of the data repository but turn out to be useful for a number of other informatics purposes.

Data Repositories for Brain Imaging

While GenBank, the PDB, and GEO illuminate some of the components of a successful database, examining data sharing in the brain imaging area offers other lessons. Magnetic resonance imaging (MRI) of the brain is a research area where lots of data are collected and many papers are published each year but where a data repository has not flourished. PubMed shows that in 2014 over 18,000 papers related to brain imaging were published. The raw

data and the analyzed data for most of those studies are not available outside the laboratory that collected the data.

There was an attempt to create a data repository for functional MRI data (Van Horn and Gazzaniga, 2013; Mennes, et al., 2013) about 15 years ago that seemed to have many of the same components that resulted in the successful repositories described above. The Journal of Cognitive Neuroscience supported the proposed data repository, called the fMRI Data Center (fMRIDC). There was not uniform support for the fMRIDC in the imaging research community, but there was support from some members of that community (Van Horn and Gazzaniga, 2013). That mixed support mirrors the early history of the PDB and GenBank. It is not clear whether the opposition to data sharing was larger than the opposition in the structural biology or sequencing communities. An additional difference that may have contributed to the poor growth of fMRIDC was that neither other journals nor the funding organizations required data to be submitted to a data repository.

The lack of a central repository for imaging data seemingly had effects on the standardization of the experiment. The standardization of imaging experiments has been difficult even at the level of data formats; without a centralized repository, different standards for acquisition, storage and analysis (Aguirre, 2012) began to proliferate. DICOM and NIFTI have emerged as widely used standards, but those standards are not implemented in identical ways across all software packages or instrument manufacturers (Poldrack, et al., 2013).

While there is some agreement about imaging file formats, until very recently, there has been little agreement about the way to conduct an imaging experiment. This was one of the key arguments made against the fMRIDC (The Governing Council of the OHBM, 2001). Even when the same pulse sequences are used in the same type of data collection instrument, there are often significant differences in brain images collected at different sites, and great efforts have gone into trying to understand and minimize those differences. (Friedman and Glover, 2006; Jack et al., 2008; Gunter et al., 2009). Finally, MRI is a field in which the major instrument manufacturers use proprietary software that makes it difficult to directly compare results obtained on different instruments (Guggenberger, et al., 2013).

Is the situation hopeless for aggregating imaging data? The failure to discover imaging biomarkers that can help distinguish individuals with disorders like schizophrenia or autism has opened a discussion about how to improve the information that can be derived from imaging data at NIMH. Those discussions have resulted in the creation of a data archive that will accept imaging data (see below). While this archive is not expected to turn into a centralized imaging data repository, it might help facilitate the establishment of data collection standards. Others have argued that the imaging community should reconsider the long standing opposition to such repositories (Van Horn and Toga, 2009). Those are hopeful signs.

What may change the opposition to data sharing in the imaging community are a number of new imaging data sharing efforts that have shown some unexpectedly positive results. One of the bolder attempts to share data has been the 1000 Functional Connectomes Project

(FCP) and its successor the International Data Sharing Initiative (Mennes, et al., 2013). The FCP began by trying to understand whether resting state fMRI data measured at different sites could be combined without prior agreement concerning data collection protocols. Following an initial data aggregation from five sites that suggested that the aggregated data might be useful, many other imaging laboratories volunteered to contribute their data as well. The first FCP data release contained limited phenotypic information and images from over 1300 participants at 30 different sites (Mennes et al., 2013). The aggregated data clearly showed that similar functional architectures could be found in data from multiple laboratories even though no effort had been made to collect data using similar protocols (Biswal et al., 2010). This result was quite surprising and may attest to the strength of the underlying signal being measured in the resting state fMRI experiment. The FCP data aggregation experiment has been done with very limited external funding using a variety of existing data sharing infrastructures such as NITRC (Luo, et al., 2009).

The OpenfMRI project (Poldrack, et al., 2013) is attempting to create a data resource for task based fMRI experiments. This is a logical extension of the 1000 Functional Connectomes Project, but requires creating a way to precisely describe the task that the subject performed during the experiment. Despite some initial attempts (Turner and Laird, 2012), describing these metadata for the task fMRI experiment remains a significant challenge. The OpenfMRI platform is not quite as mature as the FCP, but the initial indications are that the task fMRI experiment is also sufficiently robust to allow useful data aggregation from different laboratories.

A third sort of data sharing has occurred for large imaging projects that have standardized data collection at multiple locations prior to the start of the experiment. There are many examples of such projects that collect imaging data at several different sites and use phantoms and harmonized data collection protocols. Generally, these collaborations have a quality control center that aggregates data and evaluates the data to ensure that each site is performing the experiment as expected (Walker, et al. 2013).

A fourth, rather different, data sharing infrastructure related to human imaging has been created for the Human Connectome Project (Van Essen, et al., 2013). The focus of the Human Connectome Project is to use advanced imaging techniques on a state of the art scanner to determine the human connectome in 1200 healthy young adults. Structural MRI, resting state fMRI, task fMRI, and diffusion imaging are all being measured on the participants as is a significant battery of behavioral tests.

The HCP data are being released close to the time when they are measured. This may be the first time that a significant imaging data set has been made available to the public before the group measuring the data published on their results. Adopting the same rapid data sharing approach as has been done for genomic data for many years seems to have had interesting results. Nearly 100 papers have been published using data from the HCP even though only half of the final data set has been released.

Although the HCP data acquisition is much faster on certain scanners, the pulse sequences have been made available for a wide range of different data collection instruments. Because

of the interesting early results (Smith, et al., 2013), NIH has decided to expand the initial HCP experiment in several different directions. NIH has made an award for an HCP informatics coordination facility, the Connectome Coordination Facility. The contributing laboratories will have to use the published HCP data collection protocol in order to deposit data. That database will hold connectomic data from many laboratories, and will offer advice to those trying to collect data according to the HCP protocol. Comparing the resting state fMRI data from multiple laboratories in the Connectome Coordination Facility with the data in the FCP will help understand the effect that data collection harmonization has on the overall quality of the data. If it turns out that the HCP data are much more useful than the FCP data have been, the imaging community will have to consider carefully whether the time to promote standards in brain imaging has finally arrived.

In addition, NIH has published request for applications for connectomes related to human disease (PAR-14-281) and related to different age groups across the human lifespan (RFA-AG-16-004, RFA-MH-16-150, and RFA-MH-16-160). As with the original HCP project, all of these data will be released to the research community quickly via the Connectome Coordination Facility. The Connectome Coordination Facility is not a universal data repository for imaging data, but it will be the central repository for HCP data. In developments that are just unfolding while this review is being written, it appears that other large imaging efforts are voluntarily adopting the HCP data collection protocols. This may mark a real turning point for data aggregation and standardization in the imaging field.

One additional imaging data sharing experiment is the approach being explored by the ENIGMA consortium (Thompson, et al., 2014). Under ENIGMA, a collaboration between research groups with imaging and genomic data relevant to a particular biological question is created. These various ENIGMA working groups do not share their raw data with each other or with the public. They do work out exactly how their imaging data sets are going to be analyzed, and then they merge the results to see if they can answer the question. A major driving force behind this approach is the belief that those who have measured the data understand best how to analyze it because of their knowledge of the data collection protocol and the research participants. This mixed collaborative model resulted in a number of interesting papers (Stein, et al., 2012; Jahanshad, et al., 2013; Kochunov, 2014; Li, et al., 2014; Hibar, et al. 2015).

It may turn out that this limited data access/shared data analysis model will prove more productive than the other imaging data sharing models outlined above. It certainly allows groups who already have significant storage and computational resources to use their existing hardware and software without creating a centralized data storage/compute infrastructure or standards across an entire field. The limited access model could also be used for other complex data types. Whether a limited access model ever graduates to the PDB/GenBank open data sharing model with defined standards for the imaging community is an important but open question. It is possible that the discussions that occur in the various ENIGMA collaborations will speed the adoption of experimental standards. It is also possible the lack of a centralized repository will impede the creation of such standards. The results of this experiment have important implications for data repositories aimed at understanding complex diseases.

In the imaging area, the lack of a single centralized database has resulted in multiple different approaches to data sharing which are still evolving. It isn't clear whether any of these approaches are going to lead to standards like the PDB and GenBank have, and it isn't clear whether any of them will end up being adopted by a majority of the imaging community. However, it is clear that the existing methods could be the starting point for serious discussions about standards.

The funding agencies supporting brain imaging research and the journals publishing such research have been unwilling to mandate data sharing. This lack of a data sharing expectation may be the key difference between imaging data and protein structure or genomic data. It might be argued that the lack of a mature data collection technology for imaging is also a barrier, but the continued evolution of the sequencing experiment has not prevented that community from aggregating their data.

If data collection and image standards don't appear, the ENIGMA approach for pooling data may be the right model for sharing this data. While ENIGMA may end up being the best model for sharing complex data, such a result has very significant consequences for understanding issues related to imaging data reproducibility as well as for the many efforts to promote global data aggregation (see below). One troubling aspect of the ENIGMA experiment is that it does not seem to provide space for reuse of the data in ways that the original experimentalist could not have anticipated (Masseroli, et al., 2014) since someone, perhaps from a very different scientific discipline, with a novel data analysis idea will have to convince the (potentially many) groups with the data to set up a collaboration and do the computations.

For a variety of reasons, sharing imaging data has been harder than sharing data from crystal structures or sequencing or high-throughput expression experiments. ENIGMA and the HCP standard are both hopeful signs as are the development of the Connectome Computational Facility and the NIMH Data Archive, described below, that allow the aggregation of data from multiple laboratories and queries across that data.

What is less clear for the development of data infrastructures related to complex diseases is whether the failure of a successful imaging data archive to develop is due to opposition from the research community, from the failure of the journals or the funders to mandate data sharing, from the proprietary data formats used by the various instrument manufacturers, by the lack of overall standards for data collection or data analysis, or by the complexity of the imaging experiment. The current data sharing/data reuse efforts underway in the imaging community may provide answers to that question. What is clear is that the history of sharing data in the imaging community has been very different than the history of data sharing in the successful repositories described above.

Elements of an Infrastructure for Sharing Data Related to Complex Diseases

So far, we have seen that the technology exists to create a data repository related to complex diseases. The case histories illustrate features of successful data aggregation efforts, show

the value of data aggregation in some domains, and show some of the unresolved questions with different data aggregation strategies that have relevance to complex diseases. This section deals with additional issues that would need to be addressed by a data repository focused on complex diseases. The topics raised in this section have been divided into a “Consistent Elements Found in Successful Data Repositories” section and an “Open Questions” section where appropriate.

Common Data Elements/Standards

A) Consistent Elements Found in Successful Data Repositories—As they evolved, both GenBank and the PDB developed standards to describe the data and the experimental results in the databases. Such standards allowed others to develop software that used the data in the database, facilitated queries, and made the data easier to use by those who did not collect it. The development of those standards also allowed communities to discuss what sort of data should be provided to describe an experiment and results. In general, the centralized databases seemed to provide the necessary information for a community of scientists to discuss the best way to describe all of the parameters of that experiment and to begin to describe derived observations from the experiment.

The PDB and GeneBank both deal with a limited range of experiments, so the standards they created were limited in scope. Databases trying to aggregate all data related to a complex disease have a much larger challenge since many different sorts of data are likely to be part of the data collection. It is certainly possible for a leader in the field (a funding agency or a key journal) to try to define standards for that field. The lack of standard experimental protocols in MRI imaging contributes to the difficulty that non-experts have in using data in that field. However, there is real uncertainty about the best time for a research community to adopt and enforce standards. There has always been concern that early adoption of standards could limit innovation as an experimental technique develops although it isn't clear that there have been any cases where proposed standards have stifled creativity.

Recently, several different institutes at NIH have tried to set up common data elements (CDEs) for studies involving human subjects in a particular research domain (Gershon, et al., 2010; Hamilton, et al., 2011; Grinnon, et al., 2013; Conway, et al., 2014, <http://www.nlm.nih.gov/cde/>). The CDEs are often clinical assessments, but they can include other measurements such as blood pressure (Hamilton et al., 2011). The common data elements generally are meant to be collected by many laboratories in the field. This facilitates aggregation of and comparison of data from different laboratories.

Usually, these NIH supported CDEs have been created by assembling a group of experts in a reasonably mature research area and asking them to recommend a set of measures that are well validated and would be useful to a research community. Following the initial recommendation, there is outreach to the entire community to allow those who were not involved in the drafting of the CDEs to comment on them. The comments that are obtained in this community outreach effort are considered while reaching final consensus on the measures to be recommended. Establishing these CDEs has usually come with strong encouragement from NIH for all researchers to use the recommended measures going forward (NOT-DA-12-008, NOT-MH-15-012).

B) Open Questions—The situation is a little bit different for developing technologies. The NIH BRAIN Initiative is currently focused on developing new technologies to enhance our understanding of the human brain (Jorgenson, et al., 2015). One component of that initiative is to develop a systematic inventory/census of cell types in the brain. In the first year of the BRAIN initiative, NIH funded 10 pilot awards in the cell census area. In order to develop a common description of experiments being conducted and their results, those awardees are trading data and running local data analysis programs on data from other laboratories. It is hoped that a deep and uniform understanding will emerge from this data sharing that will allow the groups to settle on a first draft of CDEs that will be useful across the field. It isn't clear whether this approach will result in CDEs more quickly than the approach used by the PDB and GenBank, but it could do so. It might also be necessary to use this consortium approach in areas where the science is rapidly developing.

These NIH supported CDE efforts are still a fairly recent development, so it isn't clear whether the targeted research communities are adopting the recommended sets of common measures or not. Unlike the history of the PDB and GenBank, journals do not appear to be requiring articles in a research domain where CDEs have been established to use them. In many cases, there does not seem to be a data repository associated with the communities for which CDEs have been established (http://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html). As a result, it will be difficult to understand how well the proposed CDEs cover the research being done by the community. The lack of a single common data repository or a method of easily discovering relevant CDEs will also hinder evaluation of the adoption and will make it harder to modify those data elements to better fit the needs of the research community.

Despite these issues, CDEs should help research communities compare data from different laboratories. Such comparisons are probably going to be necessary to resolve the reproducibility issues that are now recognized as being a significant problem throughout biomedical research (Collins and Tabak, 2014). A listing and registry of some biomedical standards related to biological data is now available (<http://www.biosharing.org/>, Min et al., 2014).

Creating Knowledge from Data

B) Open Questions—While common data element efforts will help integrate data across laboratories moving forward, there are a shocking number of data collection instruments in use today in mental health. The NIMH Data Archive (NDA, described below) is a comprehensive archive of clinical research funded by NIMH. The NDA has been forced to deal with all of these existing data collection instruments and techniques. At least in the early days, a data repository for complex diseases will have to deal with the same level of heterogeneity. NDA started with a focus on autism but has expanded in the past 24 months. Currently, investigators have defined nearly 1500 distinct data collection instruments in NDA. We anticipate that number may grow to over 3000. The very large number of data collection instruments is not unique to mental health. PhenX (<https://www.phenxtoolkit.org/index.php?pageLink=browse>), a repository of well established, broadly validated measures currently has information about nearly 1000 protocols.

Each of these protocols or instruments typically has multiple, often many, individual questions. Is it possible to launch effective queries across so many instruments? The answer to that question really isn't clear. Imagine the level of effort that would be required for anyone just to find all of the questions that dealt with psychosis among thousands of data collection instruments each with many questions.

It is certainly possible to define the relationships between instruments in a particular topical area. A very nice example of such work can be found in the recent paper defining quantitative relationships between data collected in a variety of instruments used in autism (McCray et al., 2014). That group constructed an ontology that allows data on similar concepts collected in different ways to be aggregated together. A good specific example is finding subjects who exhibit excessive compulsive behavior. That question is asked in four different data collection instruments examined by McCray et al. (2014). The range of values that correspond to such behavior are numbers in such questions (between 1 and 2 in one instrument, between 1 and 3 in a different instrument) and are descriptions in other questions (moderately, quite a bit, extremely). These ontologies are really useful, but each of these ontologies takes real effort to construct. It may not be practical to expect the bioinformatics community to map large numbers of concepts among hundreds of thousands of questions that come from thousands of data collection instruments. The emerging tools related to the semantic web (Cheung, et al., 2009) and especially to semantic normalization (Paraiso-Medina, et al., 2015) offer a path to resolving these issues but not all of the needed infrastructure is currently available.

One would expect that as data repositories become available and as the research community begins to understand the scope of the problems caused by tens of thousands of different clinical data collection measures, that consensus will emerge about when to create something new versus when to use an established instrument. Unfortunately, the research community will have to grapple with the existing myriad of data measures for some time to come.

Data Federation versus Centralized Repositories and the Business Model for Maintaining Data

A) Consistent Elements Found in Successful Data Repositories—One very important question about a data repository for complex disease revolves around the business model for keeping the data available. The NIH Public Access policy to make journal articles freely available was instituted in no small part to make the science that the US taxpayer paid for available to all (Suber, 2008). Business models to support data repositories for complex diseases that require the user to pay for the data that originated with government support seems inconsistent with that spirit. Even if they were consistent, requiring a user to pay for data access is likely to reduce the re-use of the data.

The business model question is important for both the funding agencies and for the research community. Even a relatively small database like the PDB is very expensive to maintain. As of May, 2014 the PDB held about 130 GB of data (<http://www.wwpdb.org/downloads.html>). That amount of data isn't small enough to fit on a memory stick, but it could easily fit on a desktop hard drive where 1-2 TB of data storage costs less than \$100. The cost to the federal

government for the PDB is roughly \$8M per year (NSF Award Abstract #0829586). That cost reflects more than just data storage, but it is a surprisingly large amount for a database where good standards, robust automated software to deposit data exist, and the community doesn't need too much handholding or encouragement to deposit data. Costs of that size certainly take away from the funds that federal agencies have to support the measurement of new data. Finding a business model to minimize the cost of storing data while still making it broadly available to understand complex diseases is an urgent concern.

Since many biomedical datasets are relatively small, it is often straightforward for individuals to share data using e-mail or an infrastructure like Dropbox. This allows pairwise or small group collaboration, but it does not allow others to discover that data are available. However, many authors are unwilling to share their data even though the infrastructure to do so is available (Savage and Vickers, 2009; Piwowar 2011). Even in the case of pairwise collaboration, the lack of common data elements or formal definitions of the experiment often require significant discussion between those who measured the data and those who want to reuse it to make sure that everyone understands exactly what is being reported in the columns of a spreadsheet or database.

B) Open Questions—In theory, it is possible for researchers to use a federated data sharing model to make data more broadly available. The federated model allows data to be stored at multiple locations and for queries to find the data without a central repository. A good example where this is possible is in the sharing of imaging data. For imaging data either the XNAT (Marcus et al., 2007) or COINS (Scott et al., 2011) data sharing software allows the group measuring data to make it available to a defined group of researchers or to the whole community. However, setting up such a data sharing infrastructure in a laboratory requires real effort. Responding to requests to help with questions about the data also requires significant commitment for a widely used dataset. The long term sustainability of the data is also a question. If grant funding for the lab runs out, the data may just disappear. As the costs to store data decrease, the storage costs for legacy data sets should become very inexpensive. However, it still takes time and effort to move the data to new storage formats and to set up the data sharing software and permissions in the first place. Some of the global data repositories discussed below may solve some of these issues, but the absence of well documented standards will mean that laboratories sending data to a repository will likely have to answer questions about the data.

Data sharing software like XNAT makes data access possible, but it does not make the data easy to discover. Under the Big Data to Knowledge (BD2K, <http://bd2k.nih.gov/>) initiative, NIH is trying to create a data discovery index (<http://biocaddie.org/>) that will make data discoverable and accessible. If successful, this project aims to create an index for data that works the same way that PubMed does to find journal articles. Nature's Scientific Data, the INCF Data Space (<https://github.com/INCF/ids-tools/wiki/Using-the-INCf-Dataspace>), and the Neuroscience Information Framework (Gardner, et al., 2008) have also created infrastructures to help with the data discovery problem. The business model for funding data sharing is likely to be an important topic for the BD2K initiative for some time.

No matter where the data related to complex diseases are stored, someone is going to have to pay for it. Until an outside entity sees real value in keeping a data repository available, as has happened for the data repository for crystal structures of small molecules, it seems likely that those funds are mostly going to come from funding agencies either as part of the funds provided under research grants, as indirect costs that support libraries which can host data at an academic institution, or as data repositories run directly by the funding agency. Finding the right way to pay for these costs and the right level of funding to support legacy data preservation is an urgent question.

Consents and Data Ownership

A) Consistent Elements Found in Successful Data Repositories—When thinking about data relevant to complex disease, the question about how to deal with data from human subjects who might be re-identifiable from the data is another important question. The requirement to obtain informed consent from research participants is well understood, and all ethical experiments involving human subjects obtain such consents. A presentation of plans for the distribution of data from the experiment is often a part of the informed consent process. Such consents can allow broad data sharing, or the consents can hinder broad data sharing.

For complex diseases, restrictive consents are a real problem. As the biological basis for complex diseases are better understood, unexpected linkages between certain areas are discovered. A case in point is the recent potential linkage between the gut microbiome and mental disorders (Foster & Neufeld, 2013). Data from a research subject who donated a stool sample for restricted use in a study about Crohn's disease could not be used for a study on schizophrenia without re-contacting the subject for a revised consent or without an institutional review board (IRB) allowing such data sharing. Both pathways are costly and take significant time and effort. As a result, data consented for narrow sharing is far less useful when thinking about complex diseases than data consented for broad sharing.

B) Open Questions—Some efforts are being made to develop a dynamic data sharing infrastructure that will allow a research participant to control data access. These systems allow the participant to decide which data should be shared, who should have access to that data, and to change data access over time (Kaye, et al., 2012). There is a great deal to be said for returning the data sharing authority to the research participant, but it is not clear whether these new infrastructures will be robust or whether the research community and IRBs will be willing to adopt them.

There are concerns about the potential harm that re-identification might have for those who participate in clinical research. However, it is very clear that when thinking about complex diseases that the data should be made available as widely as possible. Either broad consents in the current system or adoption of a data sharing infrastructure that gives control of the data directly to the research participant could be used. Consents that restrict the use of data to a small subset of the research community interested in a particular domain or topic will likely result in that data becoming unavailable beyond the research group that collected the data.

Where Should the Data be Held?

A) Consistent Elements Found in Successful Data Repositories—While data storage and computation on an existing dataset are relatively cheap and efficient today, data transfer continues to be relatively slow and difficult for datasets of the terabyte and larger size. As a result, a key problem for any large data repository is making the data available for computation without transferring the data. It is much more efficient to compute on large datasets without moving them. Security of data from human subjects which will be the focus in a repository devoted to complex diseases also improves when the data set is not held in multiple different locations each of which has to guard the data against unapproved data access.

There are a variety of open data repositories that will accept biomedical data. The 1000 FCP, OpenfMRI, and NITRC projects for imaging data have already been discussed. The CRCNS data sharing website (Teeters, et al., 2008) will accept high quality datasets that are useful for testing computational models of the brain as well as new analysis methods. Dryad (<http://datadryad.org/>) accepts data associated with a publication, assigns a digital object identifier (doi) to the dataset to aid in discoverability, and charges a relatively small fee to hold the data. Users who want to download the data pay no charge. Figshare (<http://figshare.com/>) performs similar functions. There are many other examples of places where a researcher can drop data off for others to use.

B) Open Questions—The need for a comprehensive solution to the data discovery problem is absolutely required since it seems unlikely that there will ever be a central data repository for data that might be of use to researchers interested in complex diseases. The digital object identifier and other persistent identifiers are likely part of the data discovery solution, but currently there is not a service that allows a biomedical researcher to launch a Google-like query and find datasets today. Such a robust data discovery service is needed to allow researchers interested in complex diseases to find data that might be of interest – whether that data are stored in a general repository, in a more specialized repository, or in a laboratory that has made data accessible. In addition, the service will need to provide a way to track the number of times that data have been accessed. Such accounting would make it very easy to see how data are being used and for what purposes. This would be very helpful in giving credit to the researcher who gathered the data and would allow journals and funding agencies to understand which data was useful to a research community.

Clearly, standards are a critical component for a robust data discovery solution as well as for making use of the data once it has been discovered. Biosharing (<http://www.biosharing.org/>) as well as efforts by the National Library of Medicine (<http://www.nlm.nih.gov/cde/>), the National Cancer Institute (<https://cbiit.nci.nih.gov/ncip/biomedical-informatics-resources/interoperability-and-semantics/metadata-and-models>), the PhenX Toolkit (<https://www.phenxtoolkit.org/>) and others are starting to tackle the problem of making common data elements available to the research community. Despite these efforts, we are still reasonably far from having a functioning data search engine, and as a result it will remain difficult to find all of the data that might be relevant to a complex disease.

How Much Data are Needed to Tackle Questions about Complex Diseases?

B) Open Questions—The answer to the question of how much data is needed to understand a complex disease is probably a great deal. If complex diseases really have multiple underlying genetic and/or environmental causes and also have an undetermined number of subgroups within a particular “disease”, a great deal of data is probably necessary to understanding what is going on. A search through aggregated datasets is not going to replace well designed experiments to understand the underlying biology, but such searches should suggest experiments that the research community would not have been obvious otherwise.

The need for heterogeneous data from a large population suggests that research participants are going to have to provide data from outside the confines of a defined experimental protocol. It also suggests that citizen scientists may be needed to look at the data to find potential clues to understanding complex diseases.

Infrastructures that allow people to volunteer to share their data already exist and have been used in interesting ways. Patients Like Me (<https://www.patientslikeme.com/>) is an example of an infrastructure that could be used to understand complex diseases. As of May, 2015 Patients Like Me (PLM) has attracted over 325,000 research participants to share some data about their medical condition. More than 60 papers have resulted from the data they have accumulated. Even assuming that a number of PLM research participants do not visit the web site frequently, it is easy to imagine how researchers could use the platform to help understand the number of subgroups that exist in a complex disease and to recruit research participants from a subgroup or a large group to provide data to test a hypothesis. A similar data infrastructure focused on autism, the Interactive Autism Network (IAN) has demonstrated that deposited by ordinary individuals can be of quality that is as good as data collected by trained scientists (Lee et al., 2010). The potential for very large sample sizes, the fact that those with a complex disease can volunteer information that a narrowly focused research study designed to produce a research paper could not spend the time and effort to collect, the ability to easily collect data related to the environment, and the high quality of volunteered data all strongly suggest that key insights needed do understand complex diseases may come directly from those affected by the disease rather than being uncovered in a small research study.

Clearly, smaller focused studies would be needed to confirm and extend anything discovered in a database composed of volunteered data. There are also serious questions about sample bias in data infrastructures such as PLM or IAN. However, the ability of those with complex diseases to share data with each other (Wicks, et al., 2010) and the ability of the research community to obtain data quickly from large cohorts at relatively low cost points to this being an important next step in understanding complex diseases.

Citizen Science

B) Open Questions—In addition to the question of whether those who have a complex disease can provide useful data outside the structure of a formal research protocol, there is also a question about who should be doing data analysis. The number of different types of

data that may be accumulated in a PLM-like data infrastructure as well as in data repositories containing data from research laboratories is going to be hard to understand. As a result, even if the perfect data discovery service existed and if all of the data had broad consents that allowed accessibility, who should have access to the data?

Until recently, this wouldn't have been a difficult question. Scientists would have been the only group allowed to interpret the data. However, the past 15 years has seen a resurgence of citizen science. The start of this trend may have been the Folding@home project from the Pande laboratory at Stanford. That project allowed anyone with a computer and an internet connection to contribute computational resources when they were not being used (Beberg, et al., 2009).

The citizen science idea has expanded to invite outsiders to participate in solving scientific problems. Sometimes this participation has been cast as a game such as Foldit (Cooper, et al., 2010). In other cases, a prize has been announced and the ability to participate has not been limited to experts in the field. Sage Bionetworks has learned a great deal about how to effectively manage such contests (Friend & Norman, 2013). Amazon's Mechanical Turk may make it relatively inexpensive to perform potentially tedious data analysis/data wrangling at moderate cost and may also provide a source of data (Buhrmester, et al., 2011).

The personal interest of those who have a complex disease as well as the need to explore unconventional approaches that would not be funded or would not lead to the advancement of an academic career strongly suggest that data about a complex disease should be made available outside the professional scientific community.

Data Aggregation in a Limited but Large Domain – A Potential Model for Other Complex Diseases

In an effort to explore some of the issues raised above and with the hope that insight may be obtained for some of the many complex mental illnesses, the NIMH has established a Data Archive for all clinical research data related to mental illness. The domain of science is clearly large, but limiting the archive to data from human subjects permits a two dimensional organizational structure. The data infrastructure is organized around global unique identifiers for each research participant (Johnson, et al., 2010) and data dictionaries that are defined by the research community and allow each laboratory to describe the experiment they have performed (Hall et al., 2012). There is a single Oracle database that serves data to multiple websites. The four web sites are the National Database for Autism Research (NDAR, <https://ndar.nih.gov/>), the NIH Study of Normal Brain Development (<http://pediatricmri.nih.gov/nihpd/info/index.html>), the National Database of Clinical Trials Related to Mental Illness (NDCT, <http://ndct.nimh.nih.gov/>), and the RDoC database (RDoC db, <http://rdocdb.nimh.nih.gov/>). Each web site is aimed at a subset of the NIMH research community, but queries launched from any web site can run across all of the data in the Oracle database.

All of the research subjects have consented to broad data sharing, so some of the issues with consents discussed above have been resolved in the NIMH Data Archive (NDA).

The global unique identifiers in the NDA allow information about a single subject seen in multiple laboratories to be aggregated without personally identifiable information about that research participant being shared between laboratories or with the data infrastructure. This GUID infrastructure is being made available to other research communities and could be a key component of a broader infrastructure for the understanding of complex diseases.

The data dictionaries that describe the research experiment represent an uneasy compromise between imposing common data elements and allowing each research laboratory to define experiments in isolation from the rest of the community. By early 2016, nearly 1500 different data dictionaries containing over 130,000 individual questions were defined. If a researcher is using a data collection instrument or methodology (clinical assessment, MRI, genomics experiment, ...) that is already defined, they must submit data according to the defined format. If an investigator is using a data collection instrument that hasn't been defined, they work with the NDA to create the new data definition supporting their research. If a researcher is using a local modification of an existing dictionary, NDA often is forced to define a new version of the data dictionary and to curate the questions in these similar instruments to allow effective queries.

Data dictionaries allow the NDA to validate all incoming data to make sure that the answer to each question is in the right format and that the valued being reported are within the defined allowable range for that question. Since data are generally deposited every six months, this permits a researcher to find and correct data collection problems close to the time when the data were collected. Although data are deposited every six months, the data are not shared with the research community until a paper has been published or an agreed upon period of time has elapsed. This data validation step has proved to be enlightening to those collecting data and managing studies. Discovery of problems with data formats are best solved in an ongoing fashion in a research study rather than trying to resolve such issues when papers are being written long after the data have been collected. Although it will be difficult to measure the impact that a formal validation tool has on experimental reproducibility, it certainly can't hurt.

The data dictionaries also permit a researcher to launch queries across data collected in multiple laboratories. Relatively simple mapping (M/F to 0/1 to describe sex) is done by NDA staff. Even this mostly straightforward curation takes a great deal of time and effort, but it is essential to make the data infrastructure useful.

A more serious problem was sketched above for finding all of the questions that deal with something like finding a cohort of subjects who exhibit excessive repetitive actions. Launching such a query would require the investigator to know which data dictionaries have questions related to repetitive actions as well as the range of values for each question that means excessive. A defined ontology can certainly help in such situations (McCray, et al., 2014), but such ontologies are likely only starting points for a discussion among the research community. A variety of different queries are currently enabled across the NDA. Input from users is welcomed to define other useful queries. Finding a robust solution to this problem is essential for any infrastructure that is going to hold data about a complex disease.

The data dictionaries and the global unique identifier effectively allow a large two dimensional matrix to be constructed with individuals on one dimension and results from experiments (either raw data or derived/analyzed data) in the other dimension. The matrix is sparse, but this seems to be a useful framework for aggregating data from clinical research. Both the availability of a large number of data elements as well as the availability of data from multiple subjects using those data elements may lead to some progress in constructing useful queries.

The NDA allows a user to define a block of data (called a study) that might be useful to others. In most cases, these data are associated with a research publication, but they do not need to be. The results from data processing pipelines are another obvious use for a study. Cohorts that could be used as the basis for comparing data analysis pipelines could also be easily imagined. The NDA assigns a doi to each study, and these dois have been relatively popular among users.

The NDA holds both genomic and imaging data in the Amazon cloud. These data are made available to the research community and allow analysis pipelines to compute on the data without moving the data. The storage costs are paid for by the NDA. Computational costs are paid for by the user. Some data analysis pipelines are made available to the research community. The NDA makes those pipelines available but generally does not try to answer questions from users about the pipelines directly. Such queries are referred to the groups that have created the pipeline. Recently, publications have appeared on global analysis of the imaging data (Torgerson, et al., 2015) and genomics data (Krumm, et al., 2015).

How Would it Actually Work?

The preceding sections outlined some of the challenges as well as potential solutions to creating a data infrastructure that would be useful in defining subpopulations in a complex disease group and in finding biomarkers for those subpopulations. In this section, a potential complex disease infrastructure is proposed and some recent results from existing data infrastructures are highlighted to suggest how the data in such an infrastructure would be used.

Collecting data from a large number of research participants directly is likely to be important for a successful complex disease database. Such an infrastructure should start by inviting those with a complex disease as well as controls to enroll in the database. It seems critical that the research participant be in control of who has access to their data and be able to make changes to that access when they want to. With a very large number of subjects, losing data because a research participant has changed his sharing preferences, should not matter. However, the dynamic availability of data will require an identifier to be assigned to the cohort providing data for a particular study to enhance reproducibility in data analysis. In addition, the research participant should be in charge of whether they are willing to be re-contacted to take part in further studies. At enrollment, a GUID or a related identifier would be generated so that those analyzing the data would not have direct access to personally identifiable information.

Once the research participant has enrolled, they could provide data in a number of different ways. The data could come from traditional clinical self-report data collection instruments and questionnaires about drugs and treatments being taken, but the participant might also provide data from an electronic medical record, from an app installed on their cell phone, or from a device that they wear to monitor their daily activity or fitness. Some research participants will take part because of their interest in a particular complex disease. Others will do so as a public service.

It is likely that a social media component will make the infrastructure much more attractive to many research participants. The social media component could also become very useful to the research community as discussions there might be a very effective vehicle to discovering unexpected correlations or subpopulations that have somewhat different versions of the disease. The social media component should be viewed as a tool that enables citizen science.

The complex disease website can significantly increase the likelihood that a research participant will return regularly by providing updates about new data that have been received, new research participants who have enrolled, and information about newly opened data collection efforts and results from studies that have been completed. Such results might come from the scientific literature, but they could also come from citizen scientists.

How would research scientists interact with this infrastructure? There are three obvious ways. The first would be to mine the data for interesting observations. This data infrastructure could eliminate many of the barriers that clinical researchers currently face. The cost to analyze preliminary data that is already in a data infrastructure is trivial. Rather than needing to devise a protocol, get IRB approval, find funding, recruit 30 subjects, and then analyze the data, all a researcher would need to do is find the group that they are looking for and see if that group has already provided the needed data. A process that might take years could be reduced to hours if the desired data were available.

Once a group has been identified, the second way a research scientist would use the infrastructure is as a recruitment tool. If an existing group has not provided the data needed for a study, those allowing re-contact could be asked if they were willing to complete a clinical assessment, provide blood, go to an imaging center to have particular scan done, or any of the other things that a researcher might want. Much of this data would go directly back to the data infrastructure. If subjects with a particular set of characteristics are not available, the social media component of the infrastructure could make it possible to find the desired cohort at a very low price.

Finally, it is important that the research scientist commit to depositing their raw data and their analyzed experimental results to the infrastructure. This will make the data broadly available to anyone who might want to use the results to extend them or to verify them. Some of the participants are likely to become very engaged with the research effort and may provide ideas/correlations that would not have occurred to the researcher.

While there are technical hurdles and policy issues would have to be addressed if such a data infrastructure was created, none of those issues are insurmountable. Would the infrastructure have any real value? Some recent papers suggest ways that it could be used to advance our

understanding of complex diseases. A recent paper by Levine et al. (2015) shows how clinically significant biomarkers for schizophrenia can be derived from a large number of research participants. Similar approaches were used by Uher et al. (2012) in depression.

The data infrastructure proposed above really is a new paradigm for doing science that involves the research participant much more in various aspects of data collection and data analysis. With a large number of users, the system could also potentially contribute ongoing surveillance of the efficacy of various treatments (Curtis, et al., 2012) and could even be used to deliver advice or treatments through a smartphone app (Tregarthen, et al., 2015).

A summary of the recommendations for those concerned with a data repository for complex diseases are below.

For the IT Infrastructure/Informatics Communities:

1. Provide an effective data discovery infrastructure.
2. Provide an infrastructure to monitor data access and reuse in order to give those who measured the data appropriate credit.
3. Find inexpensive easy to use methods to allow data queries across heterogeneous data sets.
4. Carefully monitor the growth of biomedical data relative to the growth of commodity data storage and the speed of data transmission pipelines and data analysis platforms to ensure that we do not arrive at a point where the biomedical data can't be used because the infrastructure doesn't exist.
5. Find ways to link data that resides in different repositories, as appropriate.

For the research laboratory:

1. Deposit data in an appropriate data repository with a validation tool close to the time when the data are measured.
2. Share the data quickly rather than hoard it.
3. Whenever possible use common data elements or other defined standards rather than invent new data collection methods.
4. Commit to defining the data collection methodology sufficiently completely to allow data reuse.
5. Make use of data from other laboratories.

For the funders:

1. Require the deposition of data in an appropriate repository as a condition of grant award.
2. Require the use of persistent identifiers, as appropriate.
3. Provide appropriate funding for data repositories.
4. Provide appropriate funding to support investigator driven data reuse.

5. Apply review criteria related to innovation in ways that respect the value of common data elements and other data standards.
6. Find ways to provide appropriate expertise in software development and data management to research laboratories, as needed, and in training programs.

For the journals:

1. Require the deposition of data in an appropriate repository as a condition of publication.
2. Require the use of persistent identifiers and/or links to the data in the published paper.

For the research community:

1. Avoid the use of data collection or data analysis methodologies that do not permit data to be shared.
2. Use common data elements or common data collection standards.

Conclusion

The arguments made in this article suggest that there is no technical reason that would prevent the use of data repositories to help understand complex diseases. Some preliminary attempts to do so suggest that there could be value in this approach. There are still many open questions about the structure of an effective data repository in this area. The answers to those questions aren't necessarily known, but research and/or infrastructure creation is ongoing to explore the various options. The data discovery question problem and the issues raised with finding ways for a user to launch useful queries in a heterogeneous data repository are probably the most important issues to be resolved now.

Creating an infrastructure that allows those who use data to provide an appropriate citation to those who provided the data is also an important concern. The issue of finding ways to give appropriate credit to those who measured data is related to creating a broad culture of data sharing in the biomedical research community. Such culture change is always challenging, and it may be necessary for the funding agencies and/or the journals to help hasten this change in the research community. The proposed database for complex diseases would mean an even larger culture change for the research community since the research participant would potentially have a much larger role in making data available and in data analysis.

A robust data discovery index, infrastructure(s) to hold data, meaningful data standards, and a change in culture to expect data sharing as part of the scientific process would solve the problems related to data sharing for complex diseases. The NIMH Data Archive has begun tackling many of those issues and the early results suggest that meaningful data reuse can occur. We are making progress on understanding whether data repositories can help uncover a biomarker like the C-peptide, but it is still a little too early to know whether this approach will be successful in helping to tackle a very difficult problem.

Acknowledgments

Support for and clearance of this manuscript was provided by the National Institute of Mental Health. The views expressed do not necessarily represent the views of the NIMH, the National Institutes of Health, the Department of Health and Human Services, or the United States Government. The author would like to thank Michelle Freund, Fred Friedman, Dan Hall, Jeff Muller, and Martin Wiener for helpful comments on early drafts of this manuscript.

References

- Aguirre GK. FIASCO, VoxBo, and MEDx: Behind the Code. *NeuroImage*. 2012; 62(2):765–767. [PubMed: 22348882]
- Barrett T, Edgar R. Gene Expression Omnibus (GEO): Microarray Data Storage, Submission, Retrieval, and Analysis. *Methods Enzymol*. 2006; 411:352–369. [PubMed: 16939800]
- Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO: Archive for Functional Genomics Data Sets – Update. *Nucleic Acids Res*. 2013; 41(Database Issue):D991–D995. [PubMed: 23193258]
- Beberg AL, Ensign DL, Jayachandran G, Khaliq S, Pande VS. Folding@home: Lessons from eight Years of Volunteer Distributed Computing. *Parallel & Distributed Processing, 2009 IPDPS 2009 IEEE International Symposium on*. 2009:1–8.
- Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res*. 2015; 43(Database Issue):D30–D35. [PubMed: 25414350]
- Berman HM. The Protein Data Bank: A Historical Perspective. *Acta Cryst Sect A*. 2008; A64:88–95. [PubMed: 18156675]
- Biswal BB, Mennes M, Zuo XN, Gohel S, Kelly C, Smith SM, Beckmann CF, Adelstein JS, Buckner RL, Colcombe S, Dogonowski AM, Ernst M, Fair D, Hampson M, Hoptman MJ, Hyde JS, Kiviniemi VJ, Kotter R, Li SJ, Lin CP, Lowe MJ, Mackay C, Madden DJ, Madsen KH, Margulies DS, Mayberg HS, McMahon K, Monk CS, Mostofsky SH, Nagel BJ, Pekar JJ, Peltier SJ, Petersen SE, Riedl V, Rombouts SA, Rypma B, Schlaggar BL, Schmidt S, Seidler RD, Siegle GJ, Sorg C, Teng GJ, Veijola J, Villringer A, Walter M, Wang L, Weng XC, Whitfield-Gabrieli S, Williamson P, Windischberger C, Zang YF, Zhang HY, Castellanos FX, Milham MP. Toward Discovery Science of Human Brain Function. *Proc Natl Acad Sci U S A*. 2010; 107:4734–4739. [PubMed: 20176931]
- Brunger AT. Free R Value: A Novel Statistical Quantity for Assessing the Accuracy of Crystal Structures. *Nature*. 1992; 355(6359):472–475. [PubMed: 18481394]
- Buhrmester M, Kwang T, Gosling SD. Amazon's Mechanical Turk: A New source of Inexpensive yet High-quality Data? *Perspect Psychol Sci*. 2011; 6(1):3–5. [PubMed: 26162106]
- Cheung KH, Prud'hommeaux E, Wang Y, Stephens S. Semantic Web for Health Care and Life Sciences: A Review of the State of the Art. *Brief Bioinform*. 2009; 10(2):111–113. [PubMed: 19304871]
- Collins FS, Patrinos A, Jordan E, Chakravarti A, Gesteland R, Walters L. New Goals for the U.S. Human Genome Project: 1998-2003. *Science*. 1998; 282(5389):682–689. [PubMed: 9784121]
- Collins FS, Tabak LA. Policy: NIH Plans to Enhance Reproducibility. *Nature*. 2014; 505(7485):612–613. [PubMed: 24482835]
- Conway KP, Vullo GC, Kennedy AP, Finger MS, Agrawal A, Bjork JM, Farrer LA, Hancock DB, Hussong A, Wakim P, Huggins W, Hendershot T, Nettles DS, Pratt J, Maiese D, Junkins HA, Ramos EM, Strader LC, Hamilton CM, Sher KJ. Data compatibility in the addiction sciences: An examination of measure commonality. *Drug Alcohol Depend*. 2014; 141:153–158. [PubMed: 24954640]
- Cooper S, Khatib F, Treuille A, Barbero J, Lee J, Beenen M, Leaver-Fay A, Baker D, Popovic Z. FoldIt players. Predicting Protein Structures with a Multiplayer Online Game. *Nature*. 2010; 446:756–760. [PubMed: 20686574]
- Curtis LH, Weiner MG, Boudreau DM, Cooper WO, Daniel GW, Nair VP, Raebel MA, Beaulieu NU, Rosofsky R, Woodworth TS, Brown JS. Design Considerations, Architecture, and Use of the Mini-Sentinel Distributed Data System. *Pharmacoepidemiol Drug Saf*. 2012; 21(Suppl 1):23–31. [PubMed: 22262590]

- DeMauro, A., Greco, M., Grimaldi, M. What is Big Data? A Consensual Definition and Review of Key Research Topics; 4th International Conference on Integrated Information; Madrid. 2014.
- Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene Expression and Hybridization Array Data Repository. *Nucleic Acids Research*. 2002; 30(1):207–210. [PubMed: 11752295]
- Edgar R, Barrett T. NCBI GEO Standards and Services for Microarray Data. *Nat Biotechnol*. 2006; 24(12):1471–1472. [PubMed: 17160034]
- Fitzgerald, PMD., Westbrook, JD., Bourne, PE., McMahon, B., Watenpugh, KD., Berman, HM. International Tables for Crystallography, Vol. G. Definition and Exchange of Crystallographic Data. In: Hall, SR., McMahon, B., editors. *Macromolecular Dictionary (mmCIF)*. Vol. ch. 4.5. Dordrecht: Springer; 2005. p. 295–443.
- Foster JA, Neufeld KA. Gut-Brain Axis: How the Microbiome Influences Anxiety and Depression. *Trends Neurosci*. 2013; 36(5):305–312. [PubMed: 23384445]
- Friedman L, Glover GH. Report on a Multicenter fMRI Quality Assurance Protocol. *J Magn Reson Imaging*. 2006; 23(6):827–839. [PubMed: 16649196]
- Friend SH, Norman TC. Metcalfe's Law and the Biology Information Commons. *Nature Biotechnology*. 2013; 31:297–303.
- Gardner D, Akil H, Ascoli GA, Bowden DM, Bug W, Donohue DE, Goldberg DH, Grafstein B, Grethe JS, Gupta A, Halavi M, Kennedy DN, Marengo L, Martone ME, Miller A, Müller HM, Robert A, Shepherd GM, Sternberg PW, VanEssen DC, Williams RW. The Neuroscience Information Framework: A Data and Knowledge Environment for Neuroscience. *Neuroinformatics*. 2008; 6(3): 149–60. [PubMed: 18946742]
- Geer RC, Sayers EW. Entrez: Making Use of its Power. *Brief Bioinform*. 2003; 4(2):179–184. [PubMed: 12846398]
- Gershon RC, Rothrock NE, Hanrahan RT, Jansky LJ, Harniss M, Riley W. The development of a clinical outcomes survey research application: Assessment Center. *Quality of Life Research*. 2010; 19(5):677–85. [PubMed: 20306332]
- The Governing Council of the Organization for Human Brain Mapping. *Neuroimaging Databases*. Science. 2001; 292:1672–1676.
- Green ED, Watson JD, Collins FS. Twenty-five Years of Big Biology. *Nature*. 2015; 526(7571):29–31. [PubMed: 26432225]
- Grinnon ST, Miller K, Marler JR, Lu Y, Stout A, Odenkirchen J, Kunitz S. National Institute of Neurological Disorders and Stroke Common Data Element Project – Approach and Methods. *Clin Trials*. 2012; 9(2):322–329. [PubMed: 22371630]
- Groom CR, Allen FH. The Cambridge Structural Database in Retrospect and Prospect. *Angew Chem Int Ed Engl*. 2014; 53(3):662–671. [PubMed: 24382699]
- Guggenberger R, Nanz D, Bussmann L, Chhabra A, Fischer MA, Hodler J, Pfirrmann CW, Andreisek G. Diffusion Tensor Imaging of the Median Nerve at 3.0 T Using Different MR Scanners: Agreement of FA and ADC Measurements. *Eur J Radiol*. 2013; 82(10):e590–e596.
- Gunter JL, Bernstein MA, Borowski BJ, Ward CP, Britson PJ, Felmlee JP, Schuff N, Weiner M, Jack CR. Measurement of MRI Scanner Performance with the ADNI Phantom. *Med Phys*. 2009; 36(6): 2913–2205.
- Hall D, Huerta MF, McAuliffe MJ, Farber GK. Sharing Heterogeneous Data: The National Database for Autism Research. *Neuroinform*. 2012; 10:331–339.
- Hamilton CM, Strader LC, Pratt JG, Maiese D, Hendershot T, Kwok RK, Hammond JA, Huggins W, Jackman D, Pan H, Nettles DS, Beaty TH, Farrer LA, Kraft P, Marazita ML, Ordovas JM, Pato CN, Spitz MR, Wagener D, Williams M, Junkins HA, Harlan WR, Ramos EM, Haines J. The PhenX Toolkit: Get the Most From Your Measures. *Am J Epidemiol*. 2011; 174(3):253–260. [PubMed: 21749974]
- Hibar DP, et al. Common Genetic Variants Influence Human Subcortical Brain Structures. *Nature*. 2015; 520:224–229. [PubMed: 25607358]
- Husser CS, Buchhalter JR, Raffo OS, Shabo A, Brown SH, Lee KE, Elkin PL. Standardization of Microarray and Pharmacogenomics Data. *Methods Mol Biol*. 2006; 316:111–157. [PubMed: 16671403]

- Jack CR Jr, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, Borowski B, Britson PJ, Whitwell J, Ward C, Dale AM, Felmlee JP, Gunter JL, Hill DL, Killany R, Schuff N, Fox-Bosetti S, Lin C, Studholme C, DeCarli CS, Krueger G, Ward HA, Netzer GJ, Scott KT, Mallozzi R, Blezek D, Levy J, Debbins JP, Fleisher AS, Albert M, Green R, Bartzokis G, Glover G, Mugler J, Weiner MW. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI Methods. *J Magn, Reson Imaging*. 2008; 27(4):685–691. [PubMed: 18302232]
- Jagadish HV, Gehrke J, Labrinidis A, Papakonstantinou Y, Patel JM, Ramakrishnan R, Shahabi C. Big Data and Its Technical Challenges. *Communications of the ACM*. 2014; 57(7):86–94.
- Jahanshad N, Kochunov PV, Sprooten E, Mandl RC, Nichols TE, Almasy L, Blangero J, Brouwer RM, Curran JE, deZubicaray GI, Duggirala R, Fox PT, Hong LE, Landman BA, Martin NG, McMahon KL, Medland SE, Mitchell BD, Olvera RL, Peterson CP, Starr JM, Sussmann JE, Toga AW, Wardlaw JM, Wright MJ, HulshoffPol HE, Bastin ME, McIntosh AM, Deary IJ, Thompson PM, Glahn DC. Multi-site genetic analysis of diffusion images and voxelwise heritability analysis: a pilot project of the ENIGMA-DTI working group. *Neuroimage*. 2013; 81:455–69. [PubMed: 23629049]
- Johnson SB, Whitney G, McAuliffe M, Wang H, McCreedy E, Rozenblit L, Evans CC. Using Global Unique Identifiers to Link Autism Collections. *J Am Med Inform Assoc*. 2010; 17:689–695. [PubMed: 20962132]
- Jorgenson LA, Newsome WT, Anderson DJ, Bargmann CI, Brown EN, Deisseroth K, Donoghue JP, Hudson KL, Ling GSF, MacLeish PR, Marder E, Normann RA, Sanes JR, Schnitzer MJ, Sejnowski TJ, Tank DW, Tsien RY, Ugurbil K, Wingfield JC. The BRAIN Initiative: Developing Technology to Catalyze Neuroscience Discovery. *Phil Trans R Soc B*. 2015; 370:20140164. [PubMed: 25823863]
- Kaye J, Curren L, Anderson N, Edwards K, Fullerton SM, Kanellopoulou N, Lund D, MacArthur DG, Msczaloni D, Shepherd J, Taylor PL, Terry SF, Winter SF. From Patients to Partners: Participant-Centric Initiatives in Biomedical Research. *Nat Rev Genet*. 2012; 13(5):371–376. [PubMed: 22473380]
- Kochunov P, Jahanshad N, Sprooten E, Nichols TE, Mandl RC, Almasy L, Booth T, Brouwer RM, Curran JE, deZubicaray GI, Dimitrova R, Duggirala R, Fox PT, ElliotHong L, Landman BA, Lemaitre H, Lopez LM, Martin NG, McMahon KL, Mitchell BD, Olvera RL, Peterson CP, Starr JM, Sussmann JE, Toga AW, Wardlaw JM, Wright MJ, Wright SN, Bastin ME, McIntosh AM, Boomsma DI, Kahn RS, den Braber A, deGeus EJ, Deary IJ, HulshoffPol HE, Williamson DE, Blangero J, van 't Ent D, Thompson PM, Glahn DC. Multi-site study of additive genetic effects on fractional anisotropy of cerebral white matter: Comparing meta and mega-analytical approaches for data pooling. *Neuroimage*. 2014; 95:136–50. [PubMed: 24657781]
- Kodama Y, Shumway M, Leinonen R. International Nucleotide Sequence Database Collaboration. The Sequence Read Archive: Explosive Growth of Sequence Data. *Nucleic Acids Res*. 2012; 40(Database Issue):D54–6. [PubMed: 22009675]
- Komorowski, M. Hard Drive Cost Per Gigabyte. 2014. <http://www.mkomo.com/cost-per-gigabyte-update>
- Krumm N, Turner TN, Baker C, Vives L, Mohajeri K, Witherspoon K, Raja A, Coe BP, Stessman HA, He ZX, Leal SM, Bernier R, Eichler EE. Excess of Rare, Inherited Truncating Mutations in Autism. *Nature Genetics*. 2015; published online 11 May 2015. doi: 10.1038/ng.3303
- Lee H, Marvin AR, Watson T, Piggot J, Law JK, Law PA, Constantino JN, Nelson SF. Accuracy of Phenotyping of Autistic Children Based on Internet Implemented Parent Report. *Am J Met Genet B Neuropsychiatr Genet*. 2010; 0(6):1119–1126.
- Leung, L. How Much Data does X Store?. 2014. <http://techexpectations.org/2014/05/17/how-much-data-does-x-store/>
- Levine SZ, Rabinowitz J, Uher R, Kapur S. Biomarkers of Treatment Outcome in Schizophrenia: Defining a Benchmark for Clinical Significance. *Eur Neuropsychopharmacol*. 2015; Epub ahead of print. doi: 10.1016/j.euroneuro.2015.06.008
- Li M, Luo XJ, Rietschel M, Lewis CM, Mattheisen M, Müller-Myhsok B, Jamain S, Leboyer M, Landén M, Thompson PM, Cichon S, Nöthen MM, Schulze TG, Sullivan PF, Bergen SE, Donohoe G, Morris DW, Hargreaves A, Gill M, Corvin A, Hultman C, Toga AW, Shi L, Lin Q, Shi H, Gan L, Meyer-Lindenberg A, Czamara D, Henry C, Etain B, Bis JC, Ikram MA, Fornage M, Döbetz S,

Launer LJ, Seshadri S, Erk S, Walter H, Heinz A, Bellivier F, Stein JL, Medland SE, Arias Vasquez A, Hibar DP, Franke B, Martin NG, Wright MJ, Su B. MooDS Bipolar Consortium; Swedish Bipolar Study Group; Alzheimer's Disease Neuroimaging Initiative; ENIGMA Consortium; CHARGE Consortium. Allelic differences between Europeans and Chinese for CREB1 SNPs and their implications in gene expression regulation, hippocampal structure and function, and bipolar disorder susceptibility. *Mol Psychiatry*. 2014; 19(4):452–61. [PubMed: 23568192]

Luo XZ, Kennedy DN, Cohen Z. Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC) Resource Announcement. *Neuroinformatics*. 2009; 7(1):55–56. [PubMed: 19184562]

Marcus DS, Olsen TR, Ramaratnam M, Buckner RL. The Extensible Neuroimaging Archive Toolkit (XNAT): An Informatics Platform for Managing, Exploring, and Sharing Neuroimaging Data. *Neuroinformatics*. 2007; 5:11–34. [PubMed: 17426351]

Margolis R, Derr L, Dunn M, Huerta M, Larkin J, Sheehan J, Guyer M, Green ED. The National Institutes of Health's Big Data to Knowledge (BD2K) Initiative: Capitalizing on Biomedical Big Data. *J Am Med Inform Assoc*. 2014; 21:957–957. [PubMed: 25008006]

Masseroli M, Mons B, Bongcam-Rudloff E, Ceri S, Kel A, Rechenmann F, Lisack F, Roman P. Integrated Bio-Search: Challenges and Trends for the Integration, Search, and Comprehensive Processing of Biological Information. *BMC Bioinformatics*. 2014; 151(Suppl 1):S2. [PubMed: 24564249]

McCray AT, Trevvett P, Frost HR. Modeling the Autism Spectrum Disorder. *Neuroinformatics*. 2014; 12(2):291–305. [PubMed: 24163114]

Mennes M, Biswal BB, Castellanos FX, Milham MP. Making Data Sharing Work: The FCP/INDI Experience. *NeuroImage*. 2013; 82:683–691. [PubMed: 23123682]

Min H, Ohira R, Collins MA, Bondy J, Avis NE, Tchuvatkina O, Courtney PK, Moser RP, Shaikh AR, Hesse BW, Cooper M, Reeves D, Lanese B, Helba C, Miller SM, Ross EA. Sharing Behavioral Data Through a Grid Infrastructure Using Data Standards. *J Am Med Inform Assoc*. 2014; 21(4): 642–649. [PubMed: 24076749]

Paraiso-Medina S, PerezRey D, Bucur A, Claerhout B, Alonso-Calvo R. Semantic Normalization and Query Abstraction Based on SNOMED-CT and HL7: Supporting Multicentric Clinical Trials. *IEEE J Biomed Health Inform*. 2015; 19(3):1061–1067. [PubMed: 25248204]

Piwowar HA. Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data. *PLoS ONE*. 2011; 6(7):e18657. [PubMed: 21765886]

Poldrack RA, Barch DM, Mitchell JP, Wager TD, Wagner AD, Devlin JT, Cumba C, Kpyejo O, Milham MP. Toward open sharing of task-based fMRI data: the OpenfMRI project. *Front Neuroinform*. 2013; 7(12):1–12. [PubMed: 23386828]

Savage CJ, Vickers AJ. Empirical Study of Data Sharing by Authors Publishing in PLoS Journals. *PLoS ONE*. 2009; 4(9):e7078. [PubMed: 19763261]

Schadt EE, Lindeman MD, Sorenson J, Lee L, Nolan GP. Computing Solutions to Large-Scale Data Management and Analysis. *Nature Reviews Genetics*. 2010; 11:647–657.

Scott A, Courtney W, Wood D, de la Garza R, Land S, King M, Wang R, Roberts J, Turner JA, Calhoun VD. COINS: An Innovative Informatics and Neuroimaging Tool Suite Built for Large Heterogeneous Datasets. *Front Neuroinform*. 2011; 5(33):1–15. [PubMed: 21472085]

Smith SM, Vidaurre D, Beckmann CF, Glasser MF, Jenkinson M, Miller KL, Nichols TE, Robinson EC, Salimi-Khorshidi G, Woolrich MW, Barch DM, Ugurbil K, Van Essen DC. Functional Connectomics from Resting-State fMRI. *Trends in Cognitive Sciences*. 2013; 17(12):666–680. [PubMed: 24238796]

Smith TF. The History of Genetic Sequence Databases. 1990; 6:701–707.

Stein JL, et al. Identification of common variants associated with human hippocampal and intracranial volumes. *Nature Genetics*. 2012; 44(5):552–561. [PubMed: 22504417]

Strasser BJ. GenBank – Natural History in the 21st Century? *Science*. 2008; 322(5901):537–538. [PubMed: 18948528]

Strasser BJ. Collecting, Comparing, and Computing Sequences: The Making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954-1965. *J Hist Biol*. 2010; 43(4):623–660. [PubMed: 20665074]

- Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE. Big Data: Astronomical or Genomical? PLoS Biol. 2015; 13(7):e1002195. [PubMed: 26151137]
- Suber P. An Open Access Mandate for the National Institutes of Health. Open Med. 2008; 2(2):e39–e41. [PubMed: 21602938]
- Teeters JL, Harris KD, Milman J, Olshausen BA, Sommer FT. Data Sharing for Computational Neuroscience. Neuroinformatics. 2008; 6(1):47–55. [PubMed: 18259695]
- Thompson PM, et al. The ENIGMA Consortium: Large-Scale Collaborative Analysis of Neuroimaging and Genetic Data. Brain Imaging and Behavior. 2014; 8(2):153–182. [PubMed: 24399358]
- Torgerson CM, Quinn C, Dinov I, Liu Z, Petrosyan P, Pelphrey K, Haselgrove C, Kennedy DN, Toga AW, Van Horn JD. Interacting with the National Database for Autism Research (NDAR) via the LONI Pipeline Workflow Environment. Brain Imaging and Behavior. 2015; 9:89–103. [PubMed: 25666423]
- Tregarthen JP, Lock J, Darcy AM. Development of a Smartphone Application for Eating Disorder Self-Monitoring. Int J Eat Disord. 2015; 48(7):972–982. [PubMed: 26213130]
- Turner JA, Laird AR. The Cognitive Paradigm Ontology: Design and Application. Neuroinformatics. 2012; 10:57–66. [PubMed: 21643732]
- Uher R, Tansey KE, Malki K, Perlis RH. Biomarkers Predicting Treatment Outcome in Depression: What is Clinically Significant? Pharmacogenomics. 2012; 13(2):233–240. [PubMed: 22256872]
- Utturkar SM, Klingeman DM, Land ML, Schadt CW, Doktycz MJ, Pelletier DA, Brown SD. Evaluation and validation of *de novo* and hybrid assembly techniques to derive high-quality genome sequences. Bioinformatics. 2014; 30(19):2709–2716. [PubMed: 24930142]
- VanBuecken DE, Greenbaum CJ. Residual C-peptide in Type 1 diabetes: What Do We Really Know? Pediatr Diabetes. 2014; 15(2):84–90. [PubMed: 24645775]
- Van Essen DC, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K. for the WU-Minn HCP Consortium. The WU-Minn Human Connectome Project: An Overview. NeuroImage. 2013; 80:62–79. [PubMed: 23684880]
- Van Horn JD, Toga AW. Is it Time to Re-prioritize Neuroimaging Databases and Digital Repositories? NeuroImage. 2009; 47:1720–1734. [PubMed: 19371790]
- Van Horn JD, Gazzaniga MS. Why Share Data? Lessons learned from the fMRIDC. NeuroImage. 2013; 82:677–682. [PubMed: 23160115]
- Walker L, Curry M, Nayak A, Lange N, Pierpaoli C. Brain Development Cooperative Group. A Framework for the Analysis of Phantom Data in Multicenter Diffusion Tensor Imaging Studies. Hum Brain Mapp. 2013; 34(10):2439–2454. [PubMed: 22461391]
- Wicks P, Massagli M, Frost J, Brownstein C, Okun S, Vaughan T, Bradley R, Heywood J. Sharing Health Data for Better Outcomes on PatientsLikeMe. J Med Internet Res. 2010; 12(2):e19. [PubMed: 20542858]

Abbreviations

CDE	common data element
CRCNS	collaborative research in computational neuroscience
fMRIDC	fMRI data center
IAN	Interactive Autism Network
NDA	NIMH Data Archive
NDAR	National Database for Autism Research
NDCT	National Database of Clinical Trials Related to Mental Illness

PDB	Protein Data Bank
PLM	Patients Like Me
RDoC db	Research Domain Criteria database

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Highlights

There are many challenges to developing drugs for complex diseases. This review explores whether data archives can be effectively used to help find biomarkers that are related to defining sub-populations affected by a complex disease. The features of successful data repositories are outlined. Remaining challenges as well as attempts to solve those challenges are covered.