NEWS

New scheme ramps up efforts to aggregate autism data

BY **EMILY SINGER**

19 DECEMBER 2011

The nation’s largest database of autism research is about to get even bigger, thanks to a new

partnership between the **National Database for Autism Research** (NDAR) and the **Autism Genetic Research Exchange** (AGRE), a database of 2,500 families. Beginning in 2012, NDAR also plans to house data from 90 percent of autism research in the United States that is focused on humans.

NDAR is a data repository sponsored by the National Institutes of Health (NIH), and already houses data from 25,000 participants. AGRE is a collection of data from so-called multiplex families, which have more than one child with autism, and is funded by the autism science and advocacy organization Autism Speaks.

The partnership is expected to result in the largest source of genetic information, phenotypic, or observable, traits, and clinical and medical imaging data on autism in the world, enabling researchers to access and analyze information on large numbers of people with the disorder.

That aggregated data is especially important given autism's heterogeneity.

"You can't have enough data to really be able to investigate in a serious way the variability and broad range of phenotypes that we see in this disorder," says **Helen Tager-Flusberg**, professor of psychology at Boston University and president of the International Society for Autism Research. "I think we are going to see more and more analyses of these datasets that will give us much richer appreciation for the variability across different ages."

Data-sharing:

Launched in 2007, NDAR grew out of requests from Autism Speaks and other organizations to create a repository of all of NIH's autism research. In addition to AGRE, NDAR already provides access to data from Autism Speaks' **Autism Tissue Program** as well as a subset of data — from about 7,500 people — from the **Interactive Autism Network. The network houses self-reported information from more than 40,000 people with autism and their families.**

In its 2010 strategic plan, the Interagency Autism Coordinating Committee, part of the U.S. Department of Health and Human Services, set a goal for NDAR to bring together 90 percent of the nation's human autism research. The committee recommended a budget of $6,800,000 over two years.

"It's a pretty tall order," says NDAR manager **Dan Hall**. Aside from a handful of rare disorders, no single disease community has brought together the field's research so comprehensively, he says.

Scientists who win NIH funding for autism research are required to share their data on human participants through NDAR. NIH-funded grants for autism research encompass about 100,000 people — including those with autism and controls — and information on all of those individuals will be entered into NDAR. Although there are other data-sharing initiatives at the NIH, none intended

for a common disorder is this comprehensive.

“The potential gains are enormous,” says Leon Rozenblit, president of **Prometheus Research**, which manages **SFARIBase**, repository of data from the **Simons Simplex Collection**. Th collection, funded by SFARI.org’s parent organization, includes genetic samples and data from 2,700 families. “If we do it right, we will get new scientific insights without having to collect more data.” SFARI Base ultimately aims to become accessible via NDAR as well.

Rather than storing the information from other databases, NDAR and collaborating organizations have developed ways to allow scientists to access data from various sources through the NDAR portal. “Our goal is to create one entry point to access all this data, and I think we’re really on track with that goal,” says Hall.

Unique identifiers:

One of the major technical problems in combining databases is to avoid overlap: The same people may have participated in multiple studies logged in different databases, and researchers don’t want to mistakenly treat this as information from different people.

To solve this issue, NDAR and collaborators introduced global unique identifiers, called **GUIDs**, to tag individuals across different studies and databases while maintaining their anonymity.

Virtually all autism repositories are using GUIDs, says **Greg Farber**, director of both NDAR and the NIH’s Office of Technology Development and Coordination. “I think that goes to show that the autism research community really does believe in what NDAR is trying to do,” he says.

The NDAR team is also working on ways to make the growing volume of data, now approaching 100 terabytes, easy to access and analyze. “Most labs don’t have the ability to download two terabytes a day,” says Hall, so storing data locally isn’t a realistic option.

Instead, Hall says, the plan is to make the data available via ‘cloud computing.’ This refers to the use of commercial services, such as one offered by Amazon.com, to remotely store information and software. Data analysis is also done remotely, via the network.

Researchers will also need to develop tools, including data visualization or more intuitive search functions, to explore the information housed in these databases.

Indeed, the entire scheme is still something of an experiment. “If lots of different sorts of data are being measured in different laboratories in different ways,” says Farber, “is it really possible to bring it all together into a central database that would allow researchers to conduct computational experiments that they wouldn’t have been able to do otherwise?”

The answer should come soon and in turn help to solve one of the non-technical challenges for NDAR. “The biggest challenge of the system is to gain the trust of researchers for them to use this resource,” says Hall.