NEWS

Researchers debut unique identifiers for study participants

BY **DEBORAH RUDACILLE**

2 SEPTEMBER 2010

GUID vibrations: Tracking participants across studies may help researchers more efficiently validate their data

Researchers have devised a system to assign a unique identifier to each participant in an autism study. This Global Unique Identifier, or GUID, allows investigators to see which other studies participants have enrolled in, preventing them from using data from the same participants to validate the original results.

"Beyond that, GUID gives you the opportunity to fill in information from different studies that might provide new insights," says **Edwin Cook, Jr.**, director of the Laboratory of Developmental Neuroscience at the University of Illinois, Chicago. Cook was not involved in GUID's development.

GUID is the result of a three-year collaboration by the National Database for Autism Research (**NDAR**) at the National Institutes of Health (NIH), a team of researchers from Columbia University, the **Simons Foundation** (SFARI.org's parent organization) and **Prometheus Research**, a data management company based in Connecticut.

To generate a GUID, a researcher enters four variables into a server, which in three seconds returns a unique alphanumerical code. GUID can help scientists not only track an individual's

participation across studies, but link together data — genomic, behavioral and anatomical — from the same person across multiple collections, researchers say. The approach will be described in an upcoming paper in the ***Journal of the American Medical Informatics Association***.

"This gives families a great deal of hope and excitement that the research community is willing to come together and share data," says **Paul Law**, director of the Interactive Autism Network (IAN), based at the Kennedy Krieger Institute in Baltimore, and the parent of a child with autism. "It's taken a long time for that to happen."

Not for lack of enthusiasm, say others. "People want to share but there really hasn't been a mechanism for doing it correctly," says Cook.

Imperfect systems:

In the 1990s, researchers associated with the **Collaborative Programs of Excellence in Autism** — a five-year, $45 million genetics research program — put into place a rudimentary identification system. But it was far from perfect.

"If you found what seemed to be a duplication, you would contact the researcher," Cook recalls. "At the time I remember saying, 'Somebody should come up with a better idea.'"

Until a couple of years ago, researchers couldn't create global identifiers because there was no database or format for entering identifying information in a way that protects the privacy of study participants**.** Institutional review boards rejected any approach that would store personal information in a central system.

The new approach surmounts those obstacles by using four simple variables — birth name, birth date, gender and city of birth. The four items are combined through a process called hashing, scrambling them together like ingredients in an omelet. Once hashed, the original variables cannot be retrieved. The server then compares the hash codes against an internal database, which checks for duplicates.

The technical challenges of developing the technology were straightforward, says **Matthew McAuliffe**, chief of biomedical imaging research services in the NIH's Division of Computational Bioscience. "The biggest challenge was changing the mindset of, 'This is my data and I'm keeping it.'"

The benefits of using GUID are likely to overcome any reservations. For example, a researcher may need to link genetic data to imaging results for the same individual. "You may need 20 rare genetic findings and 20 brain scans and they need to be collected from various locations. GUID will help us find them," says Cook.

GUID will also benefit study participants, Cook says. "If someone volunteers for one study and then another, they may not have to go back and give more blood or sit in a scanner again."

Tested and true:

Researchers have already field-tested GUID in the **Simons Simplex Collection** (SSC), a genetic repository of more than 2,000 families. The project returned unique identifiers for 96 percent of children in the study and 77 percent of parents. "You needed a big new study [to test it] and [the SSC] was it," says **Stephen Johnson**, informatics director at SFARI, professor of biomedical informatics at Columbia University and primary architect of the GUID software.

Major autism research initiatives such as IAN, which stores data from more than 20,000 individuals, and the **Autism Genetic Resource Exchange** have agreed to generate GUIDs for their studies. Law says he expects a great deal of overlap between individuals in the IAN database and those in other studies once IAN's data are fed into the server.

But first, "there are also a lot of legal hurdles to overcome," Law says. That includes going back to enrolled families and asking for consent.

As of last week, NDAR's server had generated 11,870 GUIDS and 27,587 'pseudo-GUIDs,' which serve as placeholders when one or more variables is missing. IAN's database, for example, doesn't include city of birth; once that information is plugged in, the server will replace the pseudo-GUID with a real one.

NDAR investigators will be required to generate GUIDs for participants in all new studies. This virtually ensures a vast interlinked repository of data, offering NDAR users "unprecedented access" to research from around the world, says McAuliffe.

"The aggregation of the data will enable researchers to collaborate, efficiently share and validate findings to speed research," he says, "ultimately helping those affected by autism."