Dark data see the light

Ellen P. Neff

Dredging up old studies yields new insights with the help of big data analytics, but is it enough to spark more data sharing for current research?

In the late 1990s, three Mayflower trucks made the trek from the heart of New York City to the suburbs of New Jersey. At the end of the journey their contents were unloaded into a storage unit near Rutgers University, and the door was shut.

Nearly twenty years later, a group of researchers and data scientists made their own trek across the country to visit that unit in New Brunswick. The expedition was led by Adam Ferguson, an investigator at the Brain and Spinal Injury Center and the University of California San Francisco, and his post-doc at the time, Jessica Nielson.

Ferguson laughs remembering when they lifted up the door, referring to what awaited them as “a data archeology site.” The light of day revealed boxes and boxes full of reams of paper data sheets and old floppy discs and hard drives. These spoils were no longer destined to continue languishing away as relics of the past but were to be recovered, digitized, and put to new use with decidedly more modern computers and analytic techniques.

This was not an easy feat. Before any computer could take a crack at them, the data had to be dug out and processed, essentially by hand. During a second trip to Rutgers in 2017, Nielson physically scanned over 3000 paper pages. The data are now in PDF format, awaiting further harmonization with previously collected files. Ferguson hopes improving optical character recognition capabilities will help speed things up, but for now, humans remain the workhorses.

The end goal marks a new approach for spinal cord injury (SCI) research, as Ferguson brings a penchant for bioinformatics to the field. This requires data—and lots of it. With Nielson and the others he assembled for his initial forays into data

DATA ARCHEOLOGY | Time spent raiding boxes in a storage unit (left) and scanning the spoils (right) gives old data new life.

Jessica Nielson

mining, he's demonstrating the potential that computer analytics can bring to basic and preclinical research¹.

An online database is in the works that's intended to serve the larger SCI community, but building a self-sustaining system means hashing out how the researchers involved feel about a new world that hinges on openness and sharing.

New analytics meet old data

That storage unit likely contains the complete set of records from an ambitious preclinical SCI project known as MASCIS, the Multicenter Animal Spinal Cord Injury Study, which was funded by the NIH from 1993 to 1996. Led by New York University Professor Wise Young (who later relocated, with his laboratory archive, to Rutgers) and run like a clinical randomized controlled trial, MASCIS tested potential drugs to

treat SCI in rats at eight different US laboratories. Despite the years of effort and coordination between the different groups involved, the drug results were ambiguous and the trials were stopped.

MASCIS did produce important standards that have since been adopted by many SCI labs, like the Basso, Beattie, Bresnahan (BBB) score to measure outcomes and controlled ways to model different injuries, but with no promise of a therapy to push on into the clinic for human trials, the data went dark, physical copies seemingly destined to fade away in the depths of file drawers and storage bins across the country.

Until about a decade later, when Ferguson joined one of the original MASCIS centers, the Ohio State University (OSU). He had just finished his PhD in behavioral neuroscience at Texas A&M University and while completing a post-doc at OSU, he developed

Adam Ferguson

AHEAD OF THE CURVE | Adam Ferguson, University of California San Francisco, has new ideas for old data in preclinical spinal cord injury research.

an interest in neurological trauma, like SCI and traumatic brain injury, and a desire to apply analytic methods common in psychology to a new field. His advisors, Michael Beattie and Jacqueline Bresnahan, had plenty of old data available to dig into, but Ferguson needed some resources to make use of what they had to offer.

After moving to San Francisco in 2006, he applied to the NIH for a grant to build a database of preclinical SCI data. The program officer of NINDS at the time, Naomi Kleitman, admits, “He baffled me when he first starting telling me what this new grant

proposal was about.” But the idea of databases and data sharing was already growing at the agency. “Requirements at NIH existed for sharing clinical data,” Kleitman says, so “why not look at this in terms of the pre-clinical data that he had available to him?”

His RO1 application was approved and he used that funding to hire others, like Nielson, with technical data experience to help him build the initial database. But first, they had to upload the data. “This was a classic kind of semi-standardized dark-data problem,” he says, “You have these retrospective records, a lot of them are highly detailed...that are somewhere in file drawers.”

Over weeks and months, paper data sheets from OSU’s MASCIS studies were retrieved and digitized into “essentially a giant spreadsheet,” he says. Where possible, the data variables the team identified and included have been aligned to NIH’s Common Data Elements, to help make connections between different studies. Primary endpoints were captured, but so were ancillary outcomes, like animal care records and medication histories.

The resulting database, known as VISION-SCI², ultimately grew to over 4,000 animals with 2,700 curated variables from 13 different labs, along with accessory data like PDFs and images from the original studies. Nielson’s efforts pitching the idea at conferences and workshops and collecting old data that other labs were willing to share has kept the database growing, as will the

Jessica Nielson

A DATABASE, VISUALIZED | The physical archive of old data that went into the initial VISION-SCI database.

inclusion of the additional MASCIS data retrieved from Rutgers.

Making sense of such huge quantities of unrelated data may seem daunting to do with conventional univariate analyses, but it’s quite possible with the right knowledge of computer analytics. In 2015, Nielson, Ferguson, and their colleagues published the results of a topographical data analysis (TDA) in *Nature Communications*¹. TDA, Ferguson explains, is a new type of graphical data analysis that analyzes data “in a completely unsupervised, data-driven way”. The technique can overcome noise in a dataset to identify novel relationships and patterns.

“We were able to query the data in a very precise way and say, are there clusters of animals that have distinct outcomes even though they had the same injury? And the answer was yes,” Ferguson says. They could then ask what accounted for those differences.

“I was quite convinced that it was going to be due to a drug. That we were going to find a big, hidden drug effect that was in the data but hadn’t been seen before,” he says.

“It turns out, I was wrong.”

The inclusion of animal care data proved fortunate. The distinct outcomes, the TDA revealed, were due to the animals’ blood pressure at the time of injury. Animals that were hypertensive or hypotensive on the operating table had poor long-term outcomes, Ferguson says. “That was interesting. First off, this was not a hypothesis we had. This wasn’t what we were looking for. The data analytics showed us this.”

He continues, “Secondly, it’s an immediately actionable, translatable finding if it’s true...That has led to a bunch of other

Data-driven discovery - MASCIS 1996 trial

Injury condition

MAKING THE CONNECTION | TDA reveals why animals with the same injury had different outcomes. Adapted from *Nat. Commun.* 6, 8581; 2015.

studies that we're doing both clinically and preclinically in a prospective way to confirm these findings." Preliminary data from operating room medical records at a level 1 trauma center suggest they may be on the right track; Ferguson hopes those results will be ready to publish soon.

Collecting the long tail

VISION-SCI exemplified the value of collecting "long tail data" in neuroscience: as opposed to concerted, standardized "big data" collection efforts, this term is used in a recent *Nature Neuroscience* commentary to refer to all the countless datasets of various sizes, published and unpublished, that have been created by independent labs and investigators³. Combining such data can improve insight into the complete "syndromic space" of an injury, as individual animal studies provide "an incomplete glimpse into the syndrome across the full spectrum of possible injury conditions and outcome metrics," Ferguson and colleagues write³. A given dataset may be small, but the collective data are large.

"My motive for assembling this database is that I had a real clear idea of the analytics that one would apply and I needed the data as raw material," Ferguson says. His results have kindled interest. "So now that I have this role, people approach me, handing me the data, and say, 'can you do that magic thing you do, and make sense of my data for me?'"

It's a role he takes seriously, but one that's hardly sustainable for an individual with his own research objectives and career to consider. All those data need a home. With a new pilot grant from the Craig H. Neilsen Foundation, Ferguson launched a project to develop the infrastructure needed for a new, public-facing database: the Open Data Commons for Spinal Cord Injury (ODC-SCI).

An important step was getting involved with the Neuroscience Information Framework (NIF), an organization born from a NIH Blueprint Grant in 2008 to help the agency keep track of the non-conventional (i.e., non-publication) neuroscience outputs of its awardees. The SCI community has become NIF's poster child.

"Adam is one of my absolute favorite people," says NIF director Maryann Martone, "He knows this [data sharing] needs to be done...and he's got the use case."

WHERE TO LOOK | Schematic illustration of long-tail data, published in *Nat. Neurosci.* **17**, 1442–1448 (2014).

Members of the SCI community have been reflecting on the failure to translate and reproduce study results for some time. As staff at NINDS observed the spinal injury field advance in animal models, they decided to launch a series of reproducibility contracts in the early 2000s, part of the Facilities of Research Excellence-Spinal Cord Injury (FORE-SCI), which ran for two five-year cycles.

Investigators at three centers—the University of California Irvine, The Miami Project to Cure Paralysis, and the Ohio State University—attempted to independently replicate 18 different preclinical SCI studies. Just one was successful⁴.

New approaches were needed, and that including some new thinking about data. "Review articles were written and so that I think really got the attention of the spinal cord injury community, that we needed to keep trying harder and harder both in documenting our metadata and our data," says Vance Lemmon, an investigator at the Miami Project. "Adam started sharing data and he's been talking about sharing data for a long time, and I think that there just slowly evolved more of an openness."

Reproducibility concerns are not a problem unique to SCI, but Martone, having seen how different communities have been responding to these challenges, commends SCI researchers for their willingness to entertain new approaches like Ferguson's. "I use them as an example of how you appropriately deal with non-reproducible science."

At its most basic, NIF is a repository of different databases, services, and software available to neuroscientists. But as Martone and her colleagues saw database

after database come into being with a new grant, only to become stale after the funding period ended, they devised a generalized platform that specific communities could then re-use as the basis for their own system. Researchers can contribute to their discipline-specific portal, and the data then feed into the larger infrastructure as well. All at much lower cost than building an original database from scratch.

"Adam I think was one of the first people to really grasp it," Martone says, and the ODC-SCI web portal has been built on top of NIF's SCI-CRUNCH infrastructure. Once data are uploaded, they become searchable or, with the right skills and tools, analyzable. The platform allows add-ons, and Ferguson hopes to incorporate user-friendly tools into the ODC-SCI "that will allow non-data people to interface with the data" he says.

"There are some very sophisticated online visualization tools that exist in other fields. They're just not yet integrated in biomedicine. That would be my end game here. I would like to not be the gatekeeper."

Data; open and FAIR?

Martone's goal for data, which Ferguson echoes, is ultimately for it to be both open and FAIR, an acronym adopted by data sharing advocates that stands for "Findable," "Accessible," "Interoperable," and "Reusable."⁵

Using established frameworks like the infrastructure supplied by NIF can go a long way, but communities still need to buy in. And that, acknowledges ODC-SCI co-chair Karim Fouad, "takes a shift in culture."

The approach being taken is a democratic one. From an initial stakeholder

meeting, which included researchers as well as funders, at the NIH in Bethesda, MD in October 2016, three goals for a preclinical SCI database emerged⁶, says Lyn Jakeman, the current program director at NINDS overseeing SCI. First, investigators are interested in a place to register and report the results of their preclinical studies. Second, they want to be able to search and find what others are doing—and have done, whether successful or not. That motivates Fouad, who studies neuroplasticity in SCI at his own lab at the University of Alberta. He would like “to create a transparent world where all these data are available...so we can see the entire picture of research,” he says. “How many labs use this drug? How many labs did have success? What’s the difference?”

The third, “which is really where Adam’s heart is,” Jakeman says, is to have the data collected in a way that would allow big data analytics—by SCI researchers as well as broader bioinformaticians—to use the data to come up with new hypotheses. That stands to benefit the animals used too (**Box 1**).

But sharing data outside of conventional publications is a new frontier for many; and better recording of who’s done what and why is different from opening up file drawers and sharing unpublished work. Minimum reporting standards could help the first point. A few years ago, Lemmon wrangled 80 authors to come up with minimum reporting standards for published papers that could be collected in a web-friendly way and indexed in an online database, in an effort to make it easier for researchers to find what others have done^{7, 8}.

Lemmon has been working with Ferguson and Fouad to determine how ODC-SCI might incorporate that kind of metadata from its users. “We would like to be able to merge our stuff and he and Karim want to use our annotation tools to make it easier for people to at least describe how they did their experiments when they are reporting their data,” Lemmon says.

To ease data entry, Fouad, with his postdoc Abel Torres-Espín, are helping Ferguson and NIF create what they hope will be an intuitive data upload platform for users. “Our vision is that this platform will become something like a mobile lab notebook,” he says. As for what data are included, “we decided to start small,” Fouad says. Though still in progress, rats are the first

BOX 1 | DATA SHARING & REDUCING ANIMAL USE

The 3Rs of animal research encourages scientists to reduce the number of lab animals used. Better data sharing could make a big difference.

A notable observation from the FORE-SCI replication studies, Kleitman explains, was “that some of these original studies are not using large enough group sizes... Reduction, replacement are part of what people are being told from the animal care point of view but in fact, I think that ultimately wastes animals in the long-term.” But by openly sharing preclinical data, “you could in fact decrease the need to do some sorts of animal studies because the data already exist from another study,” she says. New hypotheses can be generated and refined from a database before any additional animal experiments are conducted.

That’s especially true in the case of data with negative or inconclusive results that tend not to be published in conventional venues like journals. These data are still important to share. “Can you imagine how many lives of animals we’d save” asks Fouad, “when we know that this drug has been tested 100 times but nothing came out?”

model they are thinking about, and recovery the main outcome, though additional variables and parameters added over time will continue to make the upload options more and more comprehensive.

Ferguson, having seen the value of animal care data from MASCIS, hopes that will eventually make its way into the database as well. “I think a lot of animal researchers view the animal health records as a regulatory requirement rather than a primary data point [but] the opportunities there I think are huge,” he says. “If the animal care records were in a format that was easily reconciled with the primary data outcomes, this seems like it would be really translationally relevant. If you think about human records, most of them are in an electronic health record these days. They are digital... and it’s kind of funny that we bench scientists don’t necessarily consider it primary data in animals.”

But when talking about sharing actual data of any kind, uncertainty can linger. Another meeting was held recently as a satellite workshop before the annual Society for Neuroscience meeting in Washington, DC. In a small conference room in the Renaissance Hotel, about thirty participants met to discuss the road ahead. Questions remain. Who owns the data? Who has access? What about credit, or sensitive data? Jakeman and Ferguson collected feedback during the meeting, and hope to put together a clearer picture of the guidance needed to send around for the community to review.

“I think it’s important to note that we’re still in early days,” Jakeman cautions. But “the enthusiasm is there, the readiness is

there, the excitement is there to make more value out of both the resources that have gone into the research and the animals that have been used...I’m pretty excited about what’s happening in the SCI community.”

The clock is ticking. With a beta release planned early 2018, Ferguson is ready to push the button to launch the ODC-SCI for testing by registered users. He recognizes that its success won’t be measured by his motives alone. “The question is, what does the field want? And if the Open Data Commons is to be successful, it will be what the field wants. Not what I want, not what Vance wants, not what journals or the NIH wants. It will be the overlap of everybody,” he says.

The beta database will initially be seeded with what he and his team have collected from the MASCIS datasets and what others have donated to them along the way. But as one workshop attendee noted, those involved so far have been a self-selected group that are open to the idea of data sharing. What new data the larger community will contribute on its own ultimately remains to be seen.

1. Nielson, J.L. *et al. Nat. Commun.* **6**, 8581 (2015).
2. Nielson, J.L. *et al. J. Neurotrauma* **31**, 1789–1799 (2014).
3. Ferguson, A.R., Nielson, J.L., Cragin, M.H., Bandrowski, A.E. and Martone, M.E. *Nat. Neurosci.* **17**, 1442–1447 (2014).
4. Steward, O., Popovich, P.G., Dietrich, W.D. and Kleitman, N. *Exp. Neurol.* **233**, 597–605 (2012).
5. Wilkinson, M.D. *et al. Sci Data* **3**, 160018 (2016).
6. Callahan, A. *et al. Exp. Neurol.* **295**, 135–143 (2017).
7. Lemmon, V.P. *et al. J. Neurotrauma* **31**, 1354–1361 (2014).
8. Callahan, A. *et al. Database* **2016**, 1–13 (2016).