figshare • Neuroscience

Tips for early career researchers on making data available

An interview with Kirstie Whitaker, University of Cambridge

Key Points

- **Trust, but verify**
  You should be able to check my work using the data.
- **GitHub & figsharet**
  I store my data on both but use figshare for persistence.
- **Sharing private links**
  Using these links to share data with reviewers.
- **Make what you can available**
  Start with data you own, slides, and posters.

About Kirstie

I study adolescent brain development, so I put teenagers in MRI scanners and look at how the brain changes through development. I'm trying to understand what a typically-developing brain looks like and what a brain at risk of developmental health disorders looks like. This work is part of the Neuroscience in Psychiatry Network consortium.

The ultimate goal would be to try and identify people who are at risk of having symptoms of depression or schizophrenia, which tend to have their first onsets in adolescence. We think it's a developmental disorder, so it's something that continues to grow and at some point tips over into symptoms.

We believe there are early-life predictors, so we look at the brain to try and see if we can identify those people. Ideally, we would like to be able to provide treatment to prevent the first set of symptoms.

"I feel that the open science movement has sometimes had a rhetoric of doing it perfectly or not at all. We need to start being a little more inclusive - do what you can and do what works for you."

We published a paper this summer in PNAS that investigated the brain as a network, using MRI images. Instead of looking at each different part of the brain individually, you look at how the different parts of the brain relate to each other. We create a structural connectome and showed that the hubs of that network are the areas that are continuing to develop in late adolescence. Those are the parts of the brain's network that are really well connected to other parts of the brain. They also happen to be in parts of the brain that we call association cortex which are responsible for generating complex thoughts and understanding.

What data exists in this field of research and is it sensitive?

If you have an original MRI scan and you look at someone's face, you may be able to identify someone. One of the steps we do when processing the data is to take away everything that is not the brain - neck, face, cheek, scalp, everything like that. Once you only have a brain, it would be almost impossible to identify the person.

One of the biggest issues is that there is lots of data. The volume of data is a challenge to work with. We use a high-performance computing cluster to do the processing and to store all of our data.

The data for our paper, which is available on figshare, does not contain very much behavioural information because that wasn't the focus of the paper. We shared age and gender, but it's difficult to release a lot of information about a person. If we released all 300 questionnaire items we asked, you might be able to build a bit of a picture, but they're still coded as 1s and 0s. The interview data is almost certainly something that won't be released to the public in its raw form.

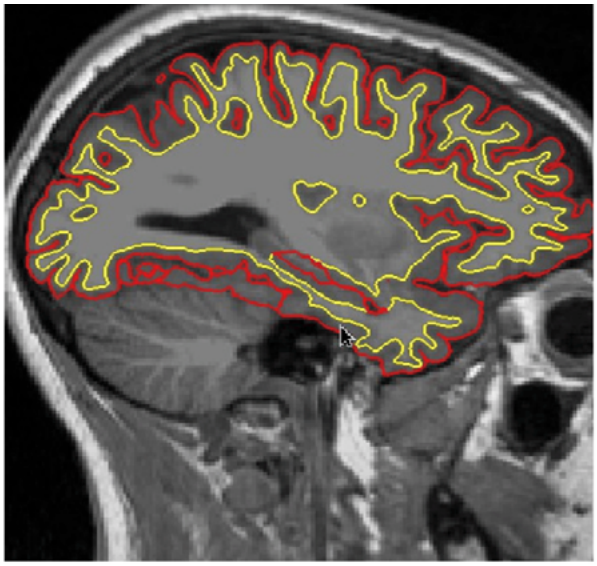
How to decide what's important to share

For me, the really important thing is that whatever I publish in my paper, you should be able to check it. You should be able to verify what I say I've done. I often appropriate Regan's catchphrase of 'trust, but verify'. I don't think you should trust any publication unless you can check their work. If you can't see the data and what we've done with it, it's just a story.

The data that's available in figshare are CSV files that have regional averages for each person. We're not releasing the raw data or everything that we have available - we're releasing a processed set of data. I'm actually working right now on releasing the whole brain data. It is an exceptional resource but much more infrastructure is needed to support it.

What I think is necessary is what we did - releasing the data that went into all of the statistical models and the code to run all of those statistical models. You can conceptualise it like this: we do some pre-processing to get the data into a format that is meant to be our independent variables, then we write some code to do the analyses, then we write a paper that presents our results. I think it's important to share all of that, but I think it's logistically harder. I'm currently writing a Scientific Data paper outlining those pre-processing steps.

I also have everything stored on GitHub with a README file explaining why things are stored on GitHub and figshare. The reason is that you can never delete anything from figshare. To me, that is hugely important: if I move onto a different job, get busy, or don't care about these results anymore and clean up my GitHub repository, our paper will still exist but people still need to access the data.



The (still) developing adolescent brain

I also use figshare for sharing slides. I gave a talk recently and was able to share the DOI to my slides on Twitter. Also, it's really easy. I think a large number of researchers are open to sharing their slides. They'll happily email them. But that's quite a barrier - there are plenty of people who won't bother with emailing to ask for them! If I provide a link to a DOI instead, anyone can have a look at them.

Using figshare to share private links

When collecting the supplementary materials for the paper we published, one feature that I thought was very valuable was the fact that we could keep the project private but send the peer reviewers links to everything. After publishing the paper, just one click made the data and code publicly available.

How researchers can plan to make their data available

I've always believed that data and code should be available along with the paper and I'm really grateful that something like figshare exists to give me an infrastructure to make that happen. I would love for people to slow down their research publication timeline a little bit more and add in a data management plan that includes how other people are going to access the data.

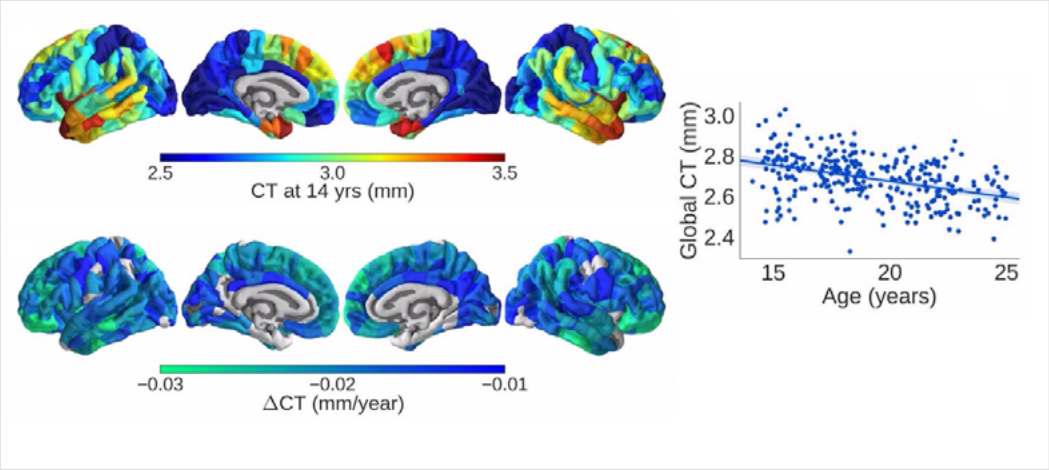
What can young researchers do to begin to share their data?

The message I would like to send to young researchers, particularly people like me who are in collaboration with other researchers, is to start small and make little things available.

For me, if you go back through my data on figshare, I made some MRI data available. That's a picture of my head - I own that data. It took me 10 minutes to upload it to figshare. I also make my talks available. Those are little things that early career researchers can do that is still really valuable and if it's available in perpetuity and has a DOI, you can get cited for it.

I feel that the open science movement has sometimes had a rhetoric of doing it perfectly or not at all. We need to start being a little more inclusive - do what you can and do what works for you.

"For me, the really important thing is that whatever I publish in my paper, you should be able to check it. What I think is necessary is what we did - releasing the data that went into all of the statistical models and the code to run all of those statistical models."



Structural brain development during adolescence and its relation to psychiatric disorders

See Kirstie's research on figshare!