Increasing the Impact of Neuroconductor

Abstract

Over the past 5 years, Neuroconductor has centralized the packages of medical image analysis in the R community. As a repository of a wide variety of analyses of diseases such as Alzheimer's disease (Ding et al. 2019) and multiple sclerosis(Valcarcel et al. 2018; E. Sweeney et al. 2013; E. M. Sweeney et al. 2013), image processing and prediction (Tustison et al. 2019; Shrivastav et al. 2018; Polzehl and Tabelow, n.d.; Shinohara et al. 2014), image visualization (Maag 2018; Mowinckel and Vidal-Piñeiro 2019) and state-of-the-art statistical analyses (Vandekar et al. 2019). For Neuroconductor to succeeed for the next 5 years and beyond, we need to grow its contributor community, and improve the stability, usability, and portability of the platform.

1. Brief Description of Neuroconductor

Neuroconductor is a platform for R package submission and repository of released packages. A user submits a package to Neuroconductor from a GitHub repository The backend of the platform then runs an initial set of sanity checks, clones the repository to the Neuroconductor GitHub repository, then checks the package on the continuous integration Travis CI (for Linux/OSX) and Appveyor (Windows). Any failures in this package checks will be reported back to the author (maintainer) of the package about what failed and what changes need to be made. As all code is on GitHub, the Neuroconductor team will also send pull requests to fix the code in many cases. The built packages are then combined every few months as a release/snapshot of all the passing packages at that time.

2. Goals

The main objectives of this proposal are:

1. improve the stability and usability of the current (N=98) packages in Neuroconductor.
2. increase the number of packages and community of Neuroconductor's contributors and developers by reaching out to seasoned users and helping them get involved more closely in the project.
3. refactor the core architecture of the Neuroconductor backend platform to handle more packages and incorporate user workflows.

Our goals will be reached if:

1. Each package has a vignette (a tutorial/long-form documentation) and all packages have a code coverage above 50% (currently 45% meet this criteria).
2. We add at least 10 new packages over the course of the project and get contributions from 20 existing contributors (out of 26). Contributions will be defined as releasing updates to packages through developers GitHub repositories or submitting additional packages to the collection.
3. Scale the Neuroconductor framework to 200 packages (currently has 98) and be able to achieve monthly releases, including releases of Docker images.

3. Improve stability of Neuroconductor packages

One of the key metrics of package stability is code coverage, which we define as the proportion of the source code that is executed when running tests, examples, or vignettes (tutorials). The `covr` package (Hester 2019) has functions to report the coverage of a package, which can be reported using services such as CodeCov (https://codecov.io/) or Coveralls (https://coveralls.io/), which is implemented in Neuroconductor.

With 4 releases (https://neuroconductor.org/releases/) in the past year (2019), the project has proved its ability to deliver checks and updates to a large number of packages. The stability of those releases has grown

over time, which is positive, but shows that improvements to the stability of the network of packages still need to be made.

John Muschelli will send a series of pull requests to packages that do not have code coverage over 50%. These pull requests will try to increase the code coverage to necessary level, but moreso provide the groundwork for maintainers to create their own tests (especially those for edge cases). Additionally, these pull requests will create vignettes of how to use the package if none exist. These vignettes will have intentional errors and breaks so that developers must change these before submitting to Neuroconductor. Documentation on increasing code coverage will be created to more formalize this process and help developers (see the next section).

By the end of the funding year, all current packages must meet these requirements to stay within the Neuroconductor repository.

4. Growing the contributor community

The Neuroconductor platform (https://neuroconductor.org/) (Muschelli et al. 2018) was started in 2014 at Johns Hopkins to centralize medical image analysis R packages, similar to the effort of Bioconductor (Huber et al. 2015). One of the main strengths of Neuroconductor is the developer community, which involves 26 maintainers residing in 5 countries, which contribute to, use, and advertise for Neuroconductor. Our plan to strengthen the community is to host a conference call/meetup twice a year to discuss developments with the maintainers, including requesting feedback for ways to improve the platform, the submission process, feature requests, or increasing user/developer engagement. With respect to users, we have had users in over 48 countries download packages from Neuroconductor.

Comparatively, the project development team is small and but also less diverse. The team is currently all white males, 2 US-born and 2 Romanian-born. Moreover, one of the developers (John Muschelli) contributes a total of 48 packages to the project. Many of these packages (18) are data packages, which include image templates or example data sets for analysis, requiring little maintenance. We would like to increase the number of developers, especially those from different countries, races, and sex. Currently, out of the 26 maintainers, only 6 are female (23%).

We propose to take the following steps to make it easier for new developers and users to contribute to the project.

4.1. Developing a code coverage user guide

We have created documentation about what changes or checks are done to packages when submitting to Neuroconductor (https://neuroconductor.org/tutorials/continuous_integration) but no clear documentation of the expectations of the package.

We propose to create:

1. Formal package guidelines that mirror closely, but not exactly, those from Bioconductor (https://www.bioconductor.org/de guidelines).
2. Examples of how to create unit tests and increase code coverage in Neuroconductor packages.
3. A contributor guide of how to create workflows and submit them to the tutorial area of Neuroconductor (https://neuroconductor.org/tutorials).

The last point is different than vignettes in a subtle way. As we require all vignettes in Neuroconductor to be built on Travis CI or Appveyor, the data must be accessible programmatically, enclosed in the package, or use data from other packages. Package maintainers and developers may make full pipelines or example analyses that may be too computation-heavy for vignettes, interact with data sources that are not open, or use many packages that are not desirable dependencies for that package. The tutorial area is for these analyses, but it is currently unclear how to create or submit these, thus we need a contributor guide.

4.2. Adding new packages

There are a number of packages that exist in R for imaging, but are only hosted on GitHub or both GitHub/CRAN. We reach out to these authors whenever we find those packages, but the effort is not always formalized. Thus, we use R Documetation (https://www.rdocumentation.org/), which aggregates packages

from all major R repositories to identify packages that are imaging-related and reach out to maintainers. Along with the every-other-month call above, we believe these steps will increase the number (and hopefully the diversity) of developers and the number of contributions of developers. We will count the number of package updates in our backend.

4.3. Hands on tutorial

Either on site in Baltimore, at a remote site, or at an imaging conference, one in-person, hands-on tutorial will be given by John Muschelli. This tutorial will be used to create an online course that will be either hosted on LeanPub or Coursera.

5. Refactor the Neuroconductor framework

5.1. Portability of Neuroconductor

Though most of the checking of packages is done on cloud-based continuous integration services, the backend is not in a true cloud platform, therefore not as portable as we would like. Currently, the Neuroconductor backend is hosted on a RedHat 6.10 server at Johns Hopkins University with the backend consisting of Drupal 8 CMS framework that relies on PHP 5.3, MySQL 14.14 and Apache 2.4. Over time, we have upgraded the backend to new operating systems, Drupal and PHP versions, and added additional third-party software to the server. Though this has worked for our current needs, we wish to migrate the backend to a cloud-based server, likely Amazon Web Services (AWS).

To address the portability of the system, we will create a custom AWS image that can be used to spool up a backend. This image will be versioned and backed up. Thus, if we wish to create another centralized platform for R packages (such as for wearable devices), this should be possible without much work.

5.2. Ease of Use of Neuroconductor

We will also create a series of Docker images that have a large percentage of the Neuroconductor packages installed, including open-source third party medical imaging software such as AFNI and ANTs (Cox 1996; Avants, Tustison, and Song 2009) and other-licensed software, such as FSL (Jenkinson et al. 2012). These images will increase the ease of use for a number of users, especially those on Windows, as many imaging software packages may not work on Windows (though almost all R packages should). An example use case is providing these Docker images for students in our "Neurohacking in R" course (https://www.coursera.org/learn/neurohacking), which currently provides a downloadable virtual machine image for use.

5.3. Scaling Neuroconductor

Though Neuroconductor can handle the 98 packages, we wish to grow the community and number of packages. Whenever a new package is submitted, that package must be checked, along with any packages that depend on that package. That creates a large number of continuous integration (CI) jobs. Thus, increasing the number of packages will require either a) more concurrent jobs on cloud systems, or b) the setup of servers for our needs. Now, we believe the CI systems we employ can handle our needs, but we need to expand above the free, single-user plans, while keeping the dependency structure bookkeeping on our server side.

6. Work plan

The first objective is to improve stability of Neuroconductor packages. John Muschelli is best placed to do this as a developer of R packages for over 10 years and contributor to many Neuroconductor packages already.

The second objective, to grow the contributor community, will be tackled by John Muschelli, Brian Caffo, and Ciprian Crainiceanu. Ciprian and Brian are best placed to grow the community due to their large networks in the statistical and imaging communities. Each has also organized meetings of many different parties, such as full conferences (ENAR - https://enar.org/).

The third objective, to scale the Neuroconductor platform, will be handled by Adrian (Adi) Gherman. As the main developer on the Neuroconductor backend, Adi is in a good position to implement the changes. He has experience with Drupal, PHP, and Docker, including carrying out large changes to the Neuroconductor system.

6.1. Improve Stability

The timescale is given on the assumption of John Muschelli working part time on this goal (0.35 full time equivalent). Our strategy to improve the stability of current packages is as follows:

1. For the 54 with insufficient code coverages, make pull requests to the packages (approximately 10 per month) - 6 months.

2. Write tutorials for prospective contributors/developers on increasing code coverage and improved submissions - 2 months
3. Improve the existing user documentation on https://neuroconductor.org/ - 3 months

6.2. Growing the contributor community

The timescale is given on the assumption of Ciprian Crainiceanu and Brian Caffo working part time on this goal (0.05 full time equivalent) and above time from John Muschelli.

1. Set up a conference call once every 2 months - 1 month
2. Reach out to R-ladies Baltimore and other R-ladies group to discuss Neuroconductor to increase diversity in the user base - 2 months
3. Organize sessions related to Neuroconductor and open source software at imaging and statistics conferences - 1 month
4. Provide one hands-on tutorial - 3 months
5. Create a survery for developers to get more detailed information, such as identified sex, race, age, and other factors such as education attained and target analysis diseases - 2 months

6.3. Refactor the Neuroconductor framework

The timescale is given on the assumption of Adrian Gherman working part time on this goal (0.5 full time equivalent).

1. Setting up AWS systems and testing - 1 month
2. Upgrading the Neuroconductor system to a newer OS version and most up-to-date Drupal/PHP/MySQL/Security - 2 months
3. Automating the creation of Docker images - 2 months
4. Implements new checks for code coverage and vignettes - 2 months

7. Potential areas for growth

Though we wish to stabilize the Neuroconductor system, we wish to integrate analyses of imaging from Neuroconductor and "omics" from Bioconductor. This integration is challenging due to the size of the data and heterogeneity of the sources, amongst many other reasons. For a review, see Antonelli et al. (2019). For example, the `limmi` package (https://github.com/muschellij2/limmi) is an attempt to coerce functional magnetic resonance imaging into a testing framework that both works with `limma` package (Ritchie et al. 2015) and into the `SummarizedExperiment` format, which both are fundamental elements from Bioconductor (Huber et al. 2015).

8. Existing support

We are currently applying for AWS credits in addition to the funding for computing here. Neuroconductor has been supported by General Funds from the Biostatistics Department at Johns Hopkins Bloomberg School of Public Health. NIH Grants XXX currently have a small amount of funding to support development of packages for the system, which have been used for package and backend development.

This Grant was generated with all materials at https://github.com/muschellij2/CZI and can be downloaded from https://johnmuschelli.com/CZI/index.pdf.

References

Antonelli, Laura, Mario Rosario Guarracino, Lucia Maddalena, and Mara Sangiovanni. 2019. "Integrating Imaging and Omics Data: A Review." *Biomedical Signal Processing and Control* 52: 264–80.

Avants, Brian B, Nick Tustison, and Gang Song. 2009. "Advanced Normalization Tools (ANTS)." *Insight J* 2: 1–35.

Cox, Robert W. 1996. "AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages." *Computers and Biomedical Research* 29 (3): 162–73.

Ding, T, AD Cohen, EE O'Connor, HT Karim, A Crainiceanu, J Muschelli, O Lopez, et al. 2019. "An Improved Algorithm of White Matter Hyperintensity Detection in Elderly Adults." *NeuroImage: Clinical*, 102151.

Hester, Jim. 2019. *covr: Test Coverage for Packages*. https://CRAN.R-project.org/package=covr.

Huber, W., V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, et al. 2015. "Orchestrating High-Throughput Genomic Analysis with Bioconductor." *Nature Methods* 12 (2): 115–21. http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html.

Jenkinson, Mark, Christian F. Beckmann, Timothy E. J. Behrens, Mark W. Woolrich, and Stephen M. Smith. 2012. "FSL." *NeuroImage* 62 (2): 782–90. https://doi.org/10.1016/j.neuroimage.2011.09.015.

Maag, Jesper LV. 2018. "gganatogram: An R Package for Modular Visualisation of Anatograms and Tissues Based on ggplot2." *F1000Research* 7.

Mowinckel, Athanasia M, and Didac Vidal-Piñeiro. 2019. "Visualisation of Brain Statistics with R-Packages Ggseg and Ggseg3d." *arXiv Preprint arXiv:1912.08200*.

Muschelli, John, Adrian Gherman, Jean-Philippe Fortin, Brian Avants, Brandon Whitcher, Jonathan D Clayden, Brian S Caffo, and Ciprian M Crainiceanu. 2018. "Neuroconductor: An R Platform for Medical Imaging Analysis." *Biostatistics*, kxx068. https://doi.org/10.1093/biostatistics/kxx068.

Polzehl, Jörg, and Karsten Tabelow. n.d. "Magnetic Resonance Brain Imaging." Springer.

Ritchie, Matthew E, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. 2015. "limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." *Nucleic Acids Research* 43 (7): e47. https://doi.org/10.1093/nar/gkv007.

Shinohara, Russell T, Elizabeth M Sweeney, Jeff Goldsmith, Navid Shiee, Farrah J Mateen, Peter A Calabresi, Samson Jarso, et al. 2014. "Statistical Normalization Techniques for Magnetic Resonance Imaging." *NeuroImage: Clinical* 6: 9–19.

Shrivastav, Kumar Dron, Ankan Mukherjee Das, Harpreet Singh, Priya Ranjan, and Rajiv Janardhanan. 2018. "Classification of Colposcopic Cervigrams Using EMD in R." In *International Symposium on Signal Processing and Intelligent Recognition Systems*, 298–308. Springer.

Sweeney, Elizabeth M, Russell T Shinohara, Navid Shiee, Farrah J Mateen, Avni A Chudgar, Jennifer L Cuzzocreo, Peter A Calabresi, Dzung L Pham, Daniel S Reich, and Ciprian M Crainiceanu. 2013. "OASIS Is Automated Statistical Inference for Segmentation, with Applications to Multiple Sclerosis Lesion Segmentation in MRI." *NeuroImage: Clinical* 2: 402–13.

Sweeney, EM, RT Shinohara, CD Shea, DS Reich, and Ciprian M Crainiceanu. 2013. "Automatic Lesion Incidence Estimation and Detection in Multiple Sclerosis Using Multisequence Longitudinal MRI." *American Journal of Neuroradiology* 34 (1): 68–73.

Tustison, Nicholas J, Andrew J Holbrook, Brian B Avants, Jared M Roberts, Philip A Cook, Zachariah M Reagh, Jeffrey T Duda, et al. 2019. "Longitudinal Mapping of Cortical Thickness Measurements: An Alzheimer's Disease Neuroimaging Initiative-Based Evaluation Study." *Journal of Alzheimer's Disease* 71 (1): 165–83.

Valcarcel, Alessandra M, Kristin A Linn, Fariha Khalid, Simon N Vandekar, Shahamat Tauhid, Theodore D Satterthwaite, John Muschelli, Melissa Lynne Martin, Rohit Bakshi, and Russell T Shinohara. 2018. "A Dual Modeling Approach to Automatic Segmentation of Cerebral T2 Hyperintensities and T1 Black Holes in Multiple Sclerosis." *NeuroImage: Clinical* 20: 1211–21.

Vandekar, Simon N, Theodore D Satterthwaite, Cedric H Xia, Azeez Adebimpe, Kosha Ruparel, Ruben C Gur, Raquel E Gur, and Russell T Shinohara. 2019. "Robust Spatial Extent Inference with a Semiparametric Bootstrap Joint Inference Procedure." *Biometrics* 75 (4): 1145–55.