Research Statement

John Muschelli

My research philosophy is to identify and solve the most relevant scientific problems. Many times, those problems require identifying, applying and refining existing methods or developing completely new statistical methods. Having a broad understanding of methodology can help with porting methods developed for a particular area of research, which can open new avenues of analysis. I see my research role as a developer of new, appropriate, and reproducible analytic methods. Methods can only be complete when accompanied by the necessary tools (i.e. software) that allow them to be used, tested, and refined. Moreover, I believe the analysis should be reproducible so that others can use the tools or results for future work.

When choosing projects, I find that those that can directly improve quality of life, patient care, or outcomes the most fulfilling. Hence, I have focused on projects rooted in well-defined clinical problems. My interests include brain imaging, high-dimensional data analysis, machine learning, the impact of data pre-processing on results, data visualization, and software development.

My overarching research goal is to build a research group on neuroimaging that is highly integrated with the various departments of the University that are conducting neuroimaging and "big data" research. The research group will be built on the same principles as the SMART (Hopkins) and PennSIVE (UPenn) research groups and will aim to establish my home department as one of the leading centers in methods development for neuroimaging.

Brain Imaging In spite of their importance, clinical images are underutilized in research. Most clinical images are often used qualitatively, at the subject-level, and are not organized for analysis. Thus, it is difficult to describe population-level metrics that are reproducible and generalizable. Moreover, methods developed for healthy brains cannot be applied directly to brain images that contain serious pathologies.

Thus, I have been interested in: 1) organizing heterogeneous brain imaging data sets collected from different centers and scanners; 2) pre-processing these data from their scanner format to analyzable formats; 3) developing pre-processing pipelines that are fully reproducible in R; 4) analyzing population level data and extracting subject-specific biomarkers; and 5) developing the methodological tools for analyzing the resulting high dimensional brain imaging data using sound statistical principles. It is fair to say that I am one of the very few Bio/Statisticians who has been involved in all stages of data collection and processing. Most importantly, I enjoy collaborating with neuroimaging scientists and I can easily initiate contacts and build and sustain the much-needed collaborations between biostatistics departments and neuroimaging researchers. Most importantly, I can create easily-used software and interpretable outputs for clinical researchers to use.

Hemorrhagic Stroke My clinical work has involved 2 stroke clinical trials: MISTIE (Phase II) and CLEAR (Phase III). My first duties involved creating reports for the data safety monitoring board (DSMB). By implementing an automated reporting framework, I reduced the turnaround time for the report from 2 weeks to generating daily reports with new data. This automated the iterative process of revision, resulting in a higher-quality report.

While analyzing endpoint outcomes, we observed the well-known relationship between stroke volume and the likelihood of positive functional outcome. However, the measurements for stroke volume are obtained either using a fast and coarse approach or a time-consuming and accurate one. The accurate method relied on human tracings on slices of a computed tomography (CT) scan to determine stroke location and size.

To reduce intra- and inter-reader variability, we developed a statistical model to automatically estimate the probability that a given voxel was part of the stroke. This allowed us to perform stroke volume estimation on a large scale. We created a set of features from the CT scan based on experience and discussion with our clinical collaborators. Based on these features, we estimated the probability that an area is part of the stroke using a logistic regression. I implemented the entire process: image conversion, image pre-processing, image analysis, and presentation of results.

One crucial step in the above segmentation procedure was to separate the brain from the rest of the image (such as bone, facial tissue, etc.). Few tools exist for segmenting the brain in CT images, but many exist in magnetic resonance imaging (MRI). In “Validated automatic brain extraction of head CT images”, we adapted a previously published method for brain extraction in MRI scans for CT scans, which was suggested previously but never fully validated.

In “Quantitative Intracerebral Hemorrhage Localization”, using the hand-traced manual segmentations, we estimated a population-level map describing where strokes occurred in the trial population. Previous work described the hemorrhage areas using coarse measurements such as the percent of patients having a stroke in a large area and few spatial maps of stroke prevalence had ever been published. We also determined anatomic regions of the brain that relate to stroke severity scores; having a stroke in these regions indicates worse stroke severity.

Multiple Sclerosis (Current Work) In patients with multiple sclerosis (MS), brain lesions that react to gadolinium, an injection agent, are markers for increased inflammation and are clinically relevant. Again, I developed relevant features for estimating gadolinium-enhancing lesions using the pre- and post-gadolinium injection T1-weighted, and pre-injection T2-weighted MRI scans. I estimated the probability that a brain area had an enhancing lesion using a random forest approach based on the proposed features. I also compared 2 intensity-based normalization schemes to determine their impact of data pre-processing on estimation performance.

Software Development The majority of my work has been implemented in R; I have been programming in R for the last 8 years. As such, I realize methods are rarely used unless well-documented code has been created and shared. I created and released a series of R packages and contributed to many others. My most significant contributions have been the fslr (presented in “fslr: Connecting the FSL Software with R”), brainR (“brainR: Interactive 3 and 4d Images of High Resolution Neuroimage Data”) and extrantsr (`https://github.com/muschellij2/extrantsr`) packages. I am currently the only developer on all three packages.

The fslr package allows R users to call FSL, a widely used and tested neuroimaging software, directly from R. This allows users to process their imaging within the same framework as their analysis without learning FSL-dependent syntax.

The brainR package allows users to interactively visualize 3-dimensional images over another dimension using a JavaScript toolkit (XTK) and simple HTML inputs. We have used it to see the progression of MS lesion incidence over time.

The extrantsr package creates convenient wrapper functions for users to pre-process structural neuroimaging data within R, calling heavily upon the ANTsR package (`https://github.com/stnava/ANTsR`). The ANTsR package is also a re-implementation of a widely used and tested neuroimaging software, advanced normalization tools (ANTs). The extrantsr package allows users to pre-process their data with simple and standardized functions with a limited amount of code. Thus, statisticians can perform a full brain imaging analysis completely within one language (R).

Future Research I am currently validating the regions that relate to higher stroke severity derived in “Quantitative Intracerebral Hemorrhage Localization” in an independent cohort of similar patients with stroke. We are also gaining access to thousands of CT scans from patients with stroke and will test our automated segmentation method on that data.

My future work is determining a general procedure for pre-processing structural imaging data, normalizing image intensities, and extracting features for image segmentation. The deliverables are R packages that apply the same general principles for image segmentation to different diseases, including multiple sclerosis, stroke, brain cancer, and traumatic brain injury.