

Statistical challenges in “big data” human neuroimaging

Stephen M Smith¹ Thomas E Nichols^{2,1,3}

¹Wellcome Trust Centre for Integrative Neuroimaging (WIN-FMRIB), University of Oxford, Oxford, UK

²Big Data Institute, University of Oxford, Oxford, UK

³Department of Statistics, University of Warwick, Coventry, UK

Introduction

Over the last 30 years, structural and functional brain imaging has become a powerful and widespread tool for clinical and basic neuroscience. However it is only now that some brain imaging studies have started to join the ranks of “big data” science. Whereas most neuroimaging studies continue to have modest sample sizes (number of subjects, $N < 50$) and modest amounts of data collected per subject, a small number of studies are now starting to embrace “big” imaging in a number of ways – with subject numbers in tens of thousands, or taking advantage of huge increases in the quantity of imaging and non-imaging collected on each subject. For example, the Human Connectome Project (HCP) (Van Essen et al. 2013) has recently completed imaging of $>1,000$ young adults, with an impressive 4 hours of scanning per subject, and utilizing vast improvements in the spatial and temporal resolutions of the acquired data. In a complementary manner, UK Biobank (UKB) (Miller et al. 2016) is acquiring more modest amounts of imaging data per subject, but is scanning 100,000 volunteers, and this imaging is just part of the much larger overall UKB project that includes genetics, biological and lifestyle measures and health outcomes from 500,000 subjects. In yet another approach to big data imaging, the ENIGMA consortium is amassing imaging and genetic data from tens of thousands of subjects by pooling many existing studies, using advanced meta-analysis techniques to overcome restrictions on sharing individual-subject data (Bearden and Thompson 2017).

In this brief review we highlight some of the challenges that “big data neuroimaging” brings. While there is no hard definition, we may consider an imaging study “big” if it has 1000 or more subjects and/or collects a substantially larger set of measurements than usual, or expands on traditional measurements (e.g., with greater resolution or duration than typical). We focus on MRI-based brain imaging, though many of the points made are also relevant for valuable complementary techniques such as EEG, MEG and PET. All aspects of neuroimaging data are rapidly becoming “bigger”. Spatial resolution and number of images are constantly increasing, enabled by enhancements in MRI hardware (higher field strength and more coil channels) and software (advanced acquisition and reconstruction techniques). Additionally, the proliferation of research scanners means that study designs are becoming more ambitious, with longer scan durations, multiple scan sessions and large subject numbers. For example, if one considers the increases in state-of-the-art fMRI imaging spatial resolution, temporal resolution and receive-coil numbers, data size has increased quite in line with Moore’s law (a doubling every two years, i.e., a thousand-fold increase in the last 20 years).

At the heart of an imaging study lies the image processing and statistical modelling, a major goal of which is to work against the “big” – to reduce the raw data to meaningful and concise information that allows the original experimental question to be answered. For example, in UKB, in addition to requiring image processing to remove artifacts and align images across modalities and individuals, imaging-derived phenotypes (IDPs) are computed - currently 2,500 distinct measures of brain structure and function. Examples of IDPs include the volume of specific brain structures, the strength of connectivity between pairs of brain regions and the dispersion of fibers in specific white-matter tracts. As a result, the 20Petabytes of raw neuroimaging data from 100,000 subjects will be converted into 300TB of pre-processed image data and finally into 1GB of IDPs (i.e., a total compression ratio of 20 million). However, it is also possible for image analysis to greatly increase data size. For example, HCP resting-state functional MRI (rfMRI) timeseries data is under 2GB per subject, but once this is processed to generate the voxels \times voxels matrix of “dense” brain connectivity, this expands twenty-fold to 33GB.

These numbers immediately illustrate perhaps the most obvious difficulty with big data neuroimaging – data sizes can quickly become hard to manage, both in terms of long-term storage and compute speed and memory. However this is just one of the challenges.

Big sample sizes - small effect sizes

A primary reason that prospective epidemiological health studies (such as UKB) need very large subject numbers is to be able to acquire data on subjects in advance of later disease development, hence allowing researchers to learn the earliest markers of (and possibly mechanisms for) disease; the large subject numbers are needed as there is no single disease focus, and only later can patient-control subsets be identified. However, another reason for acquiring data from large numbers of subjects is to allow the identification of subtle effects that are not statistically detectable in smaller groups. This has been the scenario in most genome-wide association studies, where individual genetic effects tend to be very small. The associated downside is of course exactly the same point - that statistically significant effects could be very small in terms of the biological effect (percentage variance explained in the data, Figs. 1,2). For example, in an analysis of almost 15,000 UKB subjects, even after Bonferroni correction for 14 million association tests between IDPs and non-imaging variables (such as cigarette consumption or cognitive test scores), statistical significance can be reached with much less than 1% variance in the non-imaging variables explained by the IDP (Fig. 3).

Hence, in the worst case scenario, having huge N renders “significant” individual associations meaningless. However, where many variables are considered simultaneously, reasonable total % variance prediction may be found. For example, multivariate modelling from the whole genome can account for almost 20% of hippocampal volume, whereas individual genetic variants only explain up to 1% (Bearden and Thompson 2017). Similarly, analysis of the first 5,000 UKB subjects found that population variance explained in pairwise associations (between IDPs and non-imaging variables) reached maxima of around 5%, while multivariate analyses reached up to 50% variance explained (Miller et al. 2016).

Big sample sizes – big confound trouble

A more insidious problem associated with huge N – again closely related to the upside of having high sensitivity to real effects – is high sensitivity to artifactual associations due to confounding effects. Just tiny amounts of shared confounding factors that feed into two (otherwise independent) variables of interest can induce false associations; even when a real association exists, confounds can bias the estimate of the correlation. For example, socioeconomic status might influence cheese consumption and (independently) rate of cognitive decline in aging as reflected in the volume of the hippocampus; this would induce an apparent association between cheese and hippocampus volume that is not causal, but induced by a common factor.

Confound effects are particularly problematic in big imaging studies because of the huge variety of potential imaging artifacts, including many factors that can affect both the imaging and non-imaging variables of interest. Just some of the confounding effects are: head motion, head size, changes in breathing rate and depth, scanner hardware and software changes that are much more likely to occur in big, longer-term imaging studies (Table 1, Fig. 2). It is essential therefore to develop sophisticated image processing methods for removing the effect of such confounds from imaging data; two very effective examples are structured noise removal from rfMRI (Salimi-Khorshidi et al. 2014) and diffusion MRI (dMRI) data (Andersson and Sotiropoulos 2016).

There can also be interpretability problems caused by changes in the brain. For example, structural brain changes can easily cause misinterpretation of results in other modalities through partial volume effects. This happens where voxels contain a mixture of tissue types, and atrophy (or other structural variation over time or between subjects) changes the mixture of tissues within voxels. Because MRI voxels are on the order of millimeters while the cortex is only a few millimeters thick, this is a common occurrence, in particular at the grey-white boundary. This leads to structural changes being incorrectly attributed to a change in functional activity in fMRI, or white matter microstructure in dMRI.

There are also factors that both induce imaging confounds and also relate to the effect of interest, making their removal from the data even more complex. For example, blood pressure can be a health factor of direct interest, but causes confounding effects in resting-fMRI functional connectivity estimates. In addition, blood pressure can lead to differences in dMRI measures in white matter that are driven by vascular changes rather than alterations to the fibre architecture of interest. Similarly, aging effects in the brain may be of interest, but age-related structural atrophy may also act as an imaging confound (for example, changing partial-volume effects); additionally, in an aging population, age may be a dominant source of inter-subject variability, and hence could be viewed as an important mediator in any imaging vs. non-imaging association, or could be seen as a significant source of unwanted variance.

Another concern is when a confounding effect is identified but can only be measured/estimated with error; in this case the adjustment is incomplete and increasing N just increases sensitivity to the un-adjusted variance that ‘leaks’ through; addressing this requires measures of confound uncertainty and a more complex modelling approach that propagates this uncertainty through the modelling process (Westfall and Yarkoni 2014). Another potentially useful approach is to generate an expanded set of confounds by adding transformed versions of those available (e.g. nonlinear transformations that suppress or exaggerate outliers). Adjusting the data with such an expanded set will allow more than just linear effects of confounds to be captured.

Another method for dealing with confounds is to sub-select subjects, i.e., identify a matched control for each patient or exposed individual (e.g., match for sex and age rather than by regressing those confounds out of the data). Doing this exactly is often impossible, but a soft version can be accomplished with propensity score matching (Rosenbaum and Rubin 1983). However such matching leaves out data, and so it might seem better to just “adjust for everything”, that is, include all possible confounds. However, if a variable being considered as a confound is actually a consequence of two otherwise independent variables, adjustment for this confound artefactually induces an association, in an effect known as Berkson’s paradox or “collider bias”. For the same mathematical reasons, false associations can also arise from selection (or ascertainment) bias, relating to the recruitment of subjects into a study. For example, it has been noted that individuals in UKB smoke less and hold more educational qualifications than the UK population as a whole; if it were the case that these two features actually causally influenced whether subjects participated in the UKB, a bias could be induced in any association measured between smoking and education (Fry et al. 2017; Munafò et al. 2017). Similarly, large meta-analyses can also be biased by the “file drawer problem”, where the literature is biased away from the null due to non-publication of null findings.

Multivariate modelling & machine learning – finding population patterns, making predictions

With big imaging datasets, the potential loss of statistical sensitivity associated with huge numbers of tests (across space, imaging modalities and thousands of non-imaging variables) is not the only challenge. With millions of tests comes also the challenge of interpretation of potentially thousands of significant individual results. Interpretation can be easier with multivariate analyses, where many variables are considered simultaneously; such analyses can jointly model imaging variables/voxels against a non-imaging variable (healthcare outcomes, behaviour, life factors). Alternatively, in doubly-multivariate analyses, many imaging and non-imaging variables can be jointly modelled. These methods produce a small number of dominant patterns of association that can be much easier to interpret compared to millions of univariate association tests.

Machine learning uses multivariate methods to make predictions about individual subjects and discover patterns in data. All machine learning methods must balance complexity and generalizability – the more flexible the model, the greater the chance of overfitting the data. Feature selection is one approach to managing complexity; for example, penalized regression methods (such as LASSO or Elastic Net) automatically drop variables from the model, leaving just the most important ones. However, for imaging data, where many thousands of voxels/variables feed into sparse regression models, performance can be further improved with additional pre-selection, feeding into the regression only the most promising variables (Le Floch et al. 2012). “Univariate filtering” with a target variable is one such supervised feature selection method. For example, when building a predictive model of IQ from functional connectivity, univariate filtering consists of individually correlating IQ with each connectivity “edge”, and only those edges with some evidence for an association are retained for building a multivariate predictive model.

Another approach to managing complexity is feature extraction, where variables are combined in some way to create a new, more informative variable. PCA (principal component analysis) is a common feature extraction method that is unsupervised (does not use knowledge of the target variable). However, for such unsupervised data-driven feature reduction, it can be useful to have already reduced the data in sensible ways; PCA on millions of voxels uses no prior knowledge, and may be less successful than dimension reduction based on PCA of thousands of IDPs that were explicitly designed to each reflect a meaningful quantity. For example, when looking for imaging vs. non-imaging associations in the HCP, (Smith et al. 2015) used PCA to reduce 20,000 IDPs to 100 variables. Big imaging studies naturally start with many more measurements (potentially billions of voxels/timepoints across multiple modalities) than typical non-imaging studies, hence the particular importance of such data reduction. PCA is an example of a linear feature extraction method, which has the advantage that each feature can easily be mapped back to the original variables; this is particularly important for big imaging studies where imaging confounds/artefacts can easily dominate the results (e.g., head motion in case-control studies). Nonlinear or highly reductive feature extraction methods (e.g. graph theory metrics) can be harder to interpret, including the question of whether results are confound-driven or not.

While multivariate methods typically relate multiple predictors to a single target variable, “doubly-multivariate” methods relate multiple variables from multiple sources. For example, (Miller et al. 2016) applied canonical correlation analysis to relate thousands of IDPs to thousands of non-imaging life-factor and biological measures. Such an analysis finds common patterns of population variance that exist in two datasets, which can be fundamentally different in character (e.g., imaging data and health outcomes). However, such data-driven analysis may not necessarily result in biologically-interpretable modes (patterns of population covariance). Here, Independent Components Analysis can be a powerful tool to transform the modes (e.g., from PCA or CCA) to make them more biophysically meaningful. For example, in (Miller et al. 2016), 9 significant CCA modes were “unmixed” into 9 ICA modes with enhanced reproducibility and biological interpretability, for example, relating measures of blood pressure, alcohol intake, and changes in white matter cell microstructure to each other, with up to ~20% variance explained, in 5000 subjects.

Every multivariate or predictive model is in danger of being affected by overfitting, for example, leading to over-optimistic estimates of prediction accuracy. To unambiguously demonstrate that the predictive performance is above chance a

careful null analysis is required, such as can be achieved through permutation testing. Unbiased estimates of variance explained or predictive accuracy must be assessed with rigorous cross-validation based on held-out data and, ideally, replication with another large and distinct dataset.

Significance testing - multiple comparisons, fishing trips

Since there is currently much buzz about the potential of machine learning methods, it is reasonable to consider what role existing neuroimaging statistical tools will play with big data. Currently, the workhorse tools of neuroimaging are “mass univariate”, where a regression model is fit separately at each voxel. Inferences over the image are corrected for multiple testing using random field theory (RFT) or nonparametric resampling methods (e.g., permutation or bootstrap). We believe that these tools will continue to be relevant to provide stringent control of false positives, particularly for highly focused hypotheses. RFT, an approximate analytical method, is in its element with large N; permutation methods are extremely computationally intensive with large N, but can be practical with permutation distribution approximation methods (Winkler, Ridgway, et al. 2016).

A serious problem is when researchers fly through a growing collection of imaging modalities, preprocessing options, and alternative models, searching for significance. Unless using stringent multiple testing or validation with held out data, such explorations quickly become “p-hacking”. Big open imaging datasets are explicitly designed to allow for a wide range of hypotheses to be tested, and so raise this danger further. It is vital that researchers arm themselves with detailed analysis plans that address multiplicity—ideally included as part of study plan preregistration—otherwise little confidence can be placed in the results found after such fishing expeditions. In particular, nonparametric (e.g., permutation-based) combining can be essential to account for controlling a search for effects over several modalities (Winkler, Webster, et al. 2016).

Large studies may include genetically related subjects accidentally or by design, or may use a complex sampling structure. Whether using mass univariate or multivariate methods, it is essential to control for this structure, for example with hierarchical models or constrained permutation approaches (Winkler et al. 2015). For example, the HCP includes many twins and siblings; ignoring twin structure in HCP will lead to inflated significance on heritable brain phenotypes. Similarly the ABCD study (Volkow et al. 2017) uses a stratified design, sampling schools around the country, and children within each school, which must be accounted when analyzing its data.

As large, open datasets become widely used, new complexities arise. Consider the Alzheimer’s Disease Neuroimaging Initiative (ADNI), first started in 2004, which has produced a repository of multimodal MRI for predicting the course of Alzheimer’s Disease (AD) (Mueller et al. 2005). ADNI has been very successful, and, for example, now nearly every publication on constructing imaging biomarkers to predict AD conversion (from mild cognitive impairment) is based on ADNI. However, as more and more researchers base findings on the very same ADNI data, generalizability becomes a concern: will the 100th ADNI paper on AD conversion reflect over-fitting to idiosyncrasies of this sample? Part of the solution is the continued creation of new big imaging resources, so that each new method and scientific finding can be validated on alternate and diverse populations.

Conclusions

Imaging can be a powerful way to identify phenotypes that are more “intermediate” than the measures of ultimate interest (such as health outcomes or cognitive scores), but which can be more quantitative and sensitive (Duff et al. 2015). Combining this strength of imaging with advanced analysis methods in large-scale imaging studies will allow us to find population patterns that more closely map to underlying disease mechanisms and cognitive behavior, for example as proposed in the Research Domain Criteria (RDoC) paradigm in psychiatry (Insel et al. 2010). To harness these opportunities, researchers will need to equip themselves with the tools of computer scientists and epidemiologists, learning to scale up existing tools and develop new ones that grapple with the confounds that will be as potent as the subtle effects of interest now accessible with these enormous datasets.

Acknowledgements

We would like to acknowledge all scientific contributors and participants in UK Biobank and the Human Connectome Project. In particular we are grateful to Fidel Alfaro-Almagro and Karla Miller for their huge efforts on behalf of UK Biobank. We are grateful for funding from Wellcome Trust, UK Medical Research Council and NIH.

Declaration of interests

SS is part-owner and shareholder of SBGneuro.

Figure 1. Relationship between sample size and number of variables tested, holding statistical power constant. The plot shows the sample size N needed to attain 80% power to detect 1 true association while controlling the familywise error rate (the chance of one or more false positives) over K tests. Effect size is measured in terms of % variance explained (r^2), and is shown for three small values, 1% (corresponding to a correlation of $r=0.1$), 0.1% and 0.01%. While large sample sizes are needed for just 1 test, as N increases K grows exponentially. Roughly, squaring the number of tests requires only a doubling of the sample size.

Figure 2. Voxelwise analyses of the faces-shapes contrast in the UK Biobank task-fMRI data, from 12,600 subjects. The % signal change colour overlay shows the fMRI signal change associated with the faces-shapes contrast, masked by significant voxels from mixed-effects modelling of the group-average signal (all maps shown here conservatively corrected for multiple comparisons across 1.8 million voxels using Bonferroni correction, $P<0.05$). This threshold excludes only 19% of voxels, i.e., showing a significant response to the task in most of the brain. The maximum statistical effect size (Cohen's d) is 1.57, equivalent to a one-group T statistic of 176.2. The *Sex* overlay shows significant correlation of the faces-shapes effect with the confound factor of sex; orange/blue colouring shows correlation estimated after controlling for head size, while copper/green colouring is without this adjustment. The *Head size* overlay shows significant correlation with volumetric head size; orange/blue colouring shows correlation estimated after controlling for sex, and copper/green colouring is without adjustment. Because sex and head size are highly correlated ($r = 0.63$), adjustment makes a great difference, eliminating a significant effect in some regions. Over all five results shown here, the minimal detectable correlation was $r = 0.049$. (Image intensities truncated for presentation, % signal change truncated at $\pm 0.5\%$, correlation intensities at ± 0.08 ; full ranges listed in the figure.)

Figure 3. 14 million univariate association tests between IDPs and non-brain-imaging variables in UK Biobank (14,500 subjects). In the Manhattan-style plot, 5,456 non-imaging variable are arranged on the x-axis, with 16 groups of variable types. For each variable, 7 $-\log_{10}P$ values are plotted - the most significant association of that variable with each of 7 different classes of imaging-derived phenotypes (IDPs). Approximately 100,000 associations are FDR-significant, and 15,000 are Bonferroni-significant. The histograms show the distributions of correlation size (across all 14 million tests); depending on thresholding method, the minimum detectable r is 0.03-0.05, meaning that for FDR thresholding an association with 0.1% variance explained is detectable.

Table 1. Selected examples of imaging confounds with a subset of image artefacts and potential correlates. SNR = Signal to Noise Ratio; PD = Parkinson's Disease; ADHD = Attention Deficit Hyperactivity Disorder; COPD = Chronic Obstructive Pulmonary Disease; BMI = Body Mass Index.

Confound	Example effects on MRI data	Potential artefactual correlates	Comments
Head motion	Striping, ringing, blurring, dMRI dropout, low SNR, biased connectivity	Diseases (PD, ADHD) and aging correlate with increased head motion	Relates to head size; may be estimated from and partially corrected in fMRI and dMRI
Breathing rate/depth	Changes in fMRI contrast, SNR, distortion and dropout (due to B0)	COPD, heart conditions, BMI, exercise levels, some fMRI tasks	Can cause changes in real and apparent head motion and blood oxygenation/flow
Blood pressure	BOLD contrast (fMRI) and vascular compartment size (dMRI)	Functional connectivity (fMRI), and white matter microstructure (dMRI) in disease	
Age	Structural atrophy (cortical thinning, ventricle enlargement) influences voxel partial volume effects	Non-volumetric imaging measures; interaction with disease progression	If age is not of explicit interest, it should generally be included as a confound
Scanner hardware	Differences in SNR, contrast or artefact as a function of site or date (all MRI modalities)	Other measures varying with site or date	Can occur even in studies run with "identical" hardware
Operator inconsistency	Differences in SNR, artefacts, distortion, coverage	Other measures varying with site or date	Even with automated protocol, subject placement or instructions can vary

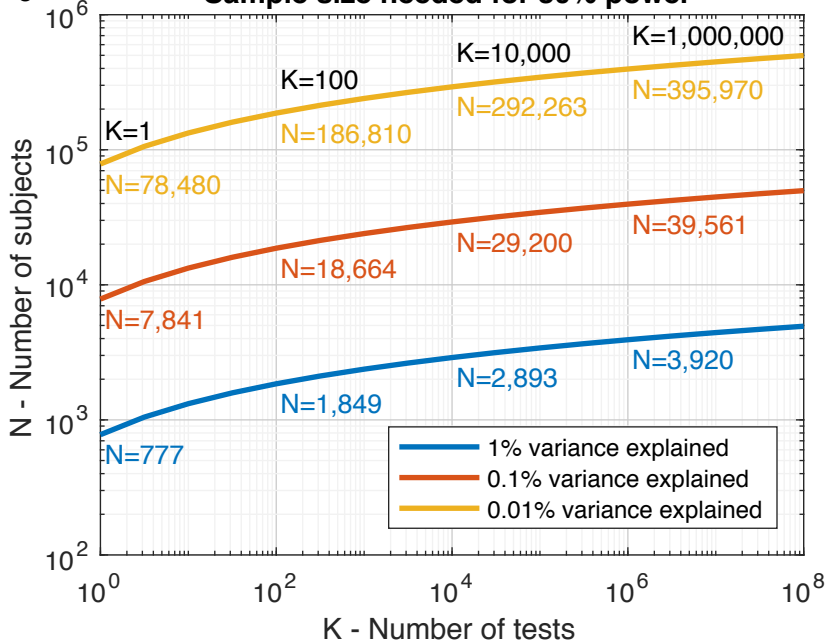
References

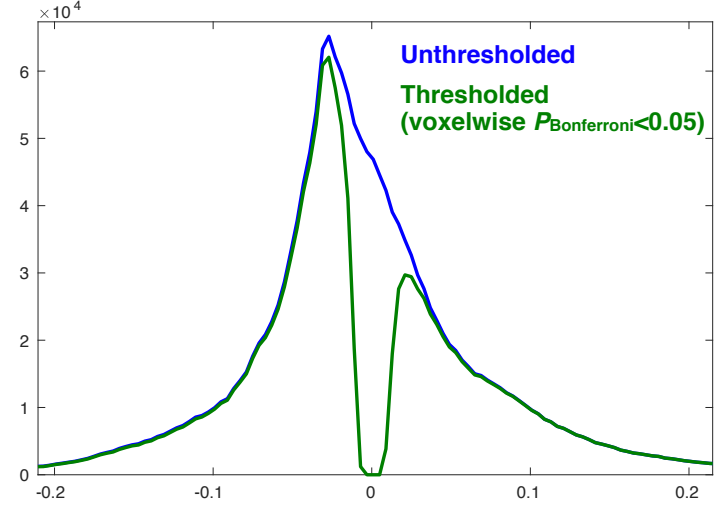
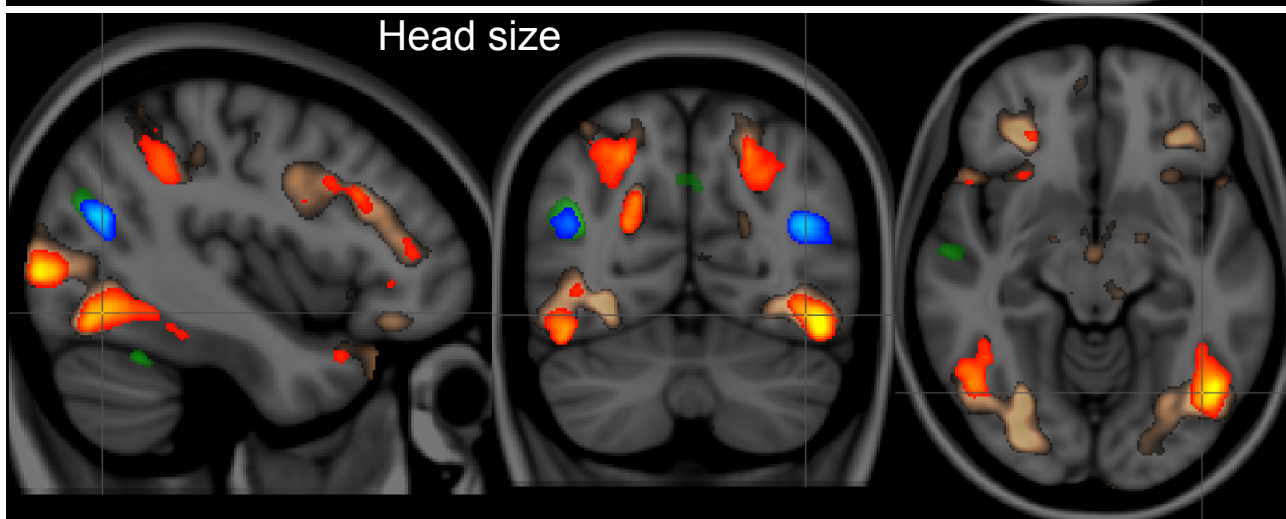
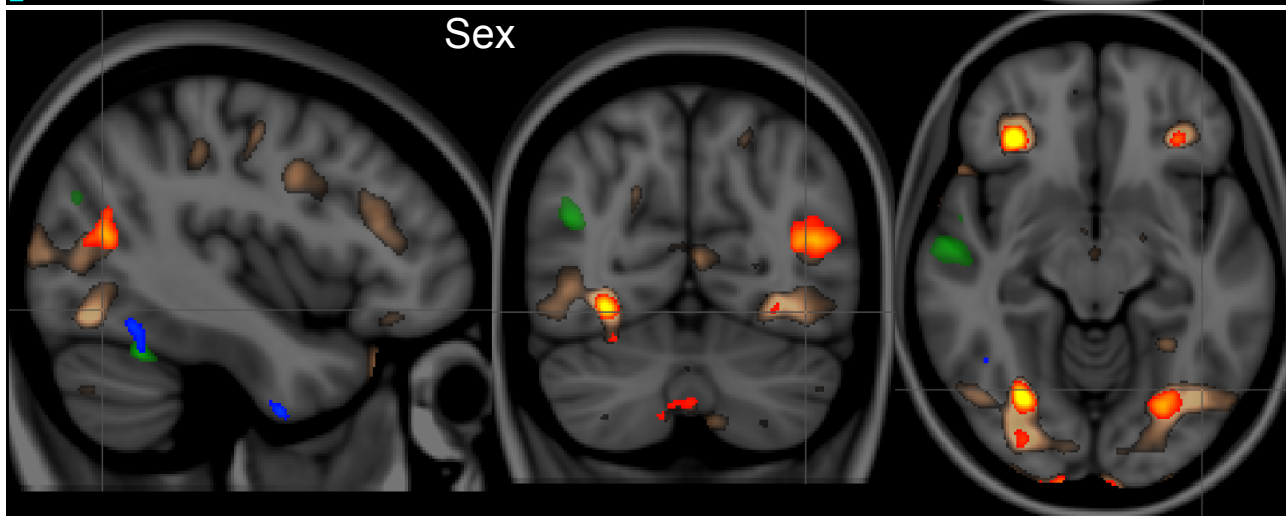
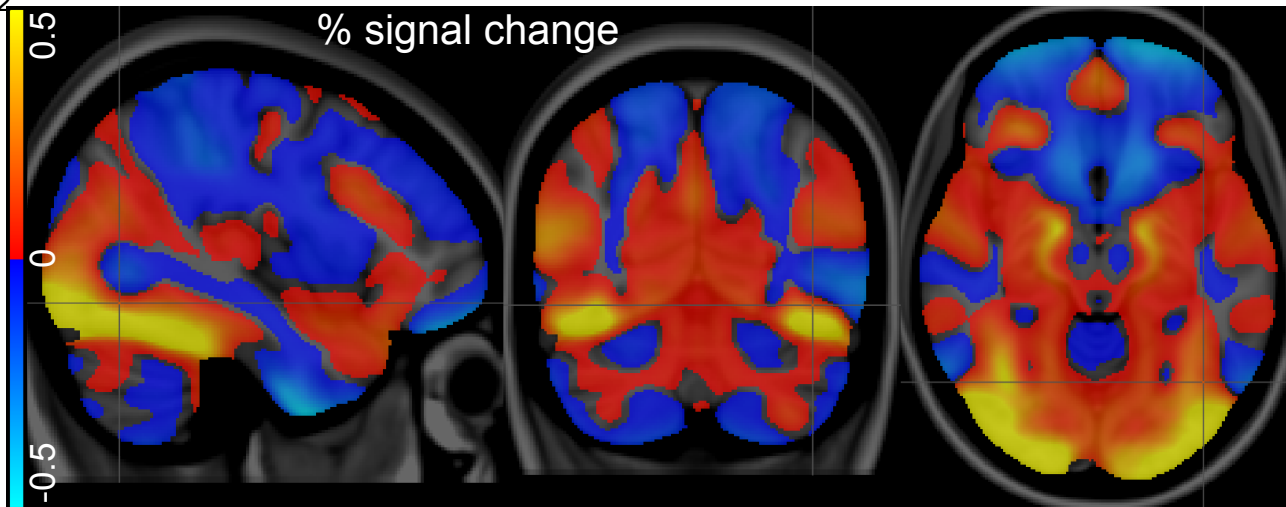
- Andersson, Jesper L.R., and Stamatios N. Sotiropoulos. 2016. "An Integrated Approach to Correction for off-Resonance Effects and Subject Movement in Diffusion MR Imaging." *NeuroImage* 125. The Authors: 1063–78. doi:10.1016/j.neuroimage.2015.10.019.
- Bearden, Carrie E, and Paul M Thompson. 2017. "NeuroView Emerging Global Initiatives in Neurogenetics : The Enhancing Neuroimaging Genetics through Meta-Analysis (ENIGMA) Consortium NeuroView." *Neuron* 94 (2). Elsevier Inc.: 232–36. doi:10.1016/j.neuron.2017.03.033.
- Duff, E. P., W. Vennart, R. G. Wise, M. A. Howard, R. E. Harris, M. Lee, K. Wartolowska, et al. 2015. "Learning to Identify CNS Drug Action and Efficacy Using Multistudy fMRI Data." *Science Translational Medicine* 7 (274): 274ra16-274ra16. doi:10.1126/scitranslmed.3008438.
- Fry, Anna, Thomas J. Littlejohns, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E. Allen. 2017. "Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population." *American Journal of Epidemiology* 186 (November): 1–9. doi:10.1093/aje/kwx246.
- Insel, Thomas, Bruce Cuthbert, Marjorie Garvey, Robert Heinssen, Daniel S. Pine, Kevin Quinn, Charles Sanislow, and Philip Wang. 2010. "Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders." *The American Journal of Psychiatry* 167 (7): 748–51. doi:10.1176/appi.ajp.2010.09091379.
- Le Floch, Edith, Vincent Guillemot, Vincent Frouin, Philippe Pinel, Christophe Lalanne, Laura Trinchera, Arthur Tenenhaus, et al. 2012. "Significant Correlation between a Set of Genetic Polymorphisms and a Functional Brain Network Revealed by Feature Selection and Sparse Partial Least Squares." *NeuroImage* 63 (1). Elsevier B.V.: 11–24. doi:10.1016/j.neuroimage.2012.06.061.
- Miller, Karla L, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, et al. 2016. "Multimodal Population Brain Imaging in the UK Biobank Prospective Epidemiological Study." *Nature Neuroscience*, no. 28: 1–9. doi:10.1038/nn.4393.
- Mueller, Susanne G., Michael W. Weiner, Leon J. Thal, Ronald C. Petersen, Clifford R. Jack, William Jagust, John Q. Trojanowski, Arthur W. Toga, and Laurel Beckett. 2005. "Ways toward an Early Diagnosis in Alzheimer's Disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI)." *Alzheimer's and Dementia* 1 (1): 55–66. doi:10.1016/j.jalz.2005.06.003.
- Munafò, Marcus R, Kate Tilling, Amy E Taylor, David M Evans, and George Davey Smith. 2017. "Collider Scope: When Selection Bias Can Substantially Influence Observed Associations." *International Journal of Epidemiology*, no. October (September): 1–10. doi:10.1093/ije/dyx206.
- Rosenbaum, Paul R, and Donald B Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41. doi:10.2307/2335942.
- Salimi-Khorshidi, Gholamreza, Gwenaëlle Douaud, Christian F. Beckmann, Matthew F. Glasser, Ludovica Griffanti, and Stephen M. Smith. 2014. "Automatic Denoising of Functional MRI Data: Combining Independent Component Analysis and Hierarchical Fusion of Classifiers." *NeuroImage* 90 (April). Elsevier B.V.: 449–68. doi:10.1016/j.neuroimage.2013.11.046.
- Smith, Stephen M, Thomas E Nichols, Diego Vidaurre, Anderson M. Winkler, Timothy E J Behrens, Matthew F Glasser, Kamil Ugurbil, Deanna M Barch, David C. Van Essen, and Karla L Miller. 2015. "A Positive-Negative Mode of Population Covariation Links Brain Connectivity, Demographics and Behavior." *Nature Neuroscience* 18 (11). Nature Publishing Group: 1565–67. doi:10.1038/nn.4125.
- Van Essen, David C., Stephen M. Smith, Deanna M. Barch, Timothy E J Behrens, Essa Yacoub, and Kamil Ugurbil. 2013. "The WU-Minn Human Connectome Project: An Overview." *NeuroImage* 80 (May). Elsevier Inc.: 62–79. doi:10.1016/j.neuroimage.2013.05.041.
- Volkow, Nora D., George F. Koob, Robert T. Croyle, Diana W. Bianchi, Joshua A. Gordon, Walter J. Koroshetz, Eliseo J. Pérez-Stable, et al. 2017. "The Conception of the ABCD Study: From Substance Use to a Broad NIH Collaboration." *Developmental Cognitive Neuroscience*, no. April. Elsevier: 1–4. doi:10.1016/j.dcn.2017.10.002.
- Westfall, Jacob, and Tal Yarkoni. 2014. "Statistically Controlling for Confounding Constructs Is Harder than You Think." *PloS One* 11 (3): e0152719. doi:10.1371/journal.pone.0152719.
- Winkler, Anderson M., Gerard R. Ridgway, Gwenaëlle Douaud, Thomas E. Nichols, and Stephen M. Smith. 2016. "Faster Permutation Inference in Brain Imaging." *NeuroImage* 141. The Authors: 502–16. doi:10.1016/j.neuroimage.2016.05.068.
- Winkler, Anderson M., Matthew A. Webster, Jonathan C. Brooks, Irene Tracey, Stephen M. Smith, and Thomas E. Nichols. 2016. "Non-Parametric Combination and Related Permutation Tests for Neuroimaging." *Human Brain Mapping* 0

(February): n/a-n/a. doi:10.1002/hbm.23115.

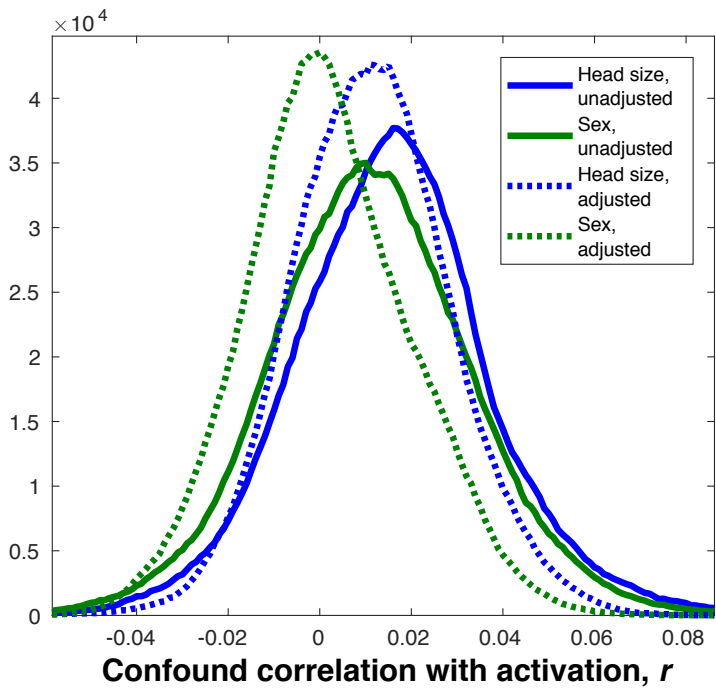
Winkler, Anderson M., Matthew a. Webster, Diego Vidaurre, Thomas E. Nichols, and Stephen M. Smith. 2015. "Multi-Level Block Permutation." *NeuroImage* 123 (December). Elsevier B.V.: 253–68. doi:10.1016/j.neuroimage.2015.05.092.

Figure 1

Sample size needed for 80% power



Functional activation % signal change
(group mean faces-shapes contrast $N = 12,600$)
range -0.5% : 1.2% min detectable signal 0.007%



Ranges	
Unadjusted correlation r	Sex -0.10 : 0.13
	Head size -0.11 : 0.12
Adjusted correlation r	Sex -0.08 : 0.10
	Head size -0.07 : 0.09

