NeuroImage: Clinical 12 (2016) 1045–1047

Contents lists available at ScienceDirect

NeuroImage: Clinical

journal homepage: www.elsevier.com/locate/ynicl

Editorial

Minimum statistical standards for submissions to Neuroimage: Clinical

The demand for reproducibility has reached fever pitch in scientific research (Baker, 2015), in particular in the field of psychology where many classic studies of human behaviour have not readily replicated (Klein et al., 2014). There are many explanations for poor replication, including subject selection bias, poor experimental control, inconsistent measurement, demand characteristics, post-hoc cherry picking of significant results, partial reporting and inadequate consideration of statistical power. While, all of these potential problems can also emerge in clinical neuroimaging, a lax statistical approach remains one of the most pernicious sources of error in our field.

Statistical issues have been discussed at length within the neuroimaging field in recent years, particularly in relation to procedures for correcting for multiple comparisons (Carter et al., 2016; Poldrack et al., 2008; Woo et al., 2014). Here we focus on multiple comparisons across image voxels, although the problem can become exacerbated when multiple hypotheses (contrasts) are tested at each voxel. While the issue of multiple comparisons is not new, it is clear that there is increasing concern that the application of relatively liberal statistical thresholds for declaring statistical significance is reducing confidence in reported effects, mainly due to weak control over the risk of false positive results. This concern is amplified in the clinical setting, where imaging plays an ever-increasing role in diagnosis and monitoring. *Neuroimage: Clinical* has therefore taken the decision to require minimum standards for correction for multiple comparisons when considering manuscripts. Manuscripts that do not meet these basic standards will normally not be considered for publication and will be returned to authors without review. Exceptions to this policy will only be made in rare cases, for example exploratory studies in particularly rare populations.

In addition to our desire to publish high quality science and reliable results, there is also a practical motivation for this decision. In our experience, manuscripts that do not meet these standards are invariably reviewed unfavourably, resulting in extra workload for authors and reviewers. We hope that by adopting a few basic standards we will not only increase the reliability of the results published in *Neuroimage: Clinical*, but also improve the efficiency of the review process for everyone.

Broadly speaking there are two areas of concern in relation to procedures used to control the family-wise error (FWE) rate (i.e. the probability of obtaining false positive results in a "family" of multiple tests): cluster-based inference and peak- or voxel-based inference. Our discussion mainly relates to magnetic resonance imaging (MRI) as the majority of submissions to our journal use this technique. However the same principles apply to other forms of neuroimaging in which large numbers of measurements are acquired (including, but not limited to: positron emission tomography, single-photon emission computed tomography, electroencephalography, magnetoencephalography and near infrared spectroscopy), or when comparisons are made across many nodes or edges of a brain network derived from neuroimaging data. For electrophysiological data modalities (magnetoencephalography and electroencephalography) we would recommend that authors adhere to the best practice guidelines outlined by Gross et al. (2013), and pay particular attention to recommendations in relation to the reporting of connectivity analyses.

1. Cluster-based inference

Cluster-based inference is one of the most commonly used methods to correct for multiple comparisons in submissions to *Neuroimage: Clinical* and is usually applied at the whole-brain level. This involves creating a statistical image (often a t-contrast) by setting an initial cluster-forming height threshold at some uncorrected *P*-value, and defining a minimum size that the resulting clusters of contiguous voxels must reach to be considered significant (though see "*Complementary Methods*" below and Smith and Nichols (2009) for discussion of an alternative approach known as threshold-free cluster enhancement: TFCE). A variety of principled procedures are available to determine whether the clusters that exceed that exceed the initial height threshold are statistically significant.

With this procedure the inference is made at the level of the entire cluster. While not central to the argument we wish to make in this editorial, it is worth briefly highlighting one of the drawbacks of cluster-based inference. If the initial height threshold results in very large clusters that encompass several brain regions then anatomical specificity is compromised, as inferences cannot be made about individual areas within clusters. While anatomical specificity may not be a central issue for some studies (i.e. those interested in whether an effect exists in the brain, rather than where), this remains a fundamental constraint on cluster-based inference. We recommend that authors do not incorrectly draw additional conclusions based on local results (e.g. reporting peak coordinate statistics) within a large cluster when using cluster-based inference.

One method for performing cluster-based inference is to set both the initial cluster-forming height threshold and the minimum cluster size arbitrarily. For example $P = 0.005$ (uncorrected), minimum cluster size 10 voxels has been recommended by some authors and is a default in some software tools (Lieberman and Cunningham, 2009). The problem with this unprincipled approach is that the corresponding FWE rate is unknown. And, more troubling, recent work has estimated the FWE rate for common implementations of this procedure to be between 60% and 90% (Eklund et al., 2016). Another problem is that clusters can vary dramatically in size depending on the size of the voxels (e.g. a 10-voxel cluster with 1 mm isotropic voxels is 27 times smaller than one with 3 mm isotropic voxels). Due to these

http://dx.doi.org/10.1016/j.nicl.2016.08.002

problems, *Neuroimage: Clinical* will not consider submissions that rely on such ad-hoc procedures to draw inferences.

It is important to note that the reporting of tables of statistical results using an arbitrary cluster-forming height threshold and arbitrary minimum cluster size is not in itself problematic. Indeed, we encourage this practice as such tables can be used for both exploratory analyses (Lieberman and Cunningham, 2009) and to facilitate meta-analysis (Eickhoff et al., 2012) (and we also encourage the sharing of unthresholded statistical maps through formal repositories such as www.neurovault.org (Gorgolewski et al., 2016) or others (Eickhoff et al., 2016)). However, inferences should only be made on results that survive correction for multiple comparisons using a statistically principled approach.

There are several different methods that are readily available that do rely on sound statistical principles to set the minimum extent threshold for cluster-based inference. These include random-field theory, permutation testing and Monte-Carlo simulation. We do not advocate any particular method or software package, and will consider manuscripts that use any of these approaches, provided that the method in question has been validated. However, the effectiveness with which some of these principled techniques control the FWE rate has been called into question, especially when the cluster-forming height threshold is liberal ($P = 0.01$ uncorrected or more liberal) (Eklund et al., 2016; Woo et al., 2014). Such liberal cluster-forming height thresholds also increase the cluster size, increasing the chance that clusters will encompass more than one region, which limits anatomical specificity as discussed above. When the cluster-forming height threshold is set at $P = 0.001$ the false positive rates, while in some cases inflated, are closer to 5% (Eklund et al., 2016; Woo et al., 2014). Therefore we encourage the use of more stringent initial height thresholds unless permutation testing is used to control the FWE rate.

2. Voxel-based inference

Voxel-level results (usually presented as the peaks of clusters), like cluster-level results, require statistically principled correction for multiple comparisons. Although it is rare for submissions to *Neuroimage: Clinical* to make inferences from uncorrected voxel-level *P*-values, or simply to use an arbitrary uncorrected *P*-value (e.g. $P = 0.001$) as a threshold for significance, it is worth stating that this is not good practice as the FWE rate is unknown. *Neuroimage: Clinical* will not consider submissions that draw inferences from uncorrected *P*-values.

In general, voxel-based correction for multiple comparisons is regarded as less powerful than cluster-based correction (Friston et al., 1996), though with this reduced power comes greater spatial precision on the location of effects. The level of correction required for whole-brain voxel-based analyses is very severe, and in our experience manuscripts submitted to *Neuroimage: Clinical* rarely use this approach. Instead, it is common for authors to apply an adjustment for small volume, which effectively limits the search space to one or more a priori specified regions of interest (ROIs), substantially reducing the severity of the correction (but also limiting anatomical inference to the volume of the ROIs). Small-volume adjustment can also be used for cluster-based inference but this is less common in our experience, presumably because clusters will often extend beyond the boundaries of small ROIs. When appropriately applied, a small-volume adjustment can be a perfectly valid inferential method. Indeed, the now mature history of fMRI experiments, across a range of cognitive tasks in healthy and pathological groups, is a reasonable argument that an ROI approach can be used in order to enhance the sensitivity of analysis. However, there are a number of practices that undermine its effectiveness.

The main issue relates to the method used to specify the ROIs. An important assumption of small-volume adjustment is that the method used to define the ROIs is independent of the analysis on which inference is made. Procedures that break this assumption suffer from “non-independence error”, which renders the resulting “corrected” *P*-values uninterpretable (Kriegeskorte et al., 2009). The most egregious practice of this type is to draw the ROIs after the analysis has been conducted, using the co-ordinates of the peak of the observed cluster to specify the centre of the ROI. *Neuroimage: Clinical* will not consider submissions that make this error. This is an example of a questionable research practice known as “selecting hypothesised areas after results are known” (SHARKing), which also occurs in less obvious forms (Poldrack et al., 2016). Therefore manuscripts submitted to *Neuroimage: Clinical* that make use of small-volume adjustment must state explicitly in the methods section how the ROIs were constructed, and additionally make it clear that they were specified independently of the results that are used to draw inferences.

Another approach for reducing the size of the search space for voxel-wise analysis involves executing an independent (orthogonal) contrast to the one used to draw inferences, and then using this as a mask (for example, only analysing interactions in regions that also show main effects). Again, this can also be applied to cluster-based correction. However, this approach is also vulnerable to non-independence error; for example if groups are of unequal sizes, any main effect across the groups will be biased towards effects that occur in the larger group (Kriegeskorte et al., 2009). Submissions to *Neuroimage: Clinical* must therefore provide sufficient information to allow the reader to understand precisely how any masks were constructed.

The final issue concerns the use of multiple independent ROIs. In itself this is not problematic – in many experiments it is quite plausible to hypothesise that a number of different regions might be activated by a task or differ between groups. However, if several ROIs are applied separately within the same analysis, this obviously increases the number of independent comparisons and therefore the FWE rate. For example, if five regions of the brain are considered to be of a priori interest, each defined on both the left and the right side, and each of these ten ROIs is used separately for small volume adjustment, then the FWE rate will be raised approximately 10-fold. There are two straightforward solutions to this issue: 1) Combining all ROIs into a single mask (which increases the severity of the correction applied within each ROI, as the search space increases in size); 2) Correcting for the number of separate ROIs applied, for example with a Bonferroni procedure.

3. Complementary methods

A fundamentally different approach for characterizing the error rates associated with multiple comparisons involves an estimation of the false discovery rate (FDR) (Benjamini and Hochberg, 1995). In the neuroimaging setting, FDR is designed to control the proportion of false positive voxels among the set of voxels that are labeled as showing significant results. It has gained rapid adoption for statistical testing in genomic studies, which face many of the same challenges resulting from multiple hypotheses testing as neuroimaging. There are a number of techniques for estimating the FDR at either the peak or cluster level of inference (Chumbley and Friston, 2009; Genovese et al., 2002).

Another alternative approach is TFCE (Smith and Nichols, 2009), a cluster-based inference method that removes the dependence on an arbitrary initial cluster-forming height threshold. TFCE considers all possible cluster-forming thresholds, and then creates an image summarising the cluster-wise evidence at each voxel. It can be more powerful than standard cluster-based methods, and is more spatially precise, as all cluster-forming thresholds are considered.

We expect these complementary approaches to be increasingly utilized by the neuroimaging community and encourage authors to become familiar with them.

In addition to these alternative statistical techniques, there are two other important strategies to reduce study bias. The first is pre-registration, in which the study design, analysis plan, predictions and boundary conditions are established a priori. This is common in clinical trials and there is no reason that strong predictions cannot be defined in clinical neuroimaging studies. Pre-registration of ROIs may be particularly useful as it guards against SHARKing. The second strategy is replication, which we have previously encouraged (Fletcher and Grafton, 2013). Replication has become a de facto requirement for many genetic studies, another field where false positive rates can be high, and *Neuroimage: Clinical* welcomes direct attempts at replications of previous studies.

4. Conclusion

We have discussed a number of statistical issues that, in our experience as editors at *Neuroimage: Clinical*, arise frequently during reviews for manuscripts submitted to our journal. Many of these practices invariably result in negative reviewer comments and rejection of manuscripts. Our intention here is not to be unduly prescriptive, and we recognise that different techniques will be appropriate for different experiments. However, manuscripts that do not meet even a bare minimum of statistical rigour, as outlined above, will normally be returned to authors without review.

Acknowledgments

We are grateful to Thomas Nichols, Michael Breakspear and Stephen Smith for comments on this editorial.

References

Baker, M.D., 2015. Over half of psychology studies fail reproducibility test. Nat. News (Aug 27, 2015).

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), pp. 289–300.

Carter, C.S., Lesh, T.A., Barch, D.M., 2016. Thresholds, power, and sample sizes in clinical neuroimaging. Biol. Psychiatry 1, 99–100.

Chumbley, J.R., Friston, K.J., 2009. False discovery rate revisited: FDR and topological inference using Gaussian random fields. NeuroImage 44, 62–70.

Eickhoff, S.B., Bzdok, D., Laird, A.R., Kurth, F., Fox, P.T., 2012. Activation likelihood estimation meta-analysis revisited. NeuroImage 59, 2349–2361.

Eickhoff, S.B., Nichols, T.E., Laird, A.R., Hoffstaedter, F., Amunts, K., Fox, P.T., Bzdok, D., Eickhoff, C.R., 2016. Behavior, sensitivity, and power of activation likelihood estimation characterized by massive empirical simulation. NeuroImage 137, 70–85.

Eklund, A., Nichols, T.E., Knutsson, H., 2016. Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. Proc. Natl. Acad. Sci. U. S. A. 113, 7900–7905.

Fletcher, P.C., Grafton, S.T., 2013. Repeat after me: replication in clinical neuroimaging is critical. Neuroimage Clin. 2, 247–248.

Friston, K.J., Holmes, A., Poline, J.-B., Price, C.J., Frith, C.D., 1996. Detecting activations in PET and fMRI: levels of inference and power. NeuroImage 40, 223–235.

Genovese, C.R., Lazar, N.A., Nichols, T.E., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 15, 870–878.

Gorgolewski, K.J., Varoquaux, G., Rivera, G., Schwartz, Y., Sochat, V.V., Ghosh, S.S., Maumet, C., Nichols, T.E., Poline, J.B., Yarkoni, T., Margulies, D.S., Poldrack, R.A., 2016. NeuroVault.org: a repository for sharing unthresholded statistical maps, parcellations, and atlases of the human brain. NeuroImage 124, 1242–1244.

Gross, J., Baillet, S., Barnes, G.R., Henson, R.N., Hillebrand, A., Jensen, O., Jerbi, K., Litvak, V., Maess, B., Oostenveld, R., Parkkonen, L., Taylor, J.R., van Wassenhove, V., Wibral, M., Schoffelen, J.M., 2013. Good practice for conducting and reporting MEG research. NeuroImage 65, 349–363.

Klein, R.A., Ratliff, K.A., Vianello, M., Adams Jr., R.B., Bahník, Š., Bernstein, M.J., Bocian, K., Brandt, M.J., Brooks, B., Brumbaugh, C.C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W.E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E.M., Hasselman, F., Hicks, J.A., Hovermale, J.F., Hunt, S.J., Huntsinger, J.R., IJzerman, H., John, M.-S., Joy-Gaba, J.A., Kappes, H.B., Krueger, L.E., Kurtz, J., Levitan, C.A., Mallett, R.K., Morris, W.L., Nelson, A.J., Nier, J.A., Packard, G., Pilati, R., Rutchick, A.M., Schmidt, K., Skorinko, J.L., Smith, R., Steiner, T.G., Storbeck, J., Van Swol, L.M., Thompson, D., van't Veer, A.E., Vaughn, L.A., Vranka, M., Wichmanm, A.L., Woodzicka, J.A., Nosek, B.A., 2014. Investigating variation in replicability. Soc. Psychol. 45, 142–152.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S., Baker, C.I., 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci. 12, 535–540.

Lieberman, M.D., Cunningham, W.A., 2009. Type I and type II error concerns in fMRI research: re-balancing the scale. Soc. Cogn. Affect. Neurosci. 4, 423–428.

Poldrack, R.A., Fletcher, P.C., Henson, R.N., Worsley, K.J., Brett, M., Nichols, T.E., 2008. Guidelines for reporting an fMRI study. NeuroImage 40, 409–414.

Poldrack, R., Baker, C.I., Durnez, J., Gorgolewski, K., Matthews, P.M., Munafo, M., Nichols, T., Poline, J.-B., Vul, E., Yarkoni, T., 2016. Scanning the Horizon: Future Challenges for Neuroimaging Research. bioRxiv.

Smith, S.M., Nichols, T.E., 2009. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. NeuroImage 44, 83–98.

Woo, C.W., Krishnan, A., Wager, T.D., 2014. Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. NeuroImage 91, 412–419.

J.P. Roiser
Institute of Cognitive Neuroscience, University College London, UK

D.E. Linden
Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, Cardiff, UK

M.L. Gorno-Tempinin
Department of Neurology, UCSF School of Medicine San Francisco, CA, USA

R.J. Moran
Virginia Tech Carilion Research Institute Roanoke, VA, USA

B.C. Dickerson
Department of Neurology, Harvard Medical School Boston, MA, USA

S.T. Grafton
Department of Psychological & Brain Sciences, University of California, Santa Barbara, California, USA
*Corresponding author.
E-mail address: scott.grafton@psych.ucsb.edu.