Everything you never wanted to know about circular analysis – but were afraid to ask

Nikolaus Kriegeskorte
MRC Cognition and Brain Sciences Unit, Cambridge, UK

Martin A. Lindquist
Department of Statistics, Columbia University; New York, USA

Thomas E. Nichols
Department of Statistics, University of Warwick; Warwick, UK
Department of Clinical Neurology, University of Oxford; Oxford, UK
Department of Biostatistics, University of Michigan; Ann Arbor, USA

Russell A. Poldrack
Department of Psychology and Neurobiology, University of Texas, Austin; Austin, USA

Edward Vul
Department of Brain and Cognitive Sciences, MIT; Cambridge, USA
Department of Psychology, UCSD; La Jolla, USA

Address ms. correspondence to: Edward Vul
                                   evul@mit.edu
                                   77 Massachusetts Ave, 46-4141
                                   Cambridge, MA, 02139
                                   USA
                                   Tel: 1 617 671 9741
                                   Fax: 1 617 258 8654

Abstract:

Over the last year a heated discussion about "circular" or "non-independent" analyses in brain imaging has emerged in the literature. An analysis is circular (or non-independent) if it is based on data that was selected for showing the effect of interest, or a related effect. The authors of this paper are researchers that have contributed to the discussion and span a range of viewpoints. In order to clarify points of agreement and disagreement in the community, we have collaboratively assembled a series of questions on circularity here, to which we provide our individual current answers in 100 words or less per question. While divergent views remain on some of the questions, there is also a substantial convergence of opinion, which we have summarized in a consensus box. The box provides the best current answers the five authors could agree upon.

Introduction

Brain imaging produces very large data sets of brain activity measurements. However, the neuroscientific conclusions in papers are typically based on a small subset of the data. The necessary selection – unless carefully accounted for in the analysis – can bias and invalidate statistical results (Vul et al. 2009; Kriegeskorte et al. 2009).

The large number of brain locations measured in parallel allows us to discover brain regions with particular functional properties. However, the more we search a noisy data set for active locations the more likely we are to find spurious effects by chance. This complicates statistical inference and decreases our sensitivity to true brain activation. In functional magnetic resonance imaging (fMRI), the goal is typically twofold: (1) to identify voxels that contain a particular effect and (2) to estimate the size of the effect, typically within a region of interest (ROI). Whether widely used analyses meet the resulting statistical challenges has been hotly debated in the past year.

Let's consider the first goal: finding brain regions that contain a particular effect. For example, we may wish to answer questions like: Which voxels respond more to faces than houses? Or, in which voxels does the face-house contrast correlate with IQ across subjects? The use of many null-hypothesis tests across brain locations presents a *multiple-testing* problem: the more voxels that are tested, the greater the *familywise error rate (FWE)*, i.e. the probability that one or more voxels will pass the significance threshold by chance even when there are no true effects *(false-alarm voxels)*. A number of statistical methods have been developed to control the FWE (for a review, see Nichols & Hayasaka 2003).

The *Bonferroni method* increases the significance threshold for each voxel to ensure that the FWE does not exceed, say, 0.05. However, since Bonferroni doesn't account for image smoothness, it is overly conservative and not optimally sensitive. *Random field theory methods* (Worsley et al. 1992; Friston et al. 1994) adjust for spatial correlation between voxels to achieve greater sensitivity (i.e., *power* -- the probability that a truly active voxel will be identified as such). While voxel-wise methods detect individual voxels, *cluster-wise methods* (Poline & Mazoyer 1993) report as significant clusters (contiguous sets of voxels that all exceed a primary threshold) that are larger than a pre-determined cluster-size threshold (chosen to ensure a 5% FWE for clusters).

Instead of limiting the *probability* of any false alarms (i.e. the FWE), *false-discovery rate (FDR)* methods (Genovese et al. 2002) limit the average *proportion* of false alarms among the voxels identified as significant. This approach promises greater sensitivity when there are effects in many voxels. When used appropriately, these methods solve the multiple-testing problem and ensure that we are unlikely to mistake an inactive region for an active one.

The second goal is estimating the size of the effect. For example we may wish to answer questions like: How strongly do these voxels respond to faces? Or, how highly does the activation contrast in this region correlate with IQ across subjects? Unfortunately, we cannot accurately address such questions by simply analyzing the selected voxels without worrying about the selection process. The effect-size statistics need to be independent of

the selection criterion; otherwise the results will be affected by "selection bias". For intuition, imagine the data were pure noise. If we select voxels by some criterion, those voxels are going to better conform to that criterion than expected by chance (for randomly selected voxels). Even if the selected voxels truly contain the effect of interest, the noise in the data will typically have pushed some voxels into the selected set and some others out of it, inflating the apparent effect in the selected set.

This problem has long been well-understood in theory, but is not always handled correctly in practice. Variants of bias due to selection among noisy effect estimates affect many parts of science. Just like *voxels* are selected by their signal level for inclusion in an *ROI*, so *studies* selected by their effect strengths for *publication* in scientific journals (Ioannidis 2005; 2008). Inflated effect estimates can results in either case.

Vul et al. (2009) suggested that cross-subject correlation neuroimaging studies in social neuroscience are affected by "non-independence" (see also Vul et al. 2010). Kriegeskorte et al. (2009) discuss the problem of "circularity" more generally as a challenge to systems neuroscience.

These authors argued that effect estimates and tests based on selected data need to be independent of the selection process, and that this can be ensured by using independent data for selection (e.g. using half of the data to select signal-carrying voxels, and the other half to estimate the signal) or by using inherently independent functional or anatomical selection criteria.

Although there is little controversy about the basic mechanism of selection bias, the 2009 papers have sparked a debate about exactly which analysis practices are affected and to what degree (Diener, 2009; Nichols & Poline, 2009; Yarkoni, 2009; Lieberman, Berkman, & Wager, 2009; Lazar, 2009; Lindquist & Gelman, 2009; Barrett, 2009; Vul et al, 2009b; Poldrack & Mumford, 2009). Here we collaboratively assembled and then individually answered a series of questions on circular analysis in order to clarify points of agreement and disagreement. Each answer is 100 words or less. We hope to contribute to a convergence within the community toward statistical practices that ensure that systems and cognitive neuroscience remain solidly grounded in empirical truth.

Scope of the problem

(1) Is circular analysis a problem in systems and cognitive neuroscience?

NK: Yes. A significant minority of papers is affected by distortions, which range from slight to severe.

ML: The term 'circular analysis' covers a wide variety of situations, whose severity range from fatal to relatively benign depending on situation and what type of information one is seeking. For example, using the same data to both train and test a classifier would be a huge problem; disqualifying any subsequent results. However, reporting the effect size over regions that survive a multiple comparisons analysis is less serious as long as the focus of the analysis is to detect regions with non-zero effect. However, even these estimates can be problematic if over-interpreted by readers/reviewers.

TN: Yes.

RP: Yes.

EV: Yes.

(2) How widespread are slight distortions and serious errors caused by circularity in the neuroscience literature?

NK: Without reanalyses the answer is elusive. My guess is that slight distortions are more widespread than severe distortions – but erroneous conclusions can result in either case.

ML: Most cases of circular analysis that I have seen in the literature relate to the practice of reporting effect size estimates over regions that survive a multiple comparisons analysis. If the focus of the analysis is null-hypothesis testing, then providing these values is not necessarily problematic. However, they must always be reported in their appropriate context, i.e. as *post hoc* estimates of effect size. The most widespread distortion is probably improper description of these estimates. This distortion has the potential of becoming serious if readers/reviewers read too much into these estimates and over-interpret their value.

TN: False positives due to circularity are minimal; biased estimates of effect size are common. False positives due to brushing off the multiple testing problem (e.g. "P<0.001 uncorrected" & crossing your fingers) remain pervasive.

RP: I think that slight distortions due to circularity are fairly common. I have no doubt that there have been serious errors due to circularity, including publication of empirical findings that are completely due to noise, but I think that this is relatively uncommon. I think that such serious problems are most likely to arise when circular analyses are combined with the use of uncorrected (or incorrectly corrected) whole-brain analyses.

EV: Slight distortions seem very common (maybe one third to one half of published fMRI papers contains some such distortion). I think that frequency falls off with the gravity of the error: Fortunately, the extreme cases (circular analysis on data selected using an inappropriately corrected selection procedure – potentially producing results from pure noise) are uncommon. Nevertheless, for the bulk of results obtained using circular methods, it is impossible to assess how large the distortions are until those results are reanalyzed using unbiased procedures.

Estimating effect sizes

(3) Are circular estimates useful measures of effect size?

NK: No.

ML: The standard approach of estimating effect sizes for voxels that survive an appropriate multiple comparisons threshold is biased and tends to overestimate the true effect size. In certain situations (e.g., when the effect size is large and/or the variance is small) this bias may be small and the estimate can still provide useful information. As a general principle, using an estimate with some bias is usually acceptable if it lowers the variance compared to other possible estimates. However, in this case I believe better unbiased estimates are available that are preferable (see Question 6).

TN: No. *However*, the bias in estimated effect size is a variable and diminishes to zero with increasing true effect size (i.e. increasing ratio of effect magnitude to standard deviation).

RP: Circular estimates will generally inflate the estimated effect size in comparison to non-circular analysis. Thus, I do not believe that circular estimates are valid measures of effect size.

EV: No – they are inflated to an unknown degree and are thus meaningless for inference.

(4) Should circular estimates of effect size be presented in papers and, if so, how?

NK: No.

ML: If the goal of the study is to estimate effect sizes then using circular estimates is not appropriate, and non-circular analyses are preferable. If the goal is hypothesis testing, I don't see an inherent problem in presenting them as long as they are placed in their appropriate context. That said, I am not sure how meaningful they are by themselves and I would urge readers not to over-interpret their value. In general, it may be useful to present any effect size estimate as confidence intervals, so readers can see for themselves how much uncertainty is related to the point estimate.

TN: Yes, if well described as circular and the detection of the effect itself was based on a valid inferential procedure.

RP: In general, they should not be presented unless accompanied by parallel non-circular effect size estimates (e.g., Poldrack & Mumford, 2009). However, if a researcher insists on presenting circular effect size estimates, then they should be presented with the clear caveat that they were estimated in a circular fashion.

EV: No. Because circular estimates can only be misleading, there is no good reason to present them. Nevertheless, if for some reason circular estimates are presented, they should come with a disclaimer to warn readers that they are looking at an un-interpretable number.

(5) Are effect size estimates important/useful for neuroscience research, and why?

NK: Yes, estimation of quantities is a key element of all sciences. Hypothesis testing by itself (A>0?, A>B?) yields an impoverished picture of the data and a loss of scientific insight. It also makes it more difficult to relate results across multiple studies and to assess the relevance of demonstrated relationships for practical applications (e.g. the diagnostic power of a given fMRI paradigm).

ML: Most neuroimaging studies to date have been focused on null-hypothesis testing and relatively little interest has been placed on estimating effect sizes. Whether that will change in the future I leave for the others to discuss. However, even in the null-hypothesis framework these estimates are useful. For example, accurate effect size estimates are needed for performing power analysis, which in turn are used to determine appropriate sample sizes for future experiments. Using biased estimates of effect size will ultimately lead to underpowered studies which can have serious ramifications.

TN: Depends on the question. Life is too short and research budgets too limited to require double experiments, n subjects to make inference on the location of an effect, another n subjects to get unbiased effect size estimates (note, splitting the data from n subjects doesn't yield two independent datasets). Hence the researcher has to choose whether they are after inference on location of an effect, or estimation of effect size assuming a known location.

RP: I think that effect size estimates may be useful in some circumstances. When we report our data, we generally wish to convey information regarding the strength of the effect, in order to provide a guide towards its importance. For example, we will be more impressed with a finding that activity in a particular region predicts 20% of the variance in a psychological trait than if it predicts 1%. That said, I don't think we have a really good notion of how big an effect needs to be in order to be considered "important" in neuroimaging.

EV: Yes, very much so. Null-hypothesis testing is insufficient for most goals of neuroscience because it can only indicate that a brain region is involved to some non-zero degree in some task contrast. This is likely to be true of most combinations of task-contrasts and brain-regions when measured with sufficient power. To determine how the brain produces cognition and behavior, cognitive neuroscience must answer questions like "which area is most responsible for this cognitive function", or "what cognitive function is this area most involved in", etc. These questions require evaluating effect sizes and comparing them across regions and tasks.

(6) What is the best way to accurately estimate effect sizes from imaging data?

NK: If the neuroscientific question requires selection of a subset of the data (e.g. a region of interest in the brain), then the selection process must not bias the effect estimate. We can either demonstrate that the effect statistic is inherently independent of the selection process and thus unaffected by selection (e.g. ROI definition by an anatomical or other statistically independent criterion), or we can use independent data (i.e. replications of the same experimental conditions) to estimate the effect for the selected subset of the data.

ML: There are a number of ways to accurately estimate effect sizes. For example, this can be done using anatomically defined ROIs or using various forms of cross-validation. We have in previous work suggested a method that selects voxels based on whether they show significant individual differences in the population (Lindquist et al. (2009)), as well as, a method based on using a multi-level Bayesian mixture model (Lindquist and Gelman (2009)). Developing methodology for accurately estimating effect sizes promises to be an area of active research in the future.

TN: Assume a known location and ROI for the effect, average data within that ROI and make univariate inference on that data.

RP: Through the use of independent/non-circular approaches. Reasonable approaches include the use of pre-specified anatomical regions of interest, independent localizer scans, split-half or cross-validation methods. It may also be reasonable to use regions identified from one contrast in a factorial design to estimate the effect of an orthogonal contrast, so long as it is confirmed that the effective regressors defined by the contrasts are truly orthogonal (which may fail for some rapid event-related designs; Kriegeskorte et al. 2009).

EV: Effect sizes should be estimated using unbiased, independent data. When considering a priori defined regions, this can be done with independent functional or anatomical localizers; if no regions of interest are known a priori, independence can be achieved by splitting a dataset and using various cross-validation methods.

(7) What makes data sets independent? Are different sets of subjects required?

NK: The statement 'data sets A and B are independent' requires qualification: What is independent between the data sets?
        (1) ...the within-subject measurement noise (requiring different sets of runs)?
        (2) ...the across-subject variation (requiring different subjects)?
        (3) ...everything?
We never want kind (3): total independence would mean there are no replicating true effects between data sets. We sometimes want independence of kind (2): for random-effects inference generalizing to the population. We always want independence of kind (1): for independent fitting and testing of single-subject models (e.g. defining a subject-specific ROI, i.e. binary weighting, and/or training a pattern classifier, i.e. continuous weighting).

ML: It depends on what type of question you want to answer. If you are interested in population inference (or prediction on future subjects) then splitting the data across subjects is required to ensure independent data sets. If, on the other hand, you are concerned with single subject inference (or prediction) then splitting within that subject will suffice. The problem of determining independence in the group setting is interesting, in part because statisticians rarely do test-training type splits and non-statisticians rarely worry about random effects. Here we need to think about both.

TN: Only datasets collected on distinct individuals can be assured to be independent. Splitting an individual's data (e.g. using run 1 and run 2 to create two datasets) doesn't yield independence at the group level, as each subject's true random effect will correlate the datasets. One exception is when the null hypothesis is true, as there are no random effects to correlate the pairs. Another exception is a fixed effects analysis, as there are no random effects to cause problems.

RP: For group analysis I think that independence will generally require different subjects, though cross-validation approaches like "leave one subject out" mean that one doesn't have to run an entirely different group. I think that simulated analyses where the null hypothesis is made true (e.g., by randomly permuting the outcome variable across subjects in a correlation analysis) are a good way to assess the validity of any proposed analysis.

EV: This depends on the inferences one wants to make at the group level. To claim that an effect exists in a specific anatomical region across the population ("Talairach coordinates (x,y,z) activate in response to reward prediction error"), one must validate with independent subjects. In contrast, to claim that a subject-specific ROI contains a certain effect across the population ("the subject-specific FFA respond more to faces than bodies"), one need only validate with independent data from the same subjects: The signal from an independently defined region can be compared at the group level using standard univariate statistics.

Beyond effect sizes: Other uses of non-independent selective analyses

(8) What information can one glean from data selected for a certain effect?

NK: Information independent of the selection criterion. However, it is not trivial to assess what aspects of the data are subtly related to the selection criterion and what aspects are strictly independent of it.

ML: It depends on how the data was selected. In the hypothesis testing framework one can state that given the assumptions, these brain areas show a non-zero effect. One should resist the temptation of trying to extract much more information than that.

TN: If the effect used for selection is nearly statistically independent from the tested effect (e.g. a main effect and an interaction), there are no limitations, as the circularity problem will be minimized. Totally dependent selection (e.g. pick the voxel with peak t-value) is essential for data visualization to ensure result is not due to outliers.

RP: These data cannot inform us regarding the strength of the effect, but we can gain general information about the nature of the effect and the degree to which assumptions of the statistical analyses may be violated.

EV: Only information independent of that selected effect. Consider data selected for (A+B)>0. Those data cannot be used to evaluate the magnitude of the A+B effect because the resulting measure will be inflated and biased. Those data also cannot be used to evaluate the dispersion or residuals (or model error) around the A+B effect because the measured dispersion will be biased to be lower. However, those data can be used to evaluate independent effects; for instance, the contrast A-B is often largely independent (given a balanced design matrix, but care must be paid to ensure that superficially orthogonal contrast vectors are indeed independent).

(9) Are visualizations of non-independent data helpful to illustrate the claims of a paper?

NK: They certainly help "tell the story" – an important part of scientific communication. To illustrate a hypothesis, however, it is entirely legitimate to include plots designed by hand. If a plot claims to present empirical evidence, the evidence should not be distorted. Nonindependent selection will tend to "clean up", appearing to give us both a view of the data and a clear illustration of our hypothesis. However, selection is akin to morphing between a data-based plot showing actual results and a hand-drawn plot illustrating the hypothesis. While each of these two is useful in its pure form, their amalgamation is misleading.

ML: I am hesitant to recommend less visualization, as they are critical for both model diagnosis and presentation of results. A visualization of non-independent data can help determine whether outliers are affecting the estimate. However, they can also lull readers' into believing the effect is stronger than in reality. Therefore it is useful to explicitly state that the plot is intended for diagnostic purposes and caution that the strength of the relationship not be over-interpreted. In general, always make visualizations as part of your analysis, but think carefully about which to include in your paper and how to present them.

TN: For experts in the field of neuroimaging, yes, such data visualization is crucial. For non-specialist (i.e. non-imaging-specific) journals, they should perhaps be relegated to supplementary material.

RP: In some cases, yes. For example, it can be difficult to interpret a significant interaction in a 2 X 3 ANOVA without visualizing the specific pattern that drives the result. In addition, I think that non-independent scatterplots should always be visualized (though not presented in a paper) for correlation analyses, in order to assess the presence of outliers. The failure to see any obvious outliers does not guarantee that the result is correct, but the presence of outliers should result in additional analyses to ensure that the result is not reliant upon a single observation (which occurs all too often with fMRI correlations).

EV: Sometimes. However, visualizations of non-independent data *are not* more helpful than hand-drawn illustrations of the effects of interest. Moreover, visualizations of non-independent data *are* more misleading, since they appear to have inferential weight, while they have none, this is exacerbated when non-independent data are plotted with measures of dispersion or error.

(10) Should data exploration be discouraged in favor of valid confirmatory analyses?

NK: No, data exploration is essential to scientific discovery. The cycle of exploration and confirmation can be closed within a single study, by using independent replications of the experiment. Central claims of a paper should be supported by valid confirmatory analyses. Additional unconfirmed exploratory results can be described as well, but it should be made clear that they are in need of future confirmation. Papers presenting only unconfirmed exploratory results should be discouraged. Unconfirmed results should never be cited without the caveat that they are yet to be confirmed.

ML: No. I don't view these types of analysis to be in direct competition with one another. There is a place for both, especially in an emerging field such as neuroimaging. The important thing is that it is clearly stated whether a result is obtained through data exploration or confirmatory analysis.

TN: No, they are both needed. Exploratory Data Analysis (EDA) is needed to ensure the data conforms to the distribution assumptions of the model (specifically, is not outlier-ridden), and discouraging EDA leads to black-box usage of software and analysis tools. If you can't trust a researcher to stick to their a priori hypotheses, can you even trust that their results aren't all Photoshop?

RP: Our understanding of brain function remains incredibly crude, and limiting research to the current set of models and methods would virtually guarantee scientific failure. Exploration of new approaches is thus critical, but the findings must be confirmed using new samples and convergent methods. For example, in the literature on resting-state fMRI, early findings using exploratory methods have been confirmed in subsequent studies using a variety of analysis techniques. Had the field insisted on a single approach to fMRI analysis (e.g., using task-based fMRI with the general linear model), the valuable insights from this body of work would have been missed.

EV: Absolutely not – data exploration is vital for science. However, it should be treated and presented as exploration: if a novel effect or phenomenon is suggested by an exploratory analysis, it should be validated by an independent confirmatory analysis before being treated as an established finding (i.e., being cited, incorporated into meta-analyses, etc.) Without independent validation, the outcomes of an exploratory analysis should be treated as well-reasoned speculation.

(11) Is a confirmatory analysis safer than an exploratory analysis in terms of drawing neuroscientific conclusions?

NK: An exploratory approach is more prone to selection bias: The more we explore, the greater the chance to find something in the noise. However, a confirmatory approach is more prone to model-misspecification bias: The more we assume, the greater the chance that our conclusions are built on sand.

ML: It depends on what type of conclusion you want to make. If the goal of the study is to test a certain null hypothesis with a certain false-positive rate than confirmatory analysis is safer. At the end of the day, confirmatory and exploratory analyses provide different ways of looking at the data. It is important to realize that the conclusions one can make are different and the results must be interpreted in their own contexts.

TN: Of course; EDA offers no inference.

RP: It depends on what risk one is protecting against. With regard to the risk of false positives, I think confirmatory analyses are probably safer. However, with regard to the more general risk of misunderstanding how the brain works, confirmatory analyses may be riskier since they bias us towards a very small portion of the hypothesis space.

EV: They each have their strengths and weaknesses. A confirmatory analysis is blind to new effects and hypotheses, while an exploratory analysis is susceptible to spurious fluctuations in the data. Both are necessary for scientific progress. However, while working within the null-hypothesis testing framework, confirmatory analyses provide the only legitimate basis for inference.

(12) What makes a whole-brain mapping analysis valid? What constitutes sufficient adjustment for multiple testing?

NK: Whole-brain analysis is exploratory with respect to brain space: we can discover new functional regions. The field has developed powerful methods that both explore the whole brain and confirm the result – with a single data set. These methods account for the multiple tests performed across locations. They are valid when the FWE or FDR is 5% or less. An uncorrected threshold of $p<0.001$ does not usually ensure this. Permutation methods can help reduce assumptions or estimate a given method's actual error rate (thus checking its assumptions).

ML: When performing a standard whole brain analysis, with separate hypothesis tests at each voxel, researchers should always use appropriate corrections for multiple comparisons. These include techniques that either control for the family-wise error rate or the false detection rate. In, addition it is important that the manner in which the correction has been performed be made explicit in the article, as this will guide the reader in interpreting the results.

TN: Control of FWE or FDR false positive risk over an a priori defined analysis mask, using an a priori specified statistic (e.g. voxel-wise or cluster-wise inference).

RP: I am satisfied with the use of any method that has been established to control familywise error or false discovery rate. I find nonparametric approaches (Nichols & Holmes, 2002) most appealing since they rely upon the fewest assumptions; they are computationally intensive, but increases in computing power have made them feasible for nearly all researchers.

EV: Although $p<0.05$ is just convention, it is a useful convention that has helped science proceed for the last 80 years, as such, I think it should be maintained until we have a justified alternative. In the meantime, I consider whole-brain analyses valid if they are corrected to keep the family-wise error rate below 5%. Unfortunately, such corrections are sometimes described cryptically, used incorrectly, and greatly reduce the power of a neuroimaging analyses. To avoid these problems, I think testing specific neuroanatomical hypotheses (using functionally or anatomically defined regions of interest) is generally preferable to exploratory whole-brain analyses.

(13) How much power should a brain-mapping analysis have to be useful?

NK: A brain-mapping analysis (or any other test) is only as useful as it is powerful (assuming that its specificity is controlled at 5% false positives). Lower power has three negative effects: (1) by definition: lower chance of finding something given that it's there, (2) lower information gained about the presence or absence of an effect (at 5% power, no information gained at all), (3) lower probability of a true effect given a positive finding (for any given prior probability of the effect), and thus a greater proportion of false positives in the literature due to publication bias for positive results.

ML: It is difficult to give an exact lower bound on the amount of power in 100 words, as it ultimately depends on the goal of the study. Perhaps the best answer is 'as much as possible'… Neuroimaging studies tend to be underpowered, as scanning subjects is expensive and time consuming. Underpowered studies can give rise to an increased number of false negatives, as well as a greater variability in effect size estimation. Hence, in studies with small sample size, large effect sizes may simply reflect the influence of random variation (Lindquist and Gelman (2009)).

TN: Chance of detecting one or more true positive voxels (or clusters) should be 80% or better. Easy to say, but such power is nearly impossible to calculate because of the myriad possible configurations of alternative hypotheses.

RP: Enough power to find an interesting effect, which begs the question of how big an effect has to be in order to be "interesting" (to which I don't think we have a good answer).

EV: Low-powered brain-mapping paints a misleading picture of neural function. Low-powered studies tend to find few punctate regions even if the underlying effect is diffuse over many, large areas; they amplify the adverse effects of publication bias; and they yield large distortions when combined with circular effect-size estimates. It is worth striving for about 80% statistical power. Such statistical power can be more easily achieved when testing specific neural hypotheses rather than conducting exploratory analyses over the whole brain.

(14) In which circumstances are non-independent selective analyses acceptable for scientific publication?

NK: Let's say 'never', and take an important step toward ensuring that our field remains solidly grounded in empirical reality. If there are aspects of the results that are demonstrably independent of the selection process, those aspects can be presented in isolation (making the selective analysis independent). If independent aspects don't exist, then the results are not useful. If isolating the independent aspects is difficult, data can almost always be either divided or replicated. Exceptions could be made if the cost outweighs the benefit of taking these steps, central claims are not concerned, and the analysis is clearly marked as circular.

ML: It ultimately depends on the purpose and goals of your study. When training a classifier I would be hard pressed to find a situation where it is acceptable. The same is true when the goal is to obtain an accurate effect size estimate. When performing null-hypothesis testing, I think it is acceptable to report an effect size as long as appropriate guidelines for interpreting the value are provided. It is, however, important that readers not be encouraged to inflate the importance or meaning of the reported estimates.

TN: Circular *inference* is never acceptable, but reporting circular effect sizes and plotting non-independent data can serve useful purposes when clearly marked as such.

RP: As the primary finding of a paper, I find non-independent analyses unacceptable. If presented in support of particular interpretations of the results, I think that they are potentially acceptable so long as they are clearly labeled as non-independent.

EV: I think non-independent selective analyses are never informative, because they will be biased to an unknown degree. However, if there is an exceptional reason for including such a biased analysis in a paper, it must be clearly demarcated as non-independent and presented with the caveat that the biased result is inferentially meaningless.

Consensus Box

(1) Is circular analysis a problem in systems and cognitive neuroscience?

> Yes.

(2) How widespread are slight distortions and serious errors caused by circularity in the neuroscience literature?

> Slight distortions are common, severe errors are less common. Insufficient correction for multiple comparisons in mapping analyses can aggravate the problem.

(3) Are circular estimates useful measures of effect size?

> No.

(4) Should circular estimates of effect size be presented in papers and, if so, how?

> Opinion is divided as to whether circular estimates should ever be presented. However, we all agree that if they are to be presented, it should be with explicit caveats regarding non-independence.

(5) Are effect size estimates important/useful for neuroscience research, and why?

> Yes. Although some questions may be answered with null-hypothesis testing alone, effect-size estimates help us judge the importance of the effect for brain function and practical applications. They are also the basis for calculations of statistical power.

(6) What is the best way to accurately estimate effect sizes from imaging data?

> The effect estimates should be independent of the selection criterion. This can be achieved by using an a-priori anatomical criterion or an independent functional contrast analysis to define the ROI. If the functional contrast is demonstrably independent of the effects to be estimated for the selected data, then the same data may be used for effect estimation. Otherwise, independent data are required to render the effect estimate independent.

(7) What makes data sets independent? Are different sets of subjects required?

> For population inference, independent subjects are required. For inference within the studied subjects (including effects in individually defined ROIs and individual regional pattern-information effects), independent data from the same subjects are sufficient. Within subjects, independence is assured by using separate sets of runs, which avoid correlation due to hemodynamic factors.

(8) What information can one glean from data selected for a certain effect?

Aspects of the data that are independent of the selection process may still be useful to examine. For example, effects demonstrated to be statistically independent of the selection criterion can be accurately estimated and interpreted. In certain circumstances, qualitative assessment of the model fit may also still be useful and help detect outliers.

(9) Are visualizations of non-independent data helpful to illustrate the claims of a paper?

While helpful for exploration and story-telling, circular data plots are misleading when presented as though they constitute empirical evidence unaffected by selection. Disclaimers and graphical indications of circularity (Kriegeskorte et al. 2009) should accompany such visualizations.

(10) Should data exploration be discouraged in favor of valid confirmatory analyses?

No!

(11) Is a confirmatory analysis safer than an exploratory analysis in terms of drawing neuroscientific conclusions?

Confirmatory analysis can support inference, exploratory analysis cannot – in that sense confirmatory analysis is safer. However, a confirmatory analysis is valid only to the extent that its assumptions hold. Exploratory analysis can help check those assumptions and generate new hypotheses.

(12) What makes a whole-brain mapping analysis valid? What constitutes sufficient adjustment for multiple testing?

Control of either the family-wise error rate (e.g., $p < 0.05$) or the false discovery rate (e.g., $q < 0.05$).

(13) How much power should a brain-mapping analysis have to be useful?

As much as possible. There is no consensus yet on the proper way to estimate power, or on the definition of a "large" effect size in neuroimaging.

(14) In which circumstances are non-independent selective analyses acceptable for scientific publication?

Inference based on non-independent selective analyses is not statistically sound and is never acceptable. If exploratory data analysis is done using non-independent analyses, these results must be presented with the appropriate disclaimers and caveats to alert readers about which conclusions can, and cannot, be drawn based on those data.

References:

Barrett , L.F. (2009) Understanding the Mind by Measuring the Brain: Lessons From Measuring Behavior (Commentary on Vul et al., 2009). Perspectives on Psychological Science, 4(3).

Friston, K.J., Worsley, K.J., Frackowiak, R.S.J., Mazziotta, J.C., and Evans, A.C. (1994) Assessing the significance of focal activations using their spatial extent. Hum. Brain Mapp. 1:214-220.

Genovese, C.R., Lazar, N.A., and Nichols T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 15(4), 870-8.

Ioannidis, J.P.A. (2005) Why Most Published Research Findings Are False. PLoS Med 2(8): e124. doi:10.1371/journal.pmed.0020124

Ioannidis, J.P. (2008) Why most discovered true associations are inflated. Epidemiology, 19(5) 640-648.

Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI. (2009) Circular analysis in systems neuroscience – the dangers of double dipping. *Nature Neuroscience* 12(5): 535-40.

Lazar, N.A. (2009) Discussion of "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition" by Vul et al. (2009). Perspectives on Psychological Science, 4(3)

Lieberman, M.D., Berkman, E.T., and Wager, T.D. (2009) Correlations in Social Neuroscience Aren't Voodoo: Commentary on Vul et al. (2009) . Perspectives on Psychological Science, 4(3)

Lindquist, M. and Gelman,A. (2009). Correlations and Multiple Comparisons in Functional Imaging: A Statistical Perspective (Commentary on Vul et al., 2009). *Perspectives on Psychological Science, 4, 310-313.*

Lindquist, M., Spicer, J., Leotti, L., Asllani, I., and Wager, T. Localizing areas with significant inter-individual variation: Testing Variance Components in a Multi-level GLM. *Human Brain Mapping Annual Meeting, 2009.*

Nichols, T.E., and Hayasaka, S. (2003). Controlling the Familywise Error Rate in Functional Neuroimaging: A Comparative Review. Statistical Methods in Medical Research, 12(5), 419–446.

Nichols, T.E. and Poline, J.-B. (2009) Commentary on Vul et al.'s (2009) "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition" Perspectives on Psychological Science, 4(3)

Poldrack, R.A. & Mumford, J.A. (2009). Independence in ROI analysis: Where is the voodoo? Social, Cognitive, and Affective Neuroscience,4, 208-213.

Poline, J.B., Mazoyer, B.M. (1993) Analysis of individual positron emission tomography activation maps by detection of high signal-to-noise-ratio pixel clusters. J Cereb Blood Flow Metab 13: 425– 437.

Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009). Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. Perspectives on Psychological Science, 4(3).

Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009b) Reply to Comments on "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition". Perspectives on Psychological Science, 4(3)

Vul, E & Kanwisher, N (2010) "Begging the question: The non-independence error in fMRI data analysis." in Hanson, S. & Bunzl, M (Eds.), Foundational issues for human brain mapping.

Worsley, K.J., Evans, A.C., Marrett, S., and Neelin, P. (1992) A three-dimensional statistical analysis for rCBF activation studies in human brain. J Cereb. Blood Flow Metab. 12:900-918.

Yarkoni, T. (2009) Big Correlations in Little Studies: Inflated fMRI Correlations Reflect Low Statistical Power. Commentary on Vul et al. (2009). Perspectives on Psychological Science, 4(3)