NeuroImage 103 (2014) 401–410

Contents lists available at ScienceDirect

NeuroImage

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

Selective correlations; not voodoo

J.D. Rosenblatt [a], Y. Benjamini [b,c,*]

[a] Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel
[b] Department of Statistics and Operations Research, The Sackler Faculty of Exact Sciences, Tel Aviv University, Israel
[c] The Sagol School of Neurosciences, Tel Aviv University, Israel

ARTICLE INFO

Article history:
Accepted 11 August 2014
Available online 19 August 2014

Keywords:
Social neuroscience
Selection bias
Effect size
Correlations
Voodoo

ABSTRACT

The problem of "voodoo" correlations–exceptionally high observed correlations in selected regions of the brain–is well recognized in neuroimaging. It arises when quantities of interest are estimated from the same data that was used to select them as interesting. In statistical terminology, the problem of inference following selection from the same data is that of *selective inference*. Motivated by the unwelcome side-effects of splitting the data–the recommended remedy–we adapt the recent developments in selective inference in order to construct confidence intervals (CIs) with good reproducibility prospects, even if selection and estimation are done with the same data. These intervals control the expected proportion of non-covered correlations in the selected voxels—the False Coverage Rate (FCR). They extend further toward zero than standard intervals, thus attenuating the impression made by highly biased observed correlations. They do so adaptively, in that they coincide with the standard CIs when far away from the selection point. We complement existing analytic proofs with a simulation, showing that the proposed intervals control the FCR in realistic social neuroscience problems. We also suggest a "confidence calibration plot", to allow the intervals to be reported in a clear and interpretable way. Applying the proposed methodology on a loss-aversion study, we demonstrate that with the sample size and selection type employed, selection bias is considerable. Finally, selective intervals are compared to the currently recommended data-splitting approach. We discover that our approach has more power and typically more informative, as no data is discarded.

Computation of the intervals is implemented in an accompanying software package.

Introduction

In the pursuit of brain regions that are highly correlated with behavioral measures (neural correlates), past practice has been to report correlations between the imaging measurements and behavioral attributes only in selected regions. These may have been selected based on the same correlation that will be reported. This practice has attracted condemnation for some time: Cureton (1950) refers to correlations reported in this manner as "baloney", with no hope of any meaningful interpretation (cited by Vul et al., 2009b). These early warnings were not echoed in the neuroimaging community until recently.

The implications of such uncontrolled *selective estimation* have been raised more recently by two provocative papers: Vul et al. (2009a) and Button et al. (2013). The problem raised by Vul et al. (2009a) is essentially that reported correlations between imaging attributes and behavioral attributes are "puzzlingly high". Using meta-analysis augmented with questionnaires, the researchers found that many published studies were likely to have applied selective estimation: the correlations reported are in locations selected based on these same correlations, thus justifying the names circular inference and double-dipping. These papers raised the awareness of the matter, not only through impressive meta-analysis, but also by provocative rhetoric. They were so influential that the original title of the former paper–"Voodoo Correlations"–has become an unofficial term for selection bias.

Initially, the many comments on Vul et al. (2009a), in the blogosphere and in the scientific literature (Fiedler, 2011; Poldrack and Mumford, 2009; Lazar, 2009; Lindquist and Gelman, 2009; Nichols and Poline, 2009; Yarkoni, 2009; Lieberman et al., 2009), were not in agreement on the source of the problem and the necessary course of action. Proposed causes included multiplicity control, reporting standards, sample size, sampling bias, and others.

To show the contribution of various different factors to this selection bias, we perform a simulation study. Fig. 1 reports the average estimated correlation following the selection stage. It demonstrates that selection bias is present whenever non-independent selection occurs, and that it can be quite considerable. Even large observed correlations, say $r = 0.8$, can stem from non-existing ones merely due to selection bias. Bias occurs in the presence or in the absence of a true effect, and will be present even if flawless control of multiplicity is performed

[*] Corresponding author at: Department of Statistics and Operations Research, The Sackler Faculty of Exact Sciences, Tel Aviv University, Israel.
E-mail address: ybenja@post.tau.ac.il (Y. Benjamini).

http://dx.doi.org/10.1016/j.neuroimage.2014.08.023
1053-8119/© 2014 Published by Elsevier Inc.

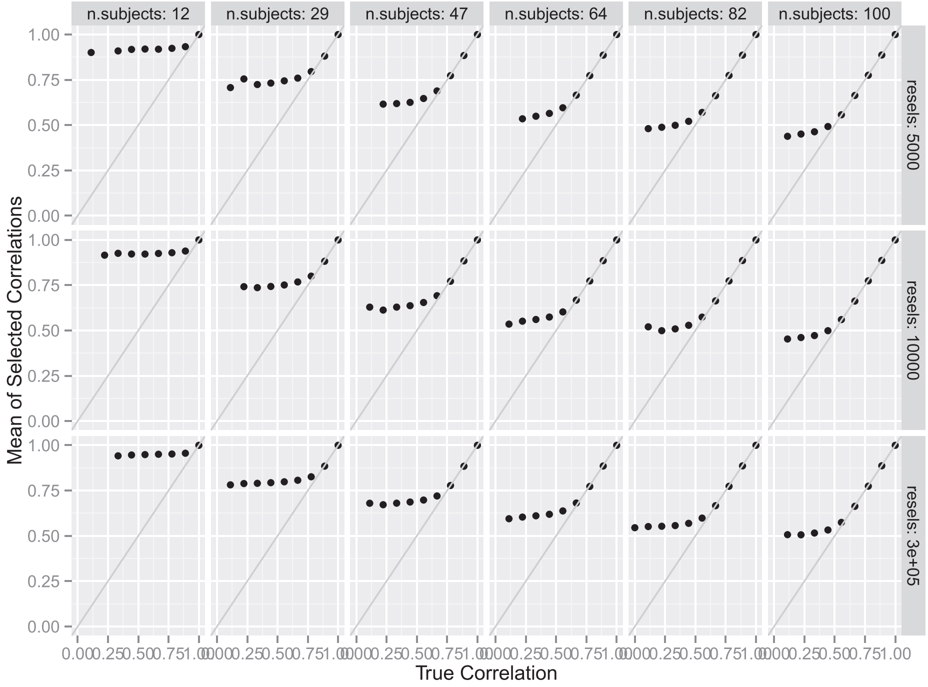

Fig. 1. The mean correlation surviving a selection stage. The figure demonstrates that selection bias is present whenever a non-independent data-driven parameter selection has been performed. The number of resolution elements (resels) varies from 5000 to $3 \cdot 10^5$ (across rows). The true underlying correlation varies from 0 to 1 (*x* axis), and is the same for all observations. The number of subjects underlying each observed correlation varies from 12 to 100 (across columns). Selection was performed so that the FWER is controlled at 0.05 using the Bonferroni procedure. See Appendix C.1 for more details. Simulation standard errors are nowhere above 0.032. Also note that the most extreme observed values are unbiased. This fact will be revisited when discussing the desired properties of our solutions.

(Bonferroni in our example). In fact, the more conservative the multiplicity control, the higher the selection threshold, so that only extreme correlations survive it. Finally, bias will occur even in very large samples, although it does decrease with sample size: the larger the sample size, the smaller the standard errors of the estimated correlation, the lower the selection threshold, and therefore the milder the selection bias. For a rough intuition regarding the effect of different parameters on the magnitude of the bias we refer the reader to Appendix A.

Imposing independence by splitting the data was the recommended remedy in Vul et al. (2009a), and shared by almost all commentators (Kriegeskorte et al., 2010; Fiedler, 2011; Poldrack and Mumford, 2009; Lazar, 2009; Lindquist and Gelman, 2009; Nichols and Poline, 2009; Yarkoni, 2009) While remedying bias, splitting the data introduces variance effects, making it an unattractive method when dealing with small samples. This matter is elaborated in the Splitting the data section.

Another unbiased approach is that of selecting parameters using the same data, but with a statistically independent criterion. This is implied in Kriegeskorte et al. (2010), and several examples of candidate statistics (albeit in a genetic setup) are suggested by Bourgon et al. (2010). If voxel-wise bias can be sacrificed for the sake of global accuracy, spatial priors in a Bayesian framework allow the spatial pooling of information for improved accuracy. We briefly comment on this view in Appendix B.

Here, we choose a different path, and demonstrate that it is possible to explicitly account for the selection stage at the estimation stage. Ultimately, we will show that:

- The bias introduced by circular inference can be accounted for by more than one way.
- It is typically preferable to account for the inherent bias in circular inference, rather than splitting a small sample to avoid it.

Our methods rest on confidence intervals (CIs) that offer coverage of population parameters, even after a biasing voxel/parameter selection. The Methods section presents two methods of selective confidence interval construction, which are directly relevant to voxel-based analysis. Both methods are sketched in the Overview subsection, leaving technical detail to following subsections, which can be skipped upon a first read. In the Results section we demonstrate the application of our intervals to the loss-aversion study by Tom et al. (2007). The Discussion section deals with shortcomings and possible extensions of the method: point estimates, cluster inference, choice of method, and duality between selection and estimation. Finally, and no less importantly, we identify areas for future research effort, which will be required in order to make selective estimation a readily available tool in researcher's arsenal. We believe this will be worthwhile because “voodoo correlations are everywhere—not only in neuroscience”, as the title of Fiedler (2011) states. We could not agree more.

Methods

Overview

A $(1 - \alpha)\%$ CI means that the population parameter will not be covered by the interval with a frequency of $\alpha\%$, over repeated experiments. When generalizing this error criterion to many parameters such as many voxel-wise correlations, or region-wise correlations, several candidate generalizations come to mind. The most natural candidates being control of the frequency of experiments where a parameter is not covered, and control of the expected proportion of non-covered parameters.[1]

The error measure we seek not only deals with a multitude of parameters, but also deals with the effect of selecting a subset of these. Fig. 2.1 depicts a case where 3 out of 20 candidate parameters were selected by a hypothesis test. When constructing 90% confidence intervals on all 20 parameters, 2 fail to cover, as expected. If focusing on the 3 parameters selected, 2 out of the 3 do not cover their underlying population parameter. This coverage is clearly worse than the 1 out of 10 error implied by the confidence level.

[1] The former leads to simultaneous coverage, and the latter is trivially satisfied by controlling the classical confidence level.

The False Coverage statement Rate criterion (FCR) (Benjamini and Yekutieli, 2005) is motivated by the above concerns and thus proposed as an error criterion.

Definition 2.1. False coverage statement rate

Define R_{CI} to be the number of selected parameters for which interval estimation is attempted, and V_{CI}, the number of such intervals that do not cover their underlying parameter. The proportion of non-covering intervals out of the intervals constructed, the false coverage statement proportion, is FCP $:= V_{CI}/R_{CI}$. Adopting the convention that no selection means no errors, then FCP $= 0$ if $R_{CI} = 0$. The false coverage statement *rate* is the expected false coverage statement proportion: FCR $= E$(FCP).

We now present two interval estimation methods which we later show to provide FCR control. These are "FCR-adjusted CIs", and "Conditional Quasi-Conventional (CQC) confidence intervals".

FCR-adjusted CIs, proposed in Benjamini and Yekutieli (2005), are simply standard intervals, the level of which is set to be more conservative by a factor that is equal to the proportion of the candidate parameters which are selected. Their construction is detailed in the FCR-adjusted confidence intervals section. Some useful properties of the FCR-adjusted confidence intervals include:

1. The FCR-adjusted CIs are dual to the Benjamini–Hochberg (BH; Benjamini and Hochberg, 1995) selection procedure: all the CIs constructed do not cover the tested parameter values. Put differently, the CIs and the selection mask agree on the locations where the inferred population correlation differs from 0.
2. For estimators that are statistically independent of one another, level α FCR-adjusted CIs provide FCR $\leq \alpha$ (Benjamini and Yekutieli, 2005, Theorem 1). This result is less appealing in neuroimaging as the spatial nature of the data will typically introduce statistical dependence between neighboring voxel-wise estimators. The next item deals with these cases.
3. For smooth Gaussian random fields (GRFs) of estimators, FCR-adjusted CIs after BH voxel selection, provide $\alpha/2 \leq$ FCR $\leq \alpha$. This follows from Benjamini and Yekutieli (2005, Theorems 2 & 3)[2], and the fact that GRFs in neuroimaging are positively regression dependent (Nichols and Poline, 2009).

When we select the estimators that are larger than a certain cutoff, we would not expect the selection bias to be symmetric, nor to be of equal magnitude around the cutoff. The bias clearly acts only in the direction of the selection, and vanishes for observations that are far away from the selection cutoff, as is evident from Fig. 1. The FCR-adjusted method, while being very simple to implement and catering to any selection rule, ignores this phenomenon: if the CI is symmetric, it remains so after the adjustment, and its conservatism does not diminish with distance from the selection boundary. This observation motivated Weinstein et al. (2013) to suggest several *conditional CIs* for the case of "above cutoff" selection, i.e., cases in which estimates are reported only if they surpass a cutoff *c*. Clearly selection by hypothesis testing is included in this definition.

The fundamental observation underlying conditional CIs is that the sampling distribution of selected estimators is simply a conditional one. It can thus be used in any way a classical sampling distribution is used: to construct CIs, to maximize likelihood, etc. We will discuss a certain type of CI constructed using this conditional distribution. Namely, Conditional Quasi-Conventional[3] (CQC) CIs.

Conditional intervals consist of all the parameter values which make observed correlations a sufficiently probable event in the presence of selection. The CQC intervals are a particular type of conditional intervals. Their uniqueness is in their aim to return parameter sets that not only render the observed correlation probable, but also have unambiguous sign (if possible).

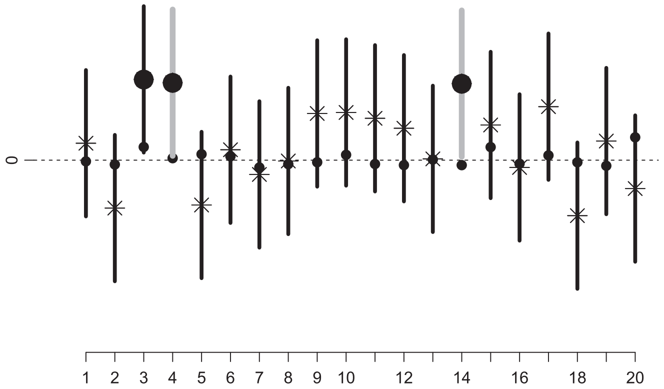

Fig. 2. Illustration of the false coverage proportion (FCP) of standard confidence intervals. Standard 1 – $\alpha = 0.9$ intervals constructed around each of 20 point estimates (asterisks) randomly drawn from non-null mean values (black dots). Two of the intervals fail to cover their underlying parameter, as expected. More importantly, out of the 3 point estimates that bear statistical significance (big black dots), 2 do not cover their underlying parameter (gray bars). The FCP is thus 2/3 which is far from the target level $\alpha = 0.1$.

Some useful properties of CQC CIs include the following:

1. CQC CIs will typically be asymmetric, reaching further toward the cutoff than away from it. Asymmetry and length decrease as the observed estimators depart from the selection cutoff. This holds for any symmetrically distributed estimator (e.g., linear contrasts), and will only be further accentuated with asymmetrically distributed estimators such as the correlation coefficients discussed in this manuscript.
2. If the selection cutoff, *c*, does not depend on the data, level α CQC CIs provide FCR $\leq \alpha$ (Weinstein et al., 2013, Section 1).
3. For estimators that are independent of one another, the selection may be done using the BH procedure with a false detection rate (FDR) $\leq q$. The cutoff, *c*, is now clearly data-dependent, and level α CQC intervals provide FCR $\leq \alpha$ (Weinstein et al., 2013, Theorem 2). This theorem is actually more general and applies to a broader class of so-called "simple" selection rules.

We now give the technical details for these two proposed approaches to selective CIs.

FCR-adjusted confidence intervals

FCR-adjusted CIs with FCR $\leq \alpha$ can make use of any valid (marginal) CI for each single parameter, as long as it is monotonic in the sense that CI's length grows as the confidence level increases. They are constructed as follows:

Algorithm 2.1. Level α FCR-adjusted CIs for simple selection rules:

1. Apply some selection rule to *m* parameter estimates, returning R_{CI} selected parameters.
2. For each selected parameter, construct the marginal CI at level

$$\alpha^* = \alpha \frac{R_{CI}}{m}.$$

Algorithm 2.1 is an adaptation of (Benjamini and Yekutieli, 2005, Definition 3) for "simple" selection rules, such as BH.

Definition 2.2. Simple selection rules

A selection rule is called *simple* if, for each parameter estimate, X_i, and when conditioning on the values of the other components, the

[2] Theorem 3 in Benjamini and Yekutieli (2005) is actually much more general, but stating it in full would require a lengthy digression.
[3] The name is taken from the quasi-conventional intervals presented by Benjamini et al. (1998).

total number of selected parameters (R_{CI}) is constant whenever X_i is among the selection.

Conditional quasi-conventional confidence intervals (CQC CIs)

We now formalize the definition of CQC CIs. Let X_i be an estimator of a parameter, θ_i, and Y_i be the same estimator conditional on it being selected above a cutoff, c:

$$Y_i = X_i \quad \text{only if} \quad |X_i| > c.$$

An acceptance region, $A_\alpha(\theta)$, of a test statistic, Y_i, is designed so that it satisfies two conditions for all θ: (i) The probability that Y_i falls within $A_\alpha(\theta)$ is α. (ii) The length of the interval that lies across 0 from θ is minimized, while penalizing for its overall length. Formally, if we denote by $B(A)$ the zero crossing part of an acceptance region, A, then we have:

$$A_\alpha(\theta) := \underset{A}{\operatorname{argmin}}\{\lambda|A| + |B(A)|\}$$

where λ is the penalty factor on the length of the interval (set to $\lambda = 2$ throughout this work).

A level α confidence set, $S_\alpha(y)$, for the parameter θ, based on an observation, y, and the above acceptance regions, $A_\alpha(\theta)$, is then:

$$S_\alpha(y) := \{\theta \quad s.t. \quad y \in A_\alpha(\theta)\}.$$

Current theory does not cover FCR control of CQC CIs in the case of a data-driven selection cutoff with statistical dependence between estimators. As dependence is clearly present in neuroimaging, we revert to simulations. We will aim to verify that CQC CIs offer FCR control on voxels selected using the BH procedure.

FCR of CQC intervals after BH selection

The main result of our simulation study shows that the FCR is indeed controlled by the CQC CIs under our observed dependence, even when the selection rule uses a data-driven threshold (BH). This can be seen from Fig. 3, where the simulated FCR (y axis) never surpasses the level for which the intervals were constructed (FCR = 0.1). The subfigures present results for different simulated scenarios, where effect magnitudes, prevalence, noise smoothness and number of subjects were varied (all ranges include the empirical values in our data). For simulation purposes, we assume that the sampling distribution of Pearson's correlation, after Gaussianization with a Fisher transformation, behaves like a smooth Gaussian stationary random field. Observations were therefore constructed from the sum of a signal field and a smooth Gaussian noise field with realistic smoothness ($FWHM_{mm} \approx \{3.3, 4.7, 5.7\}$).

Results

Armed with tools for selective interval estimation, we can now approach social neuroscience problems such as the ones discussed by Vul et al. (2009a)—in particular, the study performed by Tom et al. (2007). In this high profile study, which was revisited in the replies to Vul et al. (2009a), the authors attempted to localize brain regions associated with individuals' loss-aversion. This was done by correlating the behavioral loss-aversion index of 16 subjects with a neural loss-aversion index at each voxel. The data was organized, documented, and kindly made available via the *openfMRI* initiative at https://openfmri.org/dataset/ds000005.

In that study, high correlations were reported in 8 selected brain regions. These regions were selected using hypothesis tests on a robust version of this same correlation. Poldrack and Mumford (2009) later confirmed that these findings were indeed the result of selective

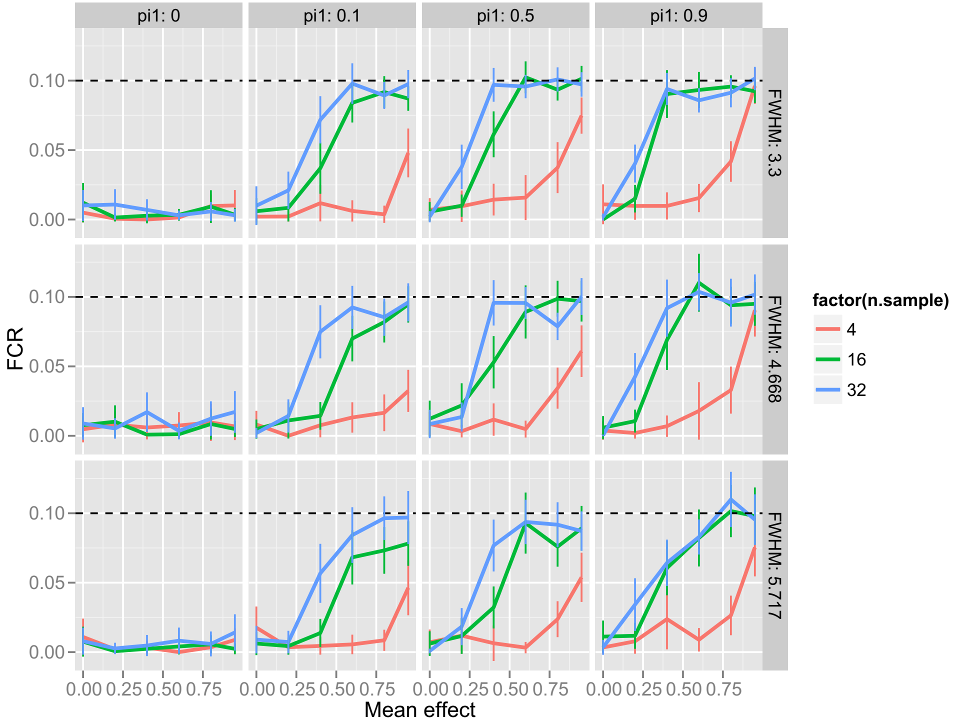

Fig. 3. Simulated FCR of the proposed CQC CIs under dependence for selected voxels. CQC CIs were constructed to control the FCR at level 0.1. Voxel selection was performed using the BH selection rule (FDR ≤ 0.1). Observations were constructed from the sum of a signal field and a smooth Gaussian noise field with a realistic smoothness varying over rows: $FWHM_{mm} \approx 3.3$, $FWHM_{mm} \approx 4.7$ (the observed smoothness), and $FWHM_{mm} \approx 5.7$. The mean effect displays the underlying correlation, ρ_i, varying from 0 to 0.9 (x axis), with the proportion of non-zero correlations, π_1, varying from 0 to 0.9 (columns). Each line represents a different number of subjects, varying through {4, 16, 32}. Simulated "brains" consist of 10 × 10 × 10 voxels. Vertical bars represent 2 standard errors of the simulated FCR. See Appendix C.2 for more details.

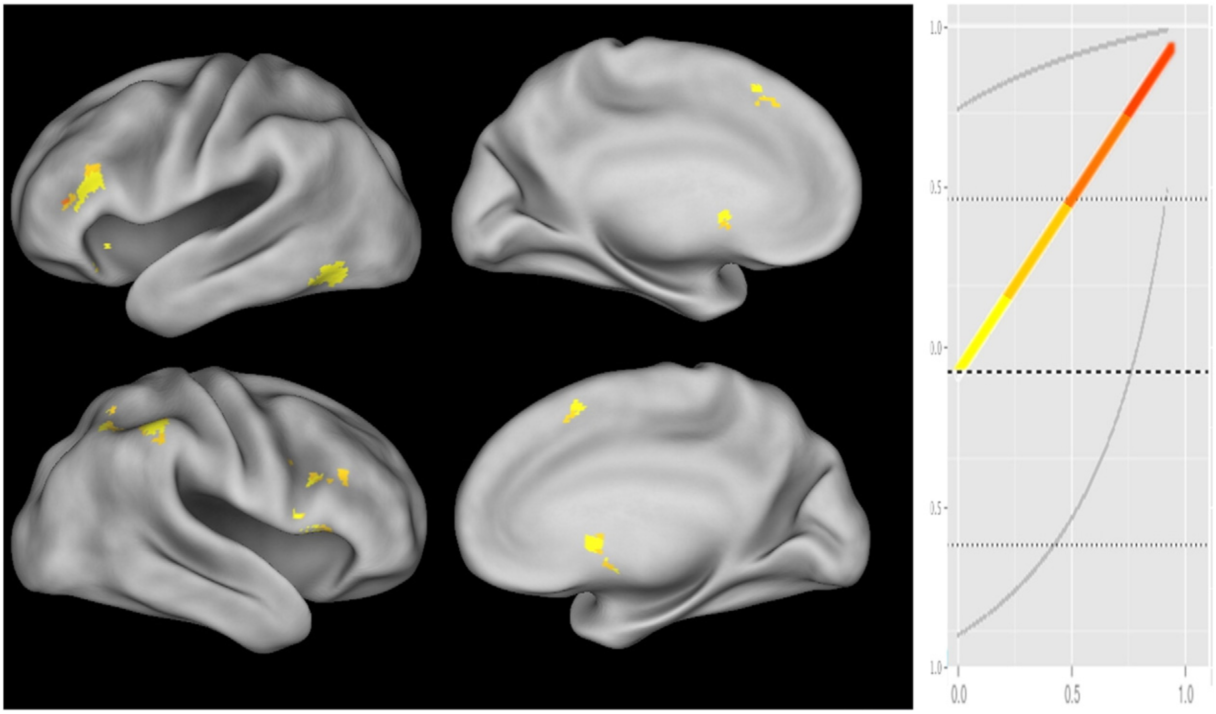

Fig. 4. A confidence calibration plot, based on observed (absolute) correlations in the voxels selected by Tom et al. (2007). The legend is adapted so that it encodes (in color) the observed values along the straight line, as well as the FCR-adjusted interval limits per observed correlation, by the curves above and below the observed correlation. This figure is equivalent to S6 in the supplementary material of Tom et al. (2007), except for reporting simple correlations instead of robust correlations. This is also why small correlations are encountered, even though all voxels surpassed a significance threshold.

estimation, and when controlling for the selection using a data split, an average upward selection bias of about 0.3 was discovered.

Applying FCR-adjusted CIs

Tom et al. (2007) used BH followed by cluster thresholding as their voxel selection rule. This rule is hard for us to condition upon, but luckily, this is not required for FCR-adjusted intervals.

For visualization of confidence limits in statistical parametric maps, we introduce the "confidence calibration plot" in Figs. 4 and 5. This is a parametric map whose legend is augmented by our confidence limits, encoded in the bands above and below the observed value. Fig. 4, for instance, informs us that an observed correlation of $r = 0.7$ could originate from a range of underlying population correlations between $\rho = -0.1$ and $\rho = 0.9$. Similarly, an observed correlation of $r = 0.8$ has bands roughly between $\rho = 0.25$ and $\rho = 0.95$.

Applying CQC CIs

FCR-adjusted intervals control the FCR in our case, but as previously stated, they do so without adapting to the degree of the bias at different observed values. As such, they are needlessly wide for strong observed correlations.[4] By contrast, CQC intervals do posses this desirable adaptive property. To demonstrate this, we use plain BH (FDR ≤ 0.1) voxel selection so that we may condition on the selection cutoff and apply CQC CIs. The result is again presented in a confidence calibration plot. The simulation results in the FCR of CQC intervals after BH selection section assure us that the CQC CIs in Fig. 5 control the FCR.

Comparing CQC and FCR-adjusted CIs

We now collapse the spatial information and compare CQC and FCR-adjusted CIs (using the same BH selection rule), as a function of the observed correlation (Fig. 6).

Both CQC CIs and FCR-adjusted CIs disagree with the preceding voxel selection. Put differently, masks constructed using BH and using any interval method would not be the same. This matter is discussed in the Duality of selection and estimation section.

The interpretation of Figs. 4 and 5 agrees in that after accounting for the selection, even large correlations can originate from very small underlying true ones. Using the CQC CIs in Fig. 6, we can conclude that with 16 subjects and BH selection at FDR ≤ 0.1, only observed correlations larger than $r \approx 0.75$ should be believed to be non-negative. Fortunately, the gradient of the CQC CIs is very steep in that region, so that observed correlations of slightly over $r = 0.8$ can be believed to originate from population correlations larger than $\rho = 0.5$.

Discussion

This work addresses the challenge of selective inference. We have shown, using simulation and social neuroscience data, that high observed correlations might be completely attributable to selection bias, and not to the true underlying correlations. This naturally holds for any noisy parameter estimate, and not only for correlations. Luckily, it can be accounted for in several ways, depending on researcher's goals, and can be easily communicated using the confidence calibration plot. We will now discuss several additional matters, namely: point estimation; how to interpret the FCR; why these CIs are better than splitting small samples; which of the selective CIs should be preferred; when the CIs can be seen as an inversion of their preceding hypothesis tests; and why more research in the field is of great importance.

[4] The tightening of the intervals in Fig. 4 is only due to the Fisher transformation of the correlations.

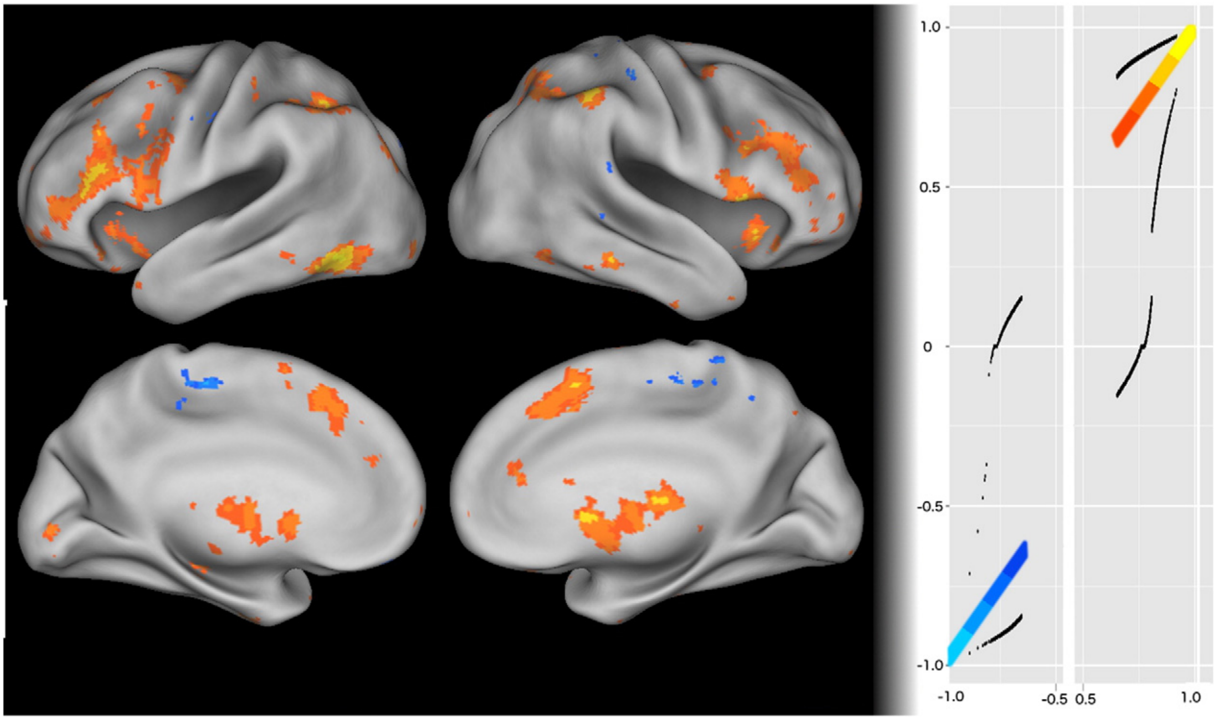

Fig. 5. A confidence calibration plot, based on observed correlations in significant voxels (BH; FDR ≤ 0.1). The legend is adapted so that it encodes (in color) the observed values along the straight line, as well as the CQC interval limits per observed correlation, by the curves above and below the observed correlation. Note that more voxels are selected than in Fig. 4, as no cluster thresholding was applied.

Point estimates

The reader might wonder why only interval estimators are suggested, and not point estimators. This is for both practical and principled reasons. In principle, we find that reporting confidence intervals is good scientific practice. We will not elaborate further on this matter as our view seems to be in line with recommended best practices in other domains (APA, 2001).

On the practical front, there is the matter of the appropriate error criterion. It turns out that adapting error criteria for selective estimation is not immediate. Once an adequate error criterion has been devised, one

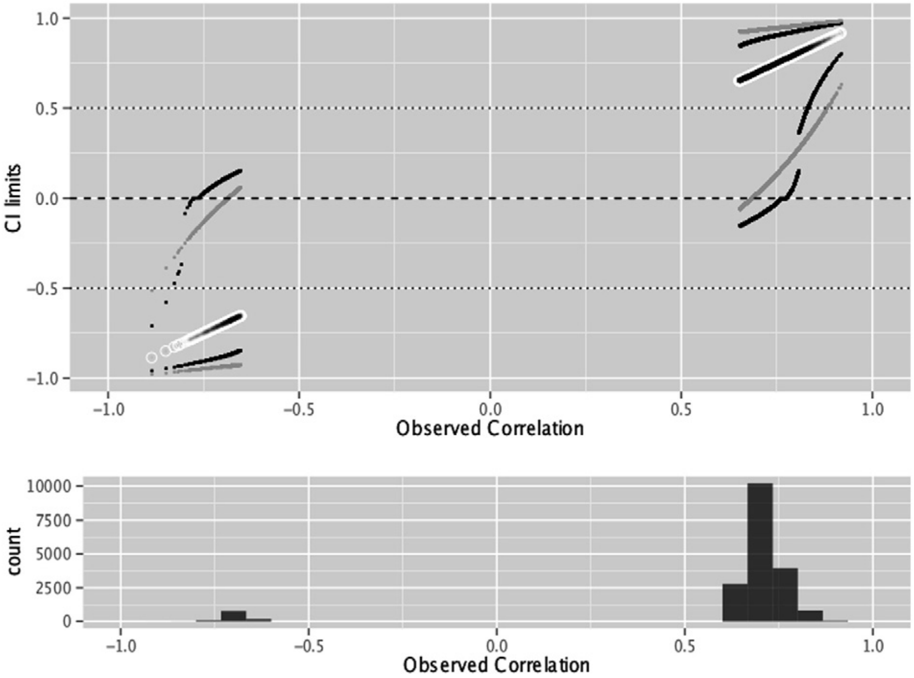

Fig. 6. Comparison of CQC and FCR-adjusted CIs (FCR ≤ 0.05) after BH voxel selection (FDR ≤ 0.1). Upper plot: The limits of the two CI methods are plotted against their underlying observed correlation. Observed correlation is on the *x* axis, with the identity line shown in black over white. CQC limits are shown in black and FCR-adjusted limits in gray. Lower plot: Histogram of observed correlations in the selected regions.

might suggest point estimators that satisfy this criterion. This is currently work in progress, as we are well aware of the preference of many researchers to report only point estimates.

Interpreting the FCR

Several clarifications regarding the meaning of the FCR are in order. First, it is an average over space and repeated experiments. It is thus possible (yet improbable) for a particular data set to be such that no interval covers its underlying parameter. Pathologies aside, we would expect that in a typical experiment the true unobserved false coverage proportion be close to the desired FCR. Under this view, one might interpret the FCR level as the probability of any particular (selected) parameter to have been covered by its corresponding interval. This view has a clear Bayesian flavor, and indeed the FCR, much like the false discovery rate, lends itself to a Bayesian interpretation.

Finally, we also stress that the FCR holds only over the entire selected set. If considering only a sub-region of selected voxels, then the false coverage proportions can be larger than desired.

Splitting the data

Using a data split, Poldrack and Mumford (2009) estimate the selection bias in Tom et al. (2007) to be in the order of + 0.3. This estimate might be misleading, since we know that different voxels incurred different biases. We expect the bias to be largest near the cutoff point, and to vanish as the point estimates depart from the selection boundary. This is not the case in Poldrack and Mumford (2009: Fig. 7 shows that the bias actually increases away from the selection. This could be explained by the fact that Poldrack and Mumford (2009) split the data along runs, not along subjects. As the authors themselves point out, this is not a truly independent analysis for group inference. The estimate of 0.3 is thus probably an underestimate.

In more general terms, let us now compare CQC CIs to standard CIs derived from split data, where voxels are selected based on a random half of the subjects, and correlations with standard CIs are computed using the second half. We start by quantifying the spatial agreement between the two methods. We denote by “SelectA(ll)” the voxel selection arrived at using BH (FDR < 0.1) and all of the subjects. Similarly, “SelectH(alf)” denotes the selection made using the same criterion on a random half of the subjects.

Inspecting the distribution of the proportions of voxels selected by both *SelectA* and *SelectH* over 50 random divisions of the subjects (half splits), we find that no more than 12% of the *SelectA* voxels were reselected by *SelectH* (and typically considerably less). Clearly, discarding half of the sample takes its toll on the power to localize activation. Cluster thresholding after selection does not add spatial stability. A cross-validation scheme will have to deal with the spatial disagreement in each data fold. Moreover, when using a cluster size threshold of 100 voxels, no voxels are selected in most of the 50 splits. This potential lack of spatial agreement over splits was also pointed out by Poldrack and Mumford (2009), but without quantifying its severity.

Now focusing on sign determination: We denote by “IntervalsA” the CQC intervals following a BH selection (FDR < 0.1) using all of the subjects. Similarly, “IntervalsH” indicates standard intervals computed based on the random half of subjects that were not used for voxel selection. It can be observed that in a situation such as ours, i.e. with 16 subjects and a threshold at $|r| > 0.65$, there is a small range of observed correlations, $|r| \in [0.72, 0.75]$, where *IntervalsH* do not cross 0 while *IntervalsA* do (Fig. 8). This is, however, where the advantage of *IntervalsH* ends, as it comes with a great cost in power to select and determine sign. As seen in Table 1, out of the voxels selected by either of the methods, voxels are more frequently selected by *SelectA* than by *SelectH* by a factor of about 5 (85.5/16.0). The sign is more frequently determined by *IntervalsA* than by *IntervalsH* by a factor of 9 (11.31/1.24).

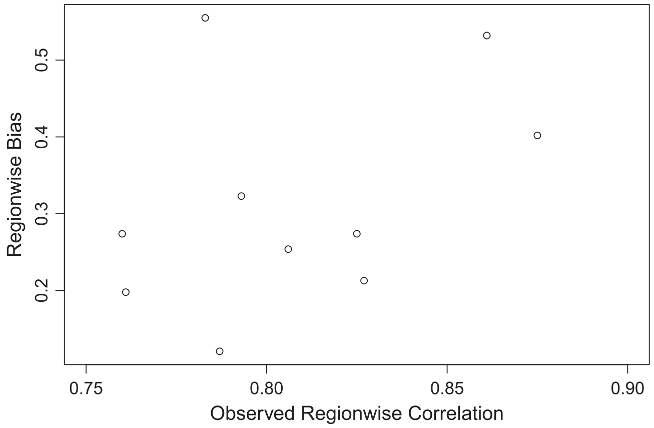

Fig. 7. The bias estimated using a data split versus the estimated region-wise correlation, as reported by Poldrack and Mumford (2009, Table 2, SOM).

Data splitting conclusions

Given that the split-data regions spatially disagree with the full-data regions, and signs are better determined using the full-data, it is difficult to recommend usage of the split-data approach with such small sample sizes, even if the split-data confidence intervals do have desirable FCR properties. Adding a cluster thresholding step will only worsen matters as more power will be lost. A robust correlation measure would not have improved things either.

Duality of selection and estimation

FCR-adjusted CIs and the BH FDR-controlling procedure are dual, in that no selected parameter will have a null-crossing CI (Benjamini and Yekutieli, 2005). In a similar way Bonferroni CIs and Bonferroni testing are dual, though the selection is more conservative and the intervals longer. The duality is often achieved by inverting the acceptance regions of tests into confidence intervals (Lehmann and Romano, 2005). If selection is done using hypothesis testing, agreement between the post-selection confidence intervals and the hypothesis test is expected. Alas, in practice, the two might not agree since the hypothesis test and the selective CIs can have different underlying assumptions or incompatible error rate controls. It might be of interest to develop CIs that are dual to the selection methods used in neuroimaging.

Returning to the duality between FCR-adjusted and BH selection: Fig. 4 reports FCR-adjusted CIs after a BH selection stage. This is a case where duality should hold, yet it clearly does not. This is because the selection was performed using a *robust* correlation and not Pearson's correlation for which the CIs were constructed. More importantly, the BH selection procedure was followed by a cluster thresholding stage, discarding clusters having less than 100 voxels. This selection is no longer a BH procedure. It has the effect of widening the FCR-adjusted CIs, so that more zero crossings will appear than would otherwise have done, had the cluster thresholding not been not applied.

Recommended selective CIs?

An important question remains: which of the two proposed selective CIs should be used? The answer depends on researcher's goals and available information. Naturally, the FCR-adjusted intervals are more general in that only the number of selected parameters matters, and not the specific selection rule. If the selection rule is such that it can be conditioned upon, then all proposed CIs are available and we examine Fig. 6 to offer the following rules of thumb: (1) For bounding parameter magnitudes distant from the null, which is natural after a two sided hypothesis test, the FCR-adjusted CIs are a good option—especially as they

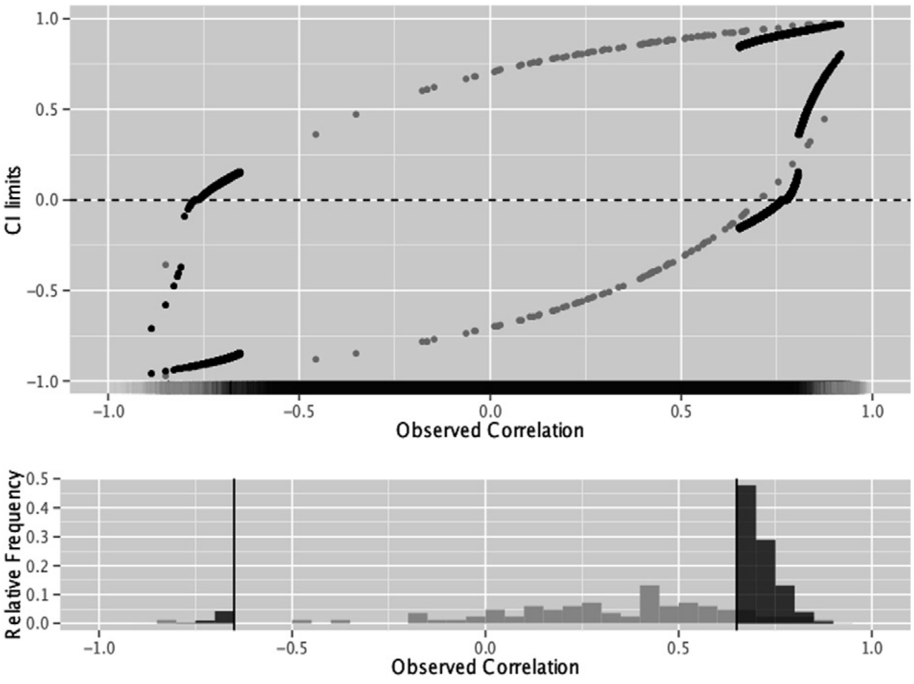

Fig. 8. CI limits versus observed correlation for CQC CIs (Intervals1, in black) and split-data CIs (Intervals2, in gray). Histograms demonstrate the empirical distribution of estimated correlations in selected regions. Note that spatial information is collapsed in this plot, and so it does not show that the same voxel has different point estimates under each strategy, nor that there is disagreement in the selected regions.

are dual to the BH selection rule. (2) If tight estimates of large observed values are of interest, conditional CIs account for the vanishing selection effect, returning smaller CIs for observed values far from the selection cutoff. This is clearly seen in Fig. 6, as the observed correlations change from $r = 0.8$ to $r = 0.85$.

Another recommendation that stems from our research is the importance of simplified selection rules. Acknowledging and reporting them can allow one to account for the selection bias at the estimation stage. By contrast, complicated selection rules yield intractable probabilistic problems.

Future research

After presenting the state of current research in Table 2, we suggest some additional research which is required to make selective confidence intervals a versatile instrument in researcher's toolbox.

Current state of research

As our smooth Gaussian random field is positively regression dependent (PRDS) (Nichols and Hayasaka, 2003), the simulation in Fig. 3 belongs to the (BH, conditional, PRDS) row.

In Tom et al. (2007), the authors selected voxels using a combination of the BH procedure followed by a cluster size threshold. We interpret this as a sequential rule: first BH and then voxel filtering (and not a simultaneous requirement of being selected by BH applied on the subset of voxels belonging to large enough clusters). This cluster size threshold is not covered by our simulations but we note that it is indeed a simple selection rule according to Definition 2.2. Being a simple rule, the voxel-wise FCR-adjusted CIs we constructed using the Tom et al. (2007) data belong to the (Simple, FCR-adjusted, PRDS) row of Table 2. If we were able to condition on the selection using the robust correlation, we could also construct conditional voxel-wise CIs that belong in the (Simple, Conditional, PRDS) row.

Table 1
Classification table of selection and sign determination, averaged over 50 random splits. A denotes a selected voxel with sign determination. B denotes a selected voxel without sign determination. C denotes non-selected voxels. Proportions are computed out of voxels selected by any of the procedures.

		IntervalsA (*CQC*)		
		Signed	Unsigned	Unselected
IntervalsH (*split*)	Signed	0.06%	0.18%	1.00%
	Unsigned	0.03%	0.41%	13.45%
	Unselected	11.22%	73.65%	0.00%

Other error measures

While very natural, the FCR is not the only candidate measure for the quality of selective estimators. The field of multiple comparisons has offered many error measures for selective *testing*, and selective *estimation* is no different in the sense that many error measures can be considered. A natural alternative is to control for the tail probability of the false coverage proportion $P(\mathrm{FCP} > \gamma) \leq \alpha$. Readers familiar with the multiple testing literature will recognize the analogy to the False Exceedance Rate criterion (FDX) (e.g., Genovese and Wasserman, 2006).

Another possibility could be the FCR over samples where something is actually selected. This intuitive criterion is impossible to control. To see the problem with such a criterion, consider repeated sampling from a single population with a single parameter. Also assume that a CI is constructed only when a type I error occurs. If the CI constructed is equivalent to the inverted hypothesis test, then the CI will never, by definition, cover the true parameter value. Readers might identify this conditional FCR as an estimation counterpart of the positive FDR (pFDR) (Storey et al., 2004), which is also discussed by Benjamini and Yekutieli (2005, p. 81).

Inference on parameter aggregates

A particular challenge in the studies presented is that the estimates were reported not on the selection units (voxels), but rather on their aggregates (clusters). A derivation of the distribution of region aggregates is thus required for interval construction. This holds true whether

Table 2
Current knowledge of FCR control of different selection–estimation schemes, and different dependency classes between estimators.

Selection	CIs	Dependence	FCR control proved	FCR control conjectured
Fixed cutoff	FCR-adjusted	Any	Yes	–
Fixed cutoff	Conditional	Any	Yes	–
Simple	FCR-adjusted	PRDS	Yes	–
Simple	Conditional	Independent	Yes	–
BH	Conditional	PRDS	–	Yes
Simple	Conditional	PRDS	–	?

regions are selected using BH or any other selection rule, such as random field theory with family wise error rate control.

Confidence intervals for selection

Once selective confidence intervals are made available, a natural use would be to select voxels so that the interval at a given voxel or region is adequately informative. This implies using the CIs for the selection itself, and not only for the estimation. This avenue is worth pursuing, with the caveat that a new layer of circularity might be introduced this way, and so may require attention.

Spatial stability of selection

We have argued that selected regions using small samples are extremely unstable. This observation was based on BH selection with 16 subjects split into two groups of 8. It is clearly worth exploring the generality of these claims, and particularly the effect of the number of subjects and the selection method.

A small sensitivity analysis suggests that with our small sample sizes, selection is extremely sensitive to the sample size. Indeed, using BH selection (no cluster thresholding) 14 subjects pick up 37% of the regions selected by 16 subjects, on average over random splits. A decrease to 12 subjects will return only 25% of the 16 subject regions. Finally, as previously stated, 8 subjects pick up no more than 12%.

Acknowledgments

Yoav Benjamini and Jonathan Rosenblatt were supported by a European Research Council under EC–EP7 European Research Council grant PSARPS-297519.

Russ Poldrack deserves special thanks for his availability and willingness to help reproduce a 6 year-old research study. Not many labs are capable of such information retrieval. Thomas Nichols for encouraging us to put our ideas in writing. Matthijs Vink and Omri Perez for helping improve the manuscript. Neomi Singer and Talma Hendler helped us to navigate endless software programs and file formats in order to reproduce previous studies. The R implementation would not have been possible without the valuable work of Jonathan Clayden for the *tractor.base* R package (Clayden et al., 2011), and of Pierre Lafaye de Micheaux on the *AnalyzeFMRI* package (Bordier et al., 2011).

Appendix A. Upper bound on selection bias

An approximate upper bound on the bias introduced by selecting voxels using their p-values can be derived by noting that: (a) correlations below the selection cutoff are never reported; and (b) given the selection cutoff, the bias is non-increasing as the effect departs from the null. This implies that the maximal bias occurs under the null hypothesis. Assuming that the number of subjects is large enough, so that the distribution of the correlations under the null is concentrated at the selection boundary, we get the following (approximate) upper bound for the bias:

$$bias(r,\rho) \preccurlyeq tanh\left(\Phi^{-1}\left(1-\frac{p_{cut}}{2}\right)/\sqrt{n-3}\right) \tag{A.1}$$

where r is the observed correlation, ρ is the underlying population correlation, $\Phi^{-1}(.)$ is the Gaussian inverse cumulative distribution function, $\preccurlyeq$ is an approximate bound, and p_{cut} is the selection cutoff in p-value scale.

For validation, compare to the simulation in Fig. 1: taking $n = 29$ in the first row, the reported correlation is between 0.7 and 0.75 for a wide range of underlying correlations, ρ. Using our formula, we calculate that the bias should be no larger than 0.7. This is indeed a tight bound if the true correlation is about $\rho = 0$, but extremely conservative if it is large.

Appendix B. A Bayesian view

Some authors suggest a Bayesian approach for dealing with selection bias (e.g. Lindquist and Gelman, 2009). There is indeed a beautiful way to correct biased estimators using Bayesian priors (Firth, 1993). This approach, however, does not deal with selection bias. The typical Bayesian approach to imaging is to shrink point-estimates to the global mean using spatial priors, i.e., to pool information from different brain regions with different parameter values. This would introduce bias everywhere in the brain. The larger the spatial variability of the parameter, the larger the bias introduced. If done before cluster selection, this would bias effects toward zero. If done after cluster selection, it would shrink all selected voxels toward the (biased) mean of the selected voxels.

There are cases where the researcher cares more about average global accuracy than voxel-wise unbiasedness. For these cases, accuracy can surely benefit from the pooling of spatial information, possibly using spatial priors in a Bayesian framework (Lindquist and Gelman, 2009). Note, however, that voxel-wise unbiasedness and global average accuracy are conflicting goals (Vaart, 1998, Lemma 8.13), and the voxel-wise bias introduced when considering global accuracy can be considerable. To see this, consider a brain with two types of correlations, balanced over voxels: either $\rho = 0$ or $\rho = 0.9$. If voxel-wise measurements are noisy, a globally accurate estimator will return mid-range values in all voxels. It would be globally accurate, but nowhere accurate in a voxel-wise sense.

Appendix C. Simulation details

C.1. Demonstration of selection bias

Here are the details of the simulation underlying Fig. 1.

Each point reports the average and two standard errors of the selected correlations in a single replication. The number of independent observations varied from 5000 to 3×10^5 across the rows. These numbers represent the number of independent resolution elements (resels) in our data and the number of voxels, which would be the number of independent resolution elements assuming no spatial correlation. The number of resels (≈ 5000) was obtained by estimating the smoothness of the correlation's z-score field. This z-score field was obtained by applying a voxel-wise Fisher transform (Fisher, 1915). We then used the *SmoothEst* function in the *AnalyzeFMRI* R package (Bordier et al., 2011) to estimate the field's smoothness, and computed the number of resels by dividing the volume (in mm^3) by the full width at half maximum (FWHM, in mm) in each direction:

$$\frac{\text{volume}}{FWHM_x \times FWHM_y \times FWHM_z}.$$

The number of subjects underlying each correlation, in the columns, varied from a small experiment ($n = 12$) to a large one ($n = 100$).

The true underlying signal, on the x-axis, varied from a null correlation ($\rho = 0$) to a perfect correlation ($\rho = 1$). This was added to the noise field after a Fisher transformation.

Correlations were selected using the Bonferroni procedure ($FWER \leq 0.05$).

Simulation code is available at https://github.com/johnros/SelectiveEstimationSimulations.

C.2. Demonstration FCR control of CQC CIs under dependence

Here are the details of the simulation underlying Fig. 3.

At each point, the averaged proportion of voxel-wise CIs covering their respective parameter is reported, with 2 standard errors.

In each of the 100 replications, a "brain" of $10 \times 10 \times 10$ voxels was generated. Observations were constructed based on the sum of a signal field and a smooth Gaussian noise field.

The noise field consists of a Gaussian white noise field convolved with a 3D isotropic Gaussian density function with varying smoothness parameter. The observed smoothness in our data corresponds to the middle row.

The signal field consists of signals of strength $\mu_i(\rho_i)$, spread as a Poisson point process with rate $\pi_1 = 1 - \pi_0$ over the whole brain volume. To convert the signal to z-scale before adding it to the noise field, we used the Fisher transform (Fisher, 1915): $\mu_i(\rho_i) = arctanh(\rho_i)$. ρ_i varied from 0 to 0.9 (x axis). $1 - \pi_0$ varied from 0 to 1 (columns). The number of subjects underlying each observations varies through {4, 16, 32}.

Selection of voxels in each replication was performed using the BH procedure (FDR $\leq$ 0.1).

Testing of the coverage was performed on the z-scale, using the BH cutoff converted to that scale.

In the selected voxels, we constructed CQC CIs using our *selectiveCI* R package, as described in the Conditional quasi-conventional confidence intervals (CQC CIs) section, and implemented at https://github.com/johnros/selectiveCI.

The selection cutoff, z_{cut}, used was the Gaussianized BH cutoff:

$$p_{cut} = \frac{\alpha}{m} \times R$$
$$z_{cut} = \Phi^{-1}\left(1 - \frac{p_{cut}}{2}\right).$$

We then tested which of the CQC CIs cover their respective parameter, and averaged over voxels to get the FCP in that replication, taking $FCP = 0$ if no voxels were selected.

Simulation code is available at https://github.com/johnros/SelectiveEstimationSimulations.

References

APA, 2001. Publication Manual of the American Psychological Association. American Psychological Association, Washington DC.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B 57, 289-289 (0).

Benjamini, Y., Yekutieli, D., 2005. False discovery rate-adjusted multiple confidence intervals for selected parameters. J. Am. Stat. Assoc. 1000 (469), 71–81 (0).

Benjamini, Y., Hochberg, Y., Stark, P.B., 1998. Confidence intervals with more power to determine the sign: two ends constrain the means. J. Am. Stat. Assoc. 930 (441), 309–317. http://dx.doi.org/10.2307/2669627 (0, Mar.. ISSN 0162-1459. URL http://www.jstor.org/stable/2669627. ArticleType: research-article / Full publication date: Mar., 1998 / Copyright © 1998 American Statistical Association).

Bordier, C., Dojat, M., Micheaux, P.L.D., 2011. Temporal and spatial independent component analysis for fMRI data sets embedded in the AnalyzeFMRI R package. J. Stat. Softw. 440 (9), 1–24 (0. URL http://www.jstatsoft.org/v44/i09/).

Bourgon, R., Gentleman, R., Huber, W., 2010. Independent filtering increases detection power for high-throughput experiments. Proc. Natl. Acad. Sci. 1070 (21), 9546–9551. http://dx.doi.org/10.1073/pnas.0914005107 (0, May. ISSN 0027-8424, 1091-6490. URL http://www.pnas.org/content/107/21/9546).

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 140 (5), 365–376. http://dx.doi.org/10.1038/nrn3475 (0, May. ISSN 1471-003X. URL http://www.nature.com/nrn/journal/v14/n5/full/nrn3475.html).

Clayden, J.D., Maniega, S.M., Storkey, A.J., King, M.D., Bastin, M.E., Clark, C.A., 2011. TractoR: magnetic resonance imaging and tractography with R. J. Stat. Softw. 440 (8), 1–18 (0. URL http://www.jstatsoft.org/v44/i08/).

Cureton, E.E., 1950. Validity, reliability and baloney. Educ. Psychol. Meas. 10, 94–96. http://dx.doi.org/10.1177/001316445001000107 (0. ISSN 1552-3888(Electronic); 0013-1644(Print)).

Fiedler, K., 2011. Voodoo correlations are everywhere—not only in neuroscience. Perspect. Psychol. Sci. 60 (2), 163–171. http://dx.doi.org/10.1177/1745691611400237 (0, Mar. ISSN 1745-6916, 1745-6924. URL http://pps.sagepub.com/content/6/2/163).

Firth, D., 1993. Bias reduction of maximum likelihood estimates. Biometrika 800 (1), 27–38. http://dx.doi.org/10.1093/biomet/80.1.27 (0, Mar. ISSN 0006-3444, 1464-3510).

Fisher, R.A., 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 100 (4), 507–521 (0).

Genovese, C., Wasserman, L., 2006. Exceedance control of the false discovery proportion. J. Am. Stat. Assoc. 1010 (476), 1408–1417. http://dx.doi.org/10.1198/016214506000000339 (0, Dec. ISSN 0162-1459, 1537-274X).

Kriegeskorte, N., Lindquist, M.A., Nichols, T.E., Poldrack, R.A., Vul, E., 2010. Everything you never wanted to know about circular analysis, but were afraid to ask. J. Cereb. Blood Flow Metab. 300 (9), 1551–1557. http://dx.doi.org/10.1038/jcbfm.2010.86 (0, Sept. ISSN 1559-7016).

Lazar, N.A., 2009. Discussion of "puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition" by vul et al. (2009). Perspect. Psychol. Sci. 40 (3), 308–309. http://dx.doi.org/10.1111/j.1745-6924.2009.01129.x (0, May. ISSN 1745-6916, 1745-6924. URL http://pps.sagepub.com/content/4/3/308).

Lehmann, E.L., Romano, J.P., 2005. Testing Statistical Hypotheses. Springer Science & Business Media 9780387988641, (Apr.).

Lieberman, M.D., Berkman, E.T., Wager, T.D., 2009. Correlations in social neuroscience aren't voodoo: commentary on vul et al. (2009). Perspect. Psychol. Sci. 40 (3), 299–307. http://dx.doi.org/10.1111/j.1745-6924.2009.01128.x (0, May. ISSN 1745-6916, 1745-6924. URL http://pps.sagepub.com/content/4/3/299).

Lindquist, M.A., Gelman, A., 2009. Correlations and multiple comparisons in functional imaging: a statistical perspective (commentary on vul et al., 2009). Perspect. Psychol. Sci. 40 (3), 310–313. http://dx.doi.org/10.1111/j.1745-6924.2009.01130.x (0, May. ISSN 1745-6916, 1745-6924. URL http://pps.sagepub.com/content/4/3/310).

Nichols, T., Hayasaka, S., 2003. Controlling the familywise error rate in functional neuroimaging: a comparative review. Stat. Methods Med. Res. 120 (5), 419–446. http://dx.doi.org/10.1191/0962280203sm341ra (0, Oct. ISSN 0962-2802, 1477-0334. URL http://smm.sagepub.com/content/12/5/419).

Nichols, T.E., Poline, J.-B., 2009. Commentary on vul et al'.s (2009) "puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition". Perspect. Psychol. Sci. 40 (3), 291–293. http://dx.doi.org/10.1111/j.1745-6924.2009.01126.x (0, May. ISSN 1745–6916, 1745–6924. URL http://pps.sagepub.com/content/4/3/291).

Poldrack, R.A., Mumford, J.A., 2009. Independence in ROI analysis: where is the voodoo? Soc. Cogn. Affect. Neurosci. 40 (2), 208–213. http://dx.doi.org/10.1093/scan/nsp011 (0, June. ISSN 1749-5016, 1749-5024. URL http://scan.oxfordjournals.org/content/4/2/208).

Storey, J.D., Taylor, J.E., Siegmund, D., 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. J. R. Stat. Soc. Ser. B 660 (1), 187–205. http://dx.doi.org/10.1111/j.1467-9868.2004.00439.x (0).

Tom, S.M., Fox, C.R., Trepel, C., Poldrack, R.A., 2007. The neural basis of loss aversion in decision-making under risk. Science 3150 (5811), 515–518. http://dx.doi.org/10.1126/science.1134239 (0, Jan. ISSN 0036-8075, 1095-9203. URL http://www.sciencemag.org/content/315/5811/515).

Vaart, A.W.V.D., 1998. Asymptotic Statistics. Cambridge University Press, Cambridge, UK; New York, NY, USA 9780521496032, (Oct.).

Vul, E., Harris, C., Winkielman, P., Pashler, H., 2009a. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspect. Psychol. Sci. 40 (3), 274–290. http://dx.doi.org/10.1111/j.1745-6924.2009.01125.x (0, May. ISSN 1745-6916, 1745-6924. URL http://pps.sagepub.com/content/4/3/274).

Vul, E., Harris, C., Winkielman, P., Pashler, H., 2009b. Reply to comments on "puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition". Perspect. Psychol. Sci. 40 (3), 319–324. http://dx.doi.org/10.1111/j.1745-6924.2009.01132.x (0, 2009. ISSN 1745-6916, 1745-6924. URL http://pps.sagepub.com/content/4/3/319).

Weinstein, A., Fithian, W., Benjamini, Y., 2013. Selection adjusted confidence intervals with more power to determine the sign. J. Am. Stat. Assoc. 1080 (501), 165–176. http://dx.doi.org/10.1080/01621459.2012.737740 (0. ISSN 0162-1459. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.2012.737740).

Yarkoni, T., 2009. Big correlations in little studies: inflated fMRI correlations reflect low statistical power—commentary on vul et al. (2009). Perspect. Psychol. Sci. 40 (3), 294–298. http://dx.doi.org/10.1111/j.1745-6924.2009.01127.x (0, May. ISSN 1745-6916, 1745-6924. URL http://pps.sagepub.com/content/4/3/294).