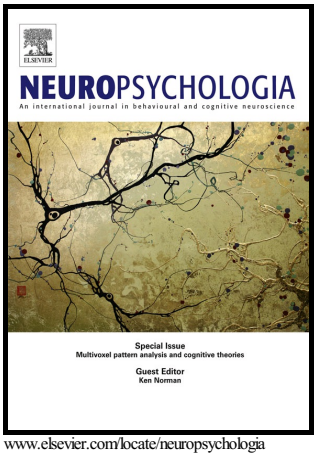
The impact of sample size on the reproducibility of voxel-based lesion-deficit mappings

Lorca-Puls, Diego L.; Gajardo-Vidal, Andrea; White, Jitrachote; Seghier, Mohamed L.; Leff, Alexander P.; Green, David W.; Crinion, Jenny T.; Ludersdorfer, Philipp; Hope, Thomas M. H.; Bowman, Howard; Price, Cathy J.

DOI:
10.1016/j.neuropsychologia.2018.03.014



www.elsevier.com/locate/neuropsychologia

The impact of sample size on the reproducibility of voxel-based lesion-deficit mappings

Diego L. Lorca-Puls, Andrea Gajardo-Vidal, Jitrachote White, Mohamed L. Seghier, Alexander P. Leff, David W. Green, Jenny T. Crinion, Philipp Ludersdorfer, Thomas M.H. Hope, Howard Bowman, Cathy J. Price

PII: S0028-3932(18)30107-6
DOI: https://doi.org/10.1016/j.neuropsychologia.2018.03.014
Reference: NSY6717

To appear in: *Neuropsychologia*

Received date: 23 June 2017
Revised date: 8 March 2018
Accepted date: 9 March 2018

Cite this article as: Diego L. Lorca-Puls, Andrea Gajardo-Vidal, Jitrachote White, Mohamed L. Seghier, Alexander P. Leff, David W. Green, Jenny T. Crinion, Philipp Ludersdorfer, Thomas M.H. Hope, Howard Bowman and Cathy J. Price, The impact of sample size on the reproducibility of voxel-based lesion-deficit mappings, *Neuropsychologia,* https://doi.org/10.1016/j.neuropsychologia.2018.03.014

The impact of sample size on the reproducibility of voxel-based lesion-deficit mappings

Diego L. Lorca-Puls[a,b], Andrea Gajardo-Vidal[a,c], Jitrachote White[a], Mohamed L. Seghier[a,d], Alexander P. Leff[e,f], David W. Green[g], Jenny T. Crinion[e], Philipp Ludersdorfer[a], Thomas M. H. Hope[a], Howard Bowman[h,i], and Cathy J. Price[a]

[a] Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, London WC1N 3BG, United Kingdom.

[b] Department of Speech, Language and Hearing Sciences, Faculty of Medicine, Universidad de Concepcion, PO Box 160-C Concepcion, Chile.

[c] Department of Speech, Language and Hearing Sciences, Faculty of Health Sciences, Universidad del Desarrollo, 4070001 Concepcion, Chile.

[d] Cognitive Neuroimaging Unit, Emirates College for Advanced Education, PO Box 126662 Abu Dhabi, United Arab Emirates.

[e] Institute of Cognitive Neuroscience, Division of Psychology and Language Sciences, University College London, London WC1N 3AR, United Kingdom.

[f] Department of Brain Repair and Rehabilitation, Institute of Neurology, University College London, London WC1N 3BG, United Kingdom.

[g] Department of Experimental Psychology, Division of Psychology and Language Sciences, University College London, London WC1H 0AP, United Kingdom.

[h] Centre for Cognitive Neuroscience and Cognitive Systems and the School of Computing, University of Kent, Canterbury CT2 7NF, United Kingdom.

[i] School of Psychology, University of Birmingham, Birmingham B15 2TT, United Kingdom.

Corresponding author: Diego L. Lorca-Puls, Wellcome Trust Centre for Neuroimaging, Institute of Neurology, UCL, 12 Queen Square, London, WC1N 3BG, UK. E-mail: d.puls@ucl.ac.uk

Abstract

This study investigated how sample size affects the reproducibility of findings from univariate voxel-based lesion-deficit analyses (e.g., voxel-based lesion-symptom mapping and voxel-based morphometry). Our effect of interest was the strength of the mapping between brain damage and speech articulation difficulties, as measured in terms of the proportion of variance explained. First, we identified a region of interest by searching on a voxel-by-voxel basis for brain areas where greater lesion load was associated with poorer speech articulation using a large sample of 360 right-handed English-speaking stroke survivors. We then randomly drew thousands of bootstrap samples from this data set that included either 30, 60, 90, 120, 180, or 360 patients. For each resample, we recorded effect size estimates and *p* values after conducting exactly the same lesion-deficit analysis within the previously identified region of interest and holding all procedures constant. The results show (1) how often small effect sizes in a heterogeneous population fail to be detected; (2) how effect size and its statistical significance varies with sample size; (3) how low-powered studies (due to small sample sizes) can greatly over-estimate as well as under-estimate effect sizes; and (4) how large sample sizes ($N \geq 90$) can yield highly significant *p* values even when effect sizes are so small that they become trivial in practical terms. The implications of these findings for interpreting the results from univariate voxel-based lesion-deficit analyses are discussed.

Keywords

voxel-based; lesion-symptom; lesion; deficit; reproducibility; stroke; speech production

1. Introduction

There is a great deal of evidence showing how both false positive and false negative results increase as sample size decreases (Bakker et al., 2012; Button et al., 2013a; Chen et al., 2018; Cremers et al., 2017; Ingre, 2013; Ioannidis, 2008) and how inadequate statistical power can lead to replication failures (Anderson et al., 2017; Bakker et al., 2012; Perugini et al., 2014; Simonsohn et al., 2014a; Szucs and Ioannidis, 2017). However, the impact of sample size on false negative and false positive rates has never been quantified in mass-univariate voxel-based lesion-deficit mapping (e.g., voxel-based lesion-symptom mapping and voxel-based morphometry). Using data from a large sample of stroke patients, we firstly estimated the magnitude of a lesion-deficit mapping of interest and then formally investigated how effect size and its statistical significance varies with sample size. In addition to demonstrating how small samples can result in over- and under-estimations of effect size, we also highlight an issue with large sample sizes whereby high statistical power dramatically increases the likelihood of detecting effects that are so small that they become uninteresting from a scientific viewpoint (i.e. the fallacy of classical inference; Friston et al., 2012). In other words, statistically significant findings when sample sizes are large can hide the fact that the effect under investigation might be of little importance in practical terms, or, even worse, the result of random chance alone and thereby a false positive (Smith and Nichols, 2018).

To investigate the effect of sample size on the results of univariate voxel-based lesion-deficit mapping, we randomly drew thousands of resamples (with a range of sample sizes) from a set of data from 360 stroke survivors who had collectively acquired a wide range of left hemisphere lesions and cognitive impairments. By using a single patient population and holding all procedures and analyses constant, we ensured that variability in the results across thousands of random resamples cannot be explained by methodological confounds - such as the use of dissimilar recruitment strategies and/or behavioural assessments - that are likely to influence the findings of studies that aggregate data from multiple independent sources (e.g., meta-analyses; Müller et al., 2018). Furthermore, by performing our statistical analyses on actual data, rather than running simulations on

synthetically-generated data, we attempt to recreate real-world scenarios that could be encountered by researchers conducting lesion-deficit mapping studies.

The goal of our resampling procedure was to estimate the degree to which the magnitude and statistical significance of the exact same lesion-deficit mapping (i.e. brain areas where damage is associated with difficulties articulating speech) changed with sample size. We report the frequency of significant and non-significant effects (using standard significance thresholds) for 6 different sample sizes: N = 30, 60, 90, 120, 180 and 360. In a real world situation where only one sample is typically analysed, results are far more likely to be published when they reach statistical significance (i.e. the associated p values are below a certain alpha threshold) than when they fail to produce any evidence in favour of the tested hypothesis. This is known as "publication bias" (e.g., Fusar-Poli et al., 2014; Ioannidis et al., 2014; Johnson et al., 2017; Simonsohn et al., 2014a). For example, the prevalence of "positive" (i.e. statistically significant) findings across a wide range of publication outlets, including neuroscience and psychology, has been shown to be well over 80% (Fanelli, 2010, 2012), which suggests that the vast majority of studies that yield "negative" findings are left unpublished. This is known as "the file drawer problem" (Franco et al., 2014; Simonsohn et al., 2014b). Moreover, the number of "positive" results in the fMRI (David et al., 2013) and brain volume abnormalities (Ioannidis, 2011) literature has been demonstrated to be significantly greater than the number expected on the basis of statistical power considerations.

By leaving non-significant results in the file drawer, it becomes increasingly difficult to ascertain which effects are true (and would replicate in subsequent studies) and which are false (and would not replicate in subsequent studies). A highly significant result from a heterogeneous population could, for example, be driven by random noise when a study selects, by chance, a sample that renders an inflated (unstandardized) effect size and under-estimated variance. In line with this rationale, it has been claimed that more than 50% of all significant effects reported in cognitive neuroscience and psychology journals are likely to correspond to false positives (Szucs and Ioannidis, 2017).

Our study therefore speaks directly to the "replication crisis" that is currently being highlighted in psychology and neuroscience (Forstmeier et al., 2017; Gelman

and Geurts, 2017; Ioannidis, 2005; Loken and Gelman, 2017; Munafò et al., 2017; Pashler and Wagenmakers, 2012). In the field of psychology, for example, a large-scale collaborative initiative reported that it could only successfully replicate less than 40% of original effects from a representative set of one hundred randomly selected studies (Open Science Collaboration, 2015). Similar failed replication attempts have also been recorded in other research areas including those investigating structural brain-behaviour correlations (Boekel et al., 2015) and the blood-oxygen-level-dependent response (Chen et al., 2018; Wende et al., 2017).

2. Materials and Methods

2.1. Participants

Data from all participants were retrieved from the Predicting Language Outcome and Recovery After Stroke (PLORAS) database (Price et al., 2010; Seghier et al., 2016). At a minimum, the data available for each patient included: a full assessment of speech and language abilities and a 3D lesion image, in standard space, created from a T1-weighted high resolution (1 mm isotropic voxels) anatomical whole-brain volume, using our automated lesion identification software (Seghier et al., 2008). The study was approved by the Joint Research Ethics Committee of the National Hospital for Neurology and Neurosurgery and the Institute of Neurology. All patients gave written informed consent prior to participation and were compensated for their time.

Our patient selection criteria included all adult stroke survivors who: (i) had a left-hemisphere lesion (as attested by a clinical neurologist: co-author A.P.L.) that was greater than 1 cm^3 (as measured by our automated lesion identification tool; Seghier et al., 2008); (ii) had no history of neurological or psychiatric illness that was not related to their stroke; (iii) were right-handed (pre-morbidly); and, (iv) were native speakers of English. Additionally, individuals who had missing scores on the tasks of interest (see below for details) were excluded from the study. These criteria were met by a total of 363 stroke patients whose data were collected between April 2003 and December 2016. To ensure that our full sample could be divided evenly into smaller resampled data sets (see below for details), we additionally excluded from any further analyses the 3 patients with the smallest lesions (i.e. 1.2, 1.3 and 1.4 cm^3

in size). See Table 1 for demographic and clinical details of the full sample of 360 stroke patients.

2.2. Behavioural assessment

All patients recruited to the PLORAS database are assessed on the Comprehensive Aphasia Test (CAT) (Swinburn et al., 2004). The CAT is a fully standardised test battery, which consists of a total of 27 different tasks. For ease of comparison across tasks, the authors of the CAT encourage the conversion (through a non-linear transformation) of raw scores into T-scores, which represent how well the patient performed relative to a reference population of 113 patients with aphasia, 56 of whom were tested more than once. For example, a T-score of 50 indicates the mean of the patient sample used to standardise the CAT, whereas a T-score of 60 represents one standard deviation above the mean. Most people without post-stroke aphasia would therefore be expected to score above the average of the patient standardisation sample on any given task from the CAT. The threshold for impairment is defined relative to a second reference population of 27 neurologically-normal controls. Specifically, it is the point below which the score would place the patient in the bottom 5% of the control population (Swinburn et al., 2004). Lower scores indicate poorer performance. Importantly, the two standardisation samples referred to before (i.e. 113 patients with aphasia and 27 neurologically-normal controls) are completely independent of the data we report in the current paper (for more details on the standardisation samples, see Swinburn et al., 2004).

As stated in the CAT manual (p. 71), the main advantages of converting raw scores into T-scores is that this allows: (i) scores from different tasks to be compared because they have been put on a common scale; and (ii) the use of parametric statistics given that T-scores are normally distributed scores with a mean of 50 and a standard deviation of 10.

The current study focused exclusively on a total of 5 tasks from the CAT. Task 1 used nonword repetition to assess the patient's ability to articulate speech. Task 2 used written picture naming to test the patient's ability to find the names of objects (lexical/phonological retrieval). Tasks 3-5 tested the patient's ability to recognise, process and remember the semantic content of pictures and auditory words. Task details were as follows:

Task 1: The CAT nonword repetition (Rep-N) task aurally presents five nonsense words (e.g., gart), one at a time, with instructions to repeat them aloud. Immediate correct responses were given a score of 2; incorrect responses were given a score of 0; correct responses after a self-correction or a delay (> 5 seconds) were given a score of 1. Articulatory errors (e.g., dysarthric distortions) not affecting the perceptual identity of the target were scored as correct. Verbal, phonemic, neologistic and apraxic errors were scored as incorrect. T-scores equal to or below 51 constitute the impaired range.

Task 2: The CAT written picture naming (Writt-PN) task visually presents five pictures of objects (e.g., tank), one at a time, with instructions to write their names down. Letters in the correct position were given a score of 1 each. Substitutions, omissions and transpositions were given a score of 0. One point was deducted from the total score if one or more letters were added to the target word. T-scores equal to or below 54 constitute the impaired range.

Task 3: The CAT semantic associations (Sem-A) task visually presents five pictures of objects simultaneously. The instructions were to match the picture at the centre (e.g., mitten) with one of four possible alternatives according to the strongest semantic association (e.g., hand, sock, jersey, and lighthouse). The inclusion of a semantically related distractor (e.g., sock) encouraged deeper levels of semantic processing/control. There are a total of ten test trials plus a practice one at the beginning. Correct responses were given a score of 1; incorrect responses were given a score of 0. T-scores equal to or below 47 constitute the impaired range.

Task 4: The CAT recognition memory (Recog-M) task visually presents each of the ten central items from the CAT semantic associations task (one at a time) along with three unrelated distractors. The instructions were to indicate which of the four pictures on display had been seen before. There are a total of ten test trials plus a practice one at the beginning. The scoring system for this task was identical to that used in the semantic associations task. T-scores equal to or below 43 constitute the impaired range.

Task 5: The CAT auditory word-to-picture matching (A_W-P) task involves hearing a word produced by the examiner and selecting the picture among four possible alternatives that best matches the meaning of the heard word. There are a total of

fifteen test trials plus a practice one at the beginning. Immediate correct responses were given a score of 2; incorrect responses were given a score of 0; correct responses after a self-correction or a delay (> 5 seconds) were given a score of 1. T-scores equal to or below 51 constitute the impaired range.

2.3. MRI data acquisition, pre-processing and lesion identification

T1-weighted high resolution anatomical whole-brain volumes were available for all patients (*n* = 360). Four different MRI scanners (Siemens Healthcare, Erlangen, Germany) were used to acquire the structural images: 167 patients were imaged on a 3T Trio scanner, 131 on a 1.5T Sonata scanner, 57 on a 1.5T Avanto scanner, and five on a 3T Allegra scanner. For anatomical images acquired on the 1.5T Avanto scanner, a 3D magnetization-prepared rapid acquisition gradient-echo (MPRAGE) sequence was used to acquire 176 sagittal slices with a matrix size of 256 × 224, yielding a final spatial resolution of 1 mm isotropic voxels (repetition time/echo time/inversion time = 2730/3.57/1000 ms). For anatomical images acquired on the other three scanners, an optimised 3D modified driven equilibrium Fourier transform (MDEFT) sequence was used to acquire 176 sagittal slices with a matrix size of 256 × 224, yielding a final spatial resolution of 1 mm isotropic voxels: repetition time/echo time/inversion time = 12.24/3.56/530 ms and 7.92/2.48/910 ms at 1.5T and 3T, respectively (Deichmann et al., 2004).

The T1-weighted anatomical whole-brain volume of each patient was subsequently analysed with our automated lesion identification toolbox using default parameters (for more details, see Seghier et al., 2008). This converts a scanner-sensitive raw image into a quantitative assessment of structural abnormality that should be independent of the scanner used. The procedure combines a modified segmentation-normalisation routine with an outlier detection algorithm according to the fuzzy logic clustering principle (for more details, see Seghier et al., 2007). The outlier detection algorithm assumes that a lesioned brain is an outlier in relation to normal (control) brains. The output includes two 3D lesion images in standard MNI space, generated at a spatial resolution of 2 x 2 x 2 mm^3. The first is a fuzzy lesion image that encodes the degree of structural abnormality on a continuous scale from 0 (completely normal) to 1 (completely abnormal) at each given voxel relative to normative data drawn from a sample of 64 neurologically-normal controls. A voxel

with a high degree of abnormality (i.e. a value near to 1 in the fuzzy lesion image) therefore means that its intensity in the segmented grey and white matter deviated markedly from the normal range. The second is a binary lesion image, which is simply a thresholded (i.e. lesion/no lesion) version of the fuzzy lesion image. All our statistical analyses were based on the fuzzy images. The binary images were used to delineate the lesions, to estimate lesion size and to create lesion overlap maps.

2.4. Lesion-deficit analyses

We used voxel-based morphometry (Ashburner and Friston, 2000; Mechelli et al., 2005) to assess lesion-deficit relationships (Mummery et al., 2000; Tyler et al., 2005), performed in SPM12 using the general linear model. The imaging data entered into the voxel-based analysis were the fuzzy (continuous) lesion images that are produced by our automated lesion identification toolbox.

The most important advantage of utilising the fuzzy lesion images (as in Price et al., 2010) over alternative methods is that they provide a quantitative measure of the degree of structural abnormality, at each and every voxel of the brain, relative to neurologically-normal controls. In contrast to fuzzy lesion images, (i) binary lesion images do not provide a continuous measure of structural abnormality and will be less sensitive to subtle changes that are below an arbitrary threshold for damage (e.g., Fridriksson et al., 2013; Gajardo-Vidal et al., 2018); (ii) normalised T1 images do not distinguish between typical and atypical (abnormal) variability in brain structure (e.g., Stamatakis and Tyler, 2005); and (iii) segmented grey or white matter probability images when used in isolation (as in standard VBM routines) do not provide a complete account of the whole of the lesion (e.g., Mehta et al. 2003).

In Analysis 1, the fuzzy lesion images were entered into a voxel-based multiple regression model with 6 different regressors (5 behavioural scores and lesion size); see Fig. 1. The regressor of interest was nonword repetition scores that are sensitive to difficulties articulating speech. In addition, the following regressors were included to factor out other sources of variance: written picture naming scores (which are sensitive to name retrieval abilities), semantic associations scores (which are sensitive to visual recognition and semantic processing), auditory word-to-picture matching scores (which are sensitive to auditory recognition and lexical-semantic processing), recognition memory scores (which are sensitive to picture recognition

and memory) and lesion size (to partial out linear effects of lesion size). For the voxel-based lesion-deficit analysis (with 360 patients), the search volume was restricted to voxels that were damaged in at least five patients (as in Fridriksson et al., 2016; for rationale, see Sperber and Karnath, 2017). For this purpose, a lesion overlap map based on the binary lesion images from all 360 patients was created, thresholded at five, and used as an inclusive mask before estimating the model (see Fig. 2A). Our statistical voxel-level threshold was set at $p < 0.05$ after family-wise error (FWE) correction for multiple comparisons (using random field theory as implemented in SPM; Flandin and Friston, 2015) across the whole search volume (for alternative approaches, see Mirman et al., 2018).

Having identified a significant lesion-deficit mapping, we quantified the strength of the association between lesion and deficit by: (i) extracting the raw signal (which indexes the degree of structural abnormality) from each statistically significant voxel; (ii) averaging the signal across voxels (i.e. a single value per patient); and, finally, (iii) computing the partial correlation between lesion load in the region of interest and nonword repetition scores, after adjusting for the effect of the covariates of no interest (i.e. 4 behavioural scores and lesion size). Our measure of effect size was the proportion of variance (= R^2) in nonword repetition scores explained uniquely by lesion load in the region of interest (i.e. the best estimate of the true population effect that we have).

In Analysis 2, we investigated how sample size affected the reproducibility of the lesion-deficit mapping within the region of interest identified in Analysis 1. Specifically, we generated 6000 bootstrap samples of the following sizes: 360, 180, 120, 90, 60 and 30 (i.e. 36000 resamples in total). These sample sizes were selected to follow as closely as possible those observed in the vast majority of published voxel-based lesion-deficit mapping studies (e.g., Dressing et al., 2018; Fridriksson et al., 2013, 2016; Halai at el., 2017; Schwartz et al., 2011, 2012). For each iteration of the resampling procedure, individuals were drawn randomly from the full set of 360 patients with replacement, meaning that the probability of being chosen remained constant throughout the selection process (i.e. the procedure satisfied the Markovian, memory-less, property). For each bootstrap sample, the partial correlation between nonword repetition scores and lesion load (averaged across voxels in the region of interest from Analysis 1) was computed. The resulting

R^2 and p values were recorded, after regressing out the variance accounted for by the covariates of no interest. Of note, when we re-ran the resampling procedure outlined above with the replacement feature disabled (i.e. sampling without replacement), virtually the same results were obtained (for more details, see Supplementary Material).

In addition, to rule out the possibility that variability in the results could simply be explained by differences in the distribution of damage across the brain, we quantified statistical power in the region of interest from Analysis 1 for a representative subset of bootstrap samples. Specifically, only those resamples that produced an R^2 value which fell exactly at a particular decile (i.e. 0th, 10th, 20th…100th) of the distribution of effect sizes were considered. This resulted in the selection of a total of 66 bootstrap samples (i.e. 11 for each sample size); see Table 2. Critically, our power calculations show where in the brain there was sufficient statistical power to detect a significant lesion-deficit association at a threshold of $p < 0.05$ after correction for multiple comparisons. The statistical power maps were generated using the "nii_powermap" function of NiiStat (https://www.nitrc.org/projects/niistat/), which is a set of Matlab scripts for analysing neuroimaging data from clinical populations.

Importantly, we have chosen to assess in-sample effect sizes, i.e. without validating in a separate data set (Friston, 2012). In this context, the effect size is providing an estimate of the strength of the particular effect identified by our analysis in our data. It may be that an out-of-sample prediction - on new data - would indicate a smaller effect size. However, this would not invalidate the logic of our reasoning, particularly since the essential point we are making here is that our effect size estimate (i.e. approximately 11% in R^2 terms) is very small. If there is inflation in this estimate, it could only mean that the out-of-sample effect size would be even less. Therefore, we have been able to show that even for an over-estimated effect size (if it would turn out to be), there are serious problems that arise from small sample sizes, the fallacy of classical inference, and publication bias. The impact of these issues on the reliability of the findings would only be worse if the effect size were to come down.

Furthermore, we have first statistically selected an ROI in a large sample of patients, with a "left-hemisphere" analysis, and then used smaller and smaller bootstrap samples that focused on the identified ROI. In this sense, we are

performing (non-orthogonal) statistical tests in a previously selected ROI, which could potentially inflate false positive rates (Brooks et al., 2017). Consequently, the results derived from the analysis of smaller samples should not be taken as robust findings: they are being presented to make important methodological points. Our best statistical estimates of the effect considered are those obtained from the full data set.

3. Results

3.1. Analysis 1: identifying a region of interest

Poorer speech articulation was significantly associated with greater lesion load (after controlling for written picture naming, recognition memory, semantic associations and auditory word-to-picture matching scores in addition to lesion size) in 549 voxels (= 4.4 cm^3 in size; see Table 3). These voxels became our region of interest (ROI) for all subsequent analyses. They were located in parts of the left ventral primary motor and somatosensory cortices (i.e. tongue, larynx, head and face regions), anterior supramarginal gyrus, posterior insula and surrounding white matter (see Fig. 2B).

This highly significant lesion-deficit relationship accounted for 11% of the variance (95% credible interval calculated using a flat prior: 0.06-0.18; Morey et al., 2016); see Fig. 3. In the following analyses, we ask how sample size affects the reproducibility of the identified effect.

3.2. Analysis 2: effect size variability and replicability

Although the mean/median effect sizes were similar across sample sizes, the mean/median *p* values changed considerably with sample size (see Fig. 4), because there was wide sample-to-sample variability in the extent to which the original effect was replicated. For instance, less than 40% of the random resamples where $N = 30$ generated significant *p* values, while this raised to virtually 100% for the resampled data sets where $N \geq 180$. Overall, R^2 values ranged between 0.00 and 0.79, whereas *p* values ranged between $6*10^{-27}$ and 1 (see Fig. 5A and B). Additionally, our analyses showed that, as sample size increased, R^2 values tended to fall closer to the mean of the effect size distribution, although a not inconsiderable degree of uncertainty regarding R^2 estimation remained (even for $N = 180$ and 360). In other

words, the dispersion of the R^2 values tended to be larger with smaller sample sizes (see Fig. 5A), resulting in less precision in the estimation of the magnitude of the true population effect.

3.2.1. Low-powered resamples can inflate effect sizes

Since studies that obtain statistically non-significant results (i.e. typically $p \geq 0.05$) are hardly ever published (also known as the file drawer problem or study publication bias), we focused directly upon the resampled data sets that produced significant p values. For $N = 30$, the mean and median effect sizes of these significant resamples (i.e. roughly 37%) were 0.26 and 0.24 (range = 0.16-0.79). Conversely, the mean and median effect sizes for the $N = 30$ resamples where the lesion-deficit mapping did not reach statistical significance (roughly 63%) were 0.07 and 0.06 (range = 0.00-0.16); see Table 4 for similar findings when $N = 60$. Critically, using a more stringent statistical threshold would only aggravate the problem (for more details, see Table 4). With larger sample sizes ($N \geq 90$), however, effect size inflation is counteracted since both over- and under-estimations of the true effect size surpassed the threshold for statistical significance, resulting in relatively accurate mean estimates (0.13, 0.12, 0.12, and 0.11 respectively).

3.2.2. High-powered resamples are sensitive to trivial/small effects

The frequency with which a significant association was observed between lesion load in the ROI and nonword repetition scores increased dramatically with sample size. For example, whereas roughly 37% of the effects for $N = 30$ would be typically regarded as statistically significant (i.e. $p < 0.05$), more than 85% of the lesion-deficit mappings for $N \geq 90$ generated equally low or even lower p values (see Table 4). More importantly, effects as small as 0.05 in R^2 terms (i.e. that only accounted for 5% of the variance) reached statistical significance for $N = 90$; and this phenomenon was even more pronounced in the presence of larger sample sizes: 0.02 for $N = 180$ (see Table 4 and Fig. 5A). Reporting point and interval estimates of effect sizes is therefore essential for assessing the importance or triviality of the identified lesion-deficit mapping, which is particularly relevant when the study uses large sample sizes.

4. Discussion

The goal of this study was to examine how sample size influences the reproducibility of voxel-based lesion-deficit mappings. First, we identified a significant lesion-deficit association and estimated its magnitude using data from a very large sample of 360 patients who were all right-handed, English speaking stroke survivors with unilateral left hemisphere damage. By repeating the same analysis on thousands of bootstrap samples of different sizes we illustrate how the estimated effect size, and its statistical significance, varied across replications. This allowed us to index the degree of uncertainty in the estimation of the true population effect as a function of sample size. As expected, effect sizes were more likely to be over-estimated or under-estimated with small sample sizes (i.e. variability in the results increased as sample size decreased). Conversely, we demonstrate how highly significant lesion-deficit mappings can be driven by a negligible proportion of the variance when the sample size is very large.

4.1. Estimating the true effect size

The first part of our investigation identified a region of interest (ROI) where damage was reliably associated with impairments in speech articulation. We then calculated what proportion of the variance in nonword repetition scores could be accounted for by the degree of damage to the identified region after factoring out confounds from auditory and visual perception, speech recognition, lexical/semantic processing and word retrieval abilities. The ROI included anatomical brain structures that have been associated with speech production in many previous lesion studies. These include the insula (Ogar et al., 2006), the precentral gyrus, the postcentral gyrus, the supramarginal gyrus and surrounding white matter (Baldo et al., 2011; Basilakos et al., 2015). It did not involve the inferior frontal gyrus/frontal operculum as reported in Hillis et al. (2004) and Baldo et al. (2011), even though our full sample incorporated plenty of patients with damage to these regions (see Fig. 2A). We do not attempt here to adjudicate whether this discrepancy was a consequence of a false negative in our study or a false positive in prior studies. Our focus was on how well the identified lesion-deficit mapping could be replicated across thousands of bootstrap samples drawn randomly from the original data set of 360 patients. For each resample, we estimated how much of the variance in nonword repetition scores could be accounted for by lesion load in the ROI (after adjusting for the effect of the covariates of no interest). These effect sizes and their statistical significance were

then compared to our best estimate of the "true" population effect size, which was found (from our full sample of 360 patients) to be 11%.

4.2. Variability in the estimated effect size and its statistical significance

The second part of our investigation showed that the probability of finding a significant lesion-deficit association in the ROI from the first analysis (with 360 participants), depended on the size of the sample. For larger samples ($N \geq 180$), the effect of interest was detected in virtually 100% of resamples. Whereas for smaller samples ($N = 30$), it was detected in less than 40% of resamples (see Table 4). We can also show that *p* values decrease as *N* increases, even when effect sizes are equated (see Fig. 4 and 50th percentile in Table 2). This observation is in line with prior reports that *p* values exhibit wide sample-to-sample variability (Cumming, 2008; Halsey et al., 2015; Vsevolozhskaya et al., 2017), particularly in the presence of small sample sizes (Hentschke and Stüttgen, 2011).

When considering the central tendency of effect size estimates, the difference between larger and smaller resamples is dramatically reduced compared to that seen for *p* values (see mean/median effect sizes in Fig. 4). Nevertheless, even if *p* values were completely abandoned (e.g., Trafimow and Marks, 2015), there is still a great deal of uncertainty in the accuracy with which effect sizes can be estimated when small samples are used. This highlights the importance of reaching a better balance between null-hypothesis significance testing and effect size estimation (Chen et al., 2017; Cumming, 2014; Morey et al., 2014). Indeed, *p* values only indicate the likelihood of observing an effect of a given magnitude (when the null hypothesis is true). As such, they cannot convey the same information provided by point and interval estimates of effect sizes (Steward, 2016; Wasserstein and Lazar, 2016), particularly since the relationship between *p* values and effect sizes is non-linear (Hentschke and Stüttgen, 2011; Simonsohn et al., 2014a, 2014b).

There are several potential reasons why the magnitude and statistical significance of the same effect varies so markedly across resamples. For example, high sample-to-sample variability could reflect (i) sampling error due to heterogeneity in the lesion-deficit association across participants (Button, 2016; Stanley and Spence, 2014), (ii) outliers that are confounding the effects (Rousselet and Pernet, 2012) or (iii) measurement error (Button, 2016; Loken and Gelman, 2017; Stanley

and Spence, 2014). In this context, the field needs to adopt informed sampling strategies that ensure representative samples and maximise the probability of identifying generalizable lesion-deficit mappings (Falk et al., 2013; LeWinn et al., 2017; Paus, 2010).

4.3. Unreliable effect sizes in smaller samples

High variance in the results of our lesion-deficit mappings with smaller samples (N = 30 and 60) demonstrates how effects can be over- as well as under-estimated (e.g., Cremers et al., 2017; Ioannidis, 2008). Indeed, we show that 85% of all significant random data sets for N = 30 yielded effect size estimates that were larger than the upper bound of the credible interval (see Table 5). This is consistent with prior observations that low-powered studies (with small sample sizes) can only consistently detect large deviations from the true population effect (Szucs and Ioannidis, 2017). Put another way, even when effect sizes are accurately estimated from small samples, they are unlikely to attain statistical significance; particularly when the magnitude of the effect under investigation is small or medium. In our data, for example, we found that more than half the analyses with N = 30 that did not reach statistical significance produced effect sizes that fell within the credible interval (i.e. accurate estimations of effect sizes resulted in false negatives). Even worse, analyses of small sample sizes can invert the direction of the effect (Gelman and Carlin, 2014) as seen in our data where we found that 5% of all results for N = 30 were in the wrong direction. Furthermore, reporting such findings as if they were accurate representations of reality would lead to misleading conclusions (Nissen et al., 2016).

Critically, the problem was not solved but became worse when we adopted a more stringent statistical threshold, which is contrary to that proposed by Johnson (2013) and Benjamin et al. (2018). For example, if we were to raise the statistical threshold from $p < 0.05$ to $p < 0.001$ for the N = 30 resamples, the statistically significant effect sizes would range from 38% to 79% of the variance (compared to 11% in the full sample of 360 patients). Increasing sample size, however, does improve accuracy, with less than 10% of significant p values associated with inflated effect sizes when $N \geq 180$ (see Table 5).

Given that results are more likely to be published if they reach statistical significance than if they do not (i.e. the file drawer problem or study publication bias), our findings highlight three important implications for future lesion-deficit mapping studies. First, low-powered studies (due to small sample sizes) could lead a whole research field to over-estimate the magnitude of the true population effect. Second, power calculations based on inflated effect sizes from studies with small samples will inevitably over-estimate the statistical power associated with small sample sizes (Anderson et al., 2017). Third, although the mean effect size measured over many studies with small sample sizes will eventually converge on the true effect size, in reality, the same study is seldom replicated exactly and null results are only rarely reported. It has therefore been advocated that, contrary to current practices, it is better to carry out a few well-designed high-powered studies than it is to assimilate the results from multiple low-powered studies (Bakker et al., 2012; Higginson and Munafò, 2016). In brief, large scale studies increase the probability that an identified lesion-deficit mapping is correct (Button et al., 2013a; Szucs and Ioannidis, 2017).

4.4. Trivial effect sizes in larger samples

Another important observation from the current study is that, when samples are sufficiently large, relatively weak lesion-deficit associations can be deemed statistically significant (i.e. $p < 0.05$). For instance, effects that only accounted for as little as 3% of the variance reached statistical significance when $N \geq 120$ - an inferential problem known as the fallacy of classical inference (Friston, 2012; Smith and Nichols, 2018). However, our findings are consistent with the view that this issue can be addressed by reporting point and interval estimates of effect sizes (Button et al., 2013b; Lindquist et al., 2013), which allow one to assess the practical significance (as opposed to statistical significance only) of the results. In other words, it can be argued that the fallacy of classical inference is specific to statistical tests (e.g., *t, F* and/or *p* values), leaving effect sizes largely unaffected (Reddan et al., 2017). Furthermore, there are two important advantages of conducting high-powered studies: (i) they greatly attenuate the impact of study publication bias as both over- and under-estimations of the true effect size will surpass the threshold for statistical significance; and (ii) the precision with which the magnitude of the true population effect can be estimated is substantially improved (Lakens and Evers, 2014; see Table 5 and Figs. 4 and 5A). Our study also indicates that, even with

sample sizes as large as N = 360, a not inconsiderable degree of uncertainty in R^2 estimation remained, which suggests that increasing sample size beyond this N will continue to bring benefit.

4.5. Study limitations

The focus of the current paper has been on establishing the degree to which the replicability of lesion-deficit mappings is influenced by sample size. To illustrate our points, we have (i) searched for brain regions where damage is significantly related to impairments in articulating speech; (ii) estimated the strength of the identified lesion-deficit association; and, (iii) run the exact same analysis on thousands of samples of varying size. However, we have not attempted to account for all possible sources of inconsistencies in univariate voxel-based lesion-deficit mapping. Nor have we investigated how our results would change if we selected another function of interest (e.g., word retrieval or phonological processing). Indeed, it has already been pointed out that higher-order functions might be associated with smaller effects than lower-level ones (Poldrack et al., 2017; Yarkoni, 2009).

We also acknowledge that there are many different ways of conducting voxel-based lesion-deficit analyses (for more information see de Haan and Karnath, 2018; Karnath et al., 2018; Rorden et al., 2007; Sperber and Karnath, 2018). We have selected one approach, using mass-univariate multiple regression on continuous measures of structural abnormality, behaviour and lesion size. However, we could have used other types of images or other behavioural regressors. For example, several recent studies have adopted dimensionality reduction techniques, such as principal component analysis (PCA), to transform a group of correlated behavioural measures into a smaller number of orthogonal (uncorrelated) factors (e.g., Butler et al., 2014; Corbetta et al., 2015; Mirman et al., 2015a). This PCA approach has made an important contribution to finding coarse-grained explanatory variables (e.g., Halai et al., 2017; Lacey et al., 2017; Mirman et al., 2015b; Ramsey et al., 2017), but some of its limitations are that it: (i) involves an arbitrary criterion for factor extraction; (ii) ignores unexplained variance when selecting a limited number of components; and, (iii) necessitates subjective, a posteriori, interpretation as to what the components might mean based on the factor loadings, which is not typically clear cut. Instead, we propose that a better solution for tackling orthogonality issues is to adopt both a

rigorous sampling strategy as well as behavioural measures that offer an optimal sensitivity-specificity balance.

Finally, we have highlighted that the reliance on small-sized samples of patients in the presence of publication bias can undermine the inferential power of univariate voxel-based lesion-deficit analyses. However, we have not attempted to provide guidance on how prospective power calculations - that correct for the various forms of bias present in scientific publications - can be conducted. Nor have we illustrated how the presence of publication and other reporting biases in the lesion-deficit mapping literature, specifically, can be ascertained. The reason simply being that others have already devoted considerable effort to developing tools that identify and deal with problems such as: (i) the excess of statistically significant findings (e.g., Ioannidis and Trikalinos, 2007); (ii) the proportion of false positives (e.g., Gronau et al., 2017); (iii) the presence of publication bias and questionable research practices (e.g., Du et al., 2017; Simonsohn et al., 2014a, 2014b); (iv) errors in the estimation of the direction and/or magnitude of a given effect (e.g., Gelman and Carlin, 2014); and, (v) sample size calculations that take into account the impact of publication bias and uncertainty on the estimation of reported effect sizes (e.g., Anderson et al., 2017). With respect to statistical power, the situation is further complicated by the fact that - in the context of univariate voxel-based lesion-deficit mapping - it not only depends on the size of the sample, the magnitude of the effect under study and the statistical threshold used (Cremers et al., 2017), but also on the distribution of damage across the brain (which is non-uniform; Inoue et al., 2014; Kimberg et al., 2007; Mah et al., 2014; Sperber and Karnath, 2017). More research on the topic will be required before prospective power calculations can be fully trusted. Until that moment, the recruitment of representative patient samples in combination with high-powered designs seems to be the best available solution to the issues discussed here.

4.6. Interpreting voxel-based lesion-deficit mappings

The strength of the lesion-deficit association that we identified in a large sample of 360 patients illustrates that the majority of the variability in speech articulation abilities was driven by factors other than the degree of damage to the ROI. A clear implication of this is that the field of lesion-deficit mapping still has a

long way to go before it can inform current clinical practice, which is arguably one of its most important goals. Future studies will need to control and understand other known sources of variance (apart from lesion site and size) such as time post-stroke, age and education in order to improve our ability to predict language outcome and recovery after stroke at the individual patient level (Price et al., 2017). Furthermore, to map all the possible ways in which brain damage can affect behaviour, it will in all likelihood be necessary to use increasingly larger samples of patients (e.g., Price et al., 2010; Seghier et al., 2016) and multivariate methods (e.g., Hope et al., 2015; Pustina et al., 2018; Yourganov et al., 2016; Zhang et al., 2014).

5. Conclusions

This study investigated the impact of sample size on the reproducibility of voxel-based lesion-deficit mappings. We showed that: (i) highly significant lesion-deficit associations can be driven by a relatively small proportion of the variance; (ii) the exact same lesion-deficit mapping can vary widely from sample to sample, even when analyses and behavioural assessments are held constant; (iii) the combination of publication bias and low statistical power can severely affect the reliability of voxel-based lesion-deficit mappings; and, finally, (iv) reporting effect size estimates is essential for assessing the importance or triviality of statistically significant findings. Solutions to the issues highlighted here will, in our view, likely involve the use of: (a) improved reporting standards; (b) increasingly larger samples of patients; (c) multivariate methods; (d) informed sampling strategies; and, (e) independent replications. Careful reflection on some deeply-rooted research practices, such as biases in favour of statistically significant findings and against null results, might also be necessary.

Acknowledgements

This work was supported by the Wellcome Trust [097720/Z/11/Z and 205103/Z/16/Z, C.J.P.], the Medical Research Council [MR/M023672/1, C.J.P., T.M.H.H., M.L.S.] and the Stroke Association [TSA 2014/02, C.J.P.]. D.L.L-P. (CONICYT BECAS-CHILE 72140131) and A.G-V. (CONICYT BECAS-CHILE 73101009) were funded by the Chilean national commission for scientific and technological research (CONICYT) through its scholarship scheme for graduate studies abroad. We would like to thank the PLORAS recruitment team (http://www.ucl.ac.uk/ploras/) for their help with data collection.

Table 1: Summary of demographic and clinical data for full sample.

Factor		***N* = 360**
Age at stroke onset (years)	*M*	54.4
	SD	12.9
	Range	17.2-86.5
Age at testing (years)	*M*	59.4
	SD	12.4
	Range	21.3-90.0
Time post-stroke (years)	*M*	4.9
	SD	5.2
	Range	0.2-36.0
Education (years)*****	*M*	14.5
	SD	3.2
	Range	10.0-30.0
Lesion size (cm^3)	*M*	85.7
	SD	87.6
	Range	1.5-386.2
Gender	Males	250
	Females	110
Rep-N	Imp/Non	132/228
	M	54.4
	SD	9.1
Writt-PN	Imp/Non	105/255
	M	58.6
	SD	8.7
Recog-M	Imp/Non	37/323
	M	53.9
	SD	7.0
Sem-A	Imp/Non	36/324
	M	56.6
	SD	6.1
A_W-P	Imp/Non	77/283
	M	57.0
	SD	6.8

Imp/Non = number of patients with impaired/non-impaired performance. *****Missing data: three patients.

Table 2: Statistical power in the region of interest.

%tile		Sample Size					
		30	**60**	**90**	**120**	**180**	**360**
0th	Power	98%	100%	100%	100%	100%	100%
	R^2	0.00	0.00	0.00	0.00	0.00	0.01
	P	0.999	0.999	0.999	0.999	0.404	0.093
10th	Power	99%	100%	100%	100%	100%	100%
	R^2	0.01	0.03	0.04	0.05	0.06	0.07
	P	0.638	0.218	0.064	0.015	0.001	0.000
20th	Power	63%	100%	100%	100%	100%	100%
	R^2	0.03	0.05	0.06	0.07	0.08	0.09
	P	0.400	0.093	0.022	0.004	0.000	0.000
30th	Power	86%	100%	100%	100%	100%	100%
	R^2	0.06	0.07	0.08	0.08	0.09	0.10
	P	0.250	0.046	0.009	0.002	0.000	0.000
40th	Power	92%	100%	100%	100%	100%	100%
	R^2	0.08	0.09	0.10	0.10	0.10	0.11
	P	0.158	0.025	0.004	0.001	0.000	0.000
50th	Power	98%	100%	100%	100%	100%	100%
	R^2	0.11	0.11	0.11	0.11	0.11	0.11
	P	0.099	0.012	0.002	0.000	0.000	0.000
60th	Power	100%	100%	100%	100%	100%	100%
	R^2	0.15	0.14	0.13	0.13	0.13	0.12
	P	0.060	0.006	0.001	0.000	0.000	0.000
70th	Power	83%	100%	100%	100%	100%	100%
	R^2	0.18	0.16	0.15	0.14	0.14	0.13
	P	0.032	0.002	0.000	0.000	0.000	0.000
80th	Power	96%	100%	100%	100%	100%	100%
	R^2	0.23	0.19	0.17	0.16	0.15	0.14
	P	0.015	0.001	0.000	0.000	0.000	0.000
90th	Power	100%	100%	100%	100%	100%	100%
	R^2	0.30	0.23	0.21	0.19	0.18	0.16
	P	0.004	0.000	0.000	0.000	0.000	0.000
100th	Power	99%	100%	100%	100%	100%	100%
	R^2	0.79	0.52	0.39	0.39	0.38	0.28
	P	0.000	0.000	0.000	0.000	0.000	0.000

The table shows that in all but one case, more than 80% of the voxels comprising the region of interest from Analysis 1 had sufficient statistical power to detect a significant lesion-deficit association at a threshold of $p < 0.05$ after correction for multiple comparisons. %tile = percentile of the effect size (R^2) distribution; Power =

percentage of voxels within the region of interest from Analysis 1 that had sufficient statistical power to detect a significant lesion-deficit association at a statistical threshold of $p < 0.05$ after correction for multiple comparisons; $R^2 = R^2$ value (at a particular decile); $P = p$ value (at a particular decile).

Table 3: Brain regions where lesion load is associated with speech articulation abilities.

Brain region	Peak coordinates			Voxel-level		Cluster-level	
	x	y	z	*Z*-score	$P_{FWE\text{-}corr}$	Extent	$P_{FWE\text{-}corr}$
Post-Central	-60	-16	12	5.8	0.000	549*	< 0.001
	-52	-14	24	4.7	0.009		
	-56	-12	18	4.6	0.012		
Posterior Insula	-40	-16	8	5.3	0.001		
Anterior SMG	-66	-30	20	4.7	0.008		
WM	-48	-24	26	4.6	0.010		

The table shows representative (peak) voxels where a significant association between stroke damage and difficulties articulating speech was found. All were in the left hemisphere and the coordinates are reported in MNI space. SMG = supramarginal gyrus; WM = white matter; $P_{FWE\text{-}corr}$ = *p* value corrected (family-wise error correction) for multiple comparisons. *At a cluster-forming voxel-wise threshold of $p < 0.05$ FWE-corrected.

Table 4: Mean and median effect size of the significant and non-significant random data sets by sample size.

R^2	Sample Size											
	30		**60**		**90**		**120**		**180**		**360**	
	s	ns	s	ns	s	ns	s	ns	s	ns	s	ns
Count	2214	3786	4272	1728	5289	711	5747	253	5974	26	5999	1
	258	5742	1279	4721	2613	3387	3911	2089	5369	631	5997	3
M	0.26	0.07	0.16	0.04	0.13	0.03	0.12	0.02	0.12	0.01	0.11	---
	0.45	0.12	0.24	0.09	0.18	0.07	0.15	0.06	0.12	0.05	0.11	0.02
Mdn	0.24	0.06	0.15	0.04	0.12	0.03	0.11	0.02	0.11	0.01	0.11	---
	0.43	0.11	0.23	0.09	0.17	0.08	0.14	0.06	0.12	0.05	0.11	0.03
Min	0.16	0.00	0.07	0.00	0.05	0.00	0.03	0.00	0.02	0.00	0.03	0.01
	0.38	0.00	0.19	0.00	0.12	0.00	0.09	0.00	0.06	0.00	0.03	0.01
Max	0.79	0.16	0.52	0.07	0.39	0.05	0.39	0.03	0.38	0.02	0.28	0.01
	0.79	0.38	0.52	0.19	0.39	0.12	0.39	0.09	0.38	0.06	0.28	0.03

For each summary statistic, the upper row indicates the corresponding value when the alpha threshold was set at 0.05, whereas the lower row indicates the corresponding value when the alpha threshold was set at 0.001. Count = the number of resampled data sets that generated significant or non-significant R^2 values; s = significant (i.e. $p < \alpha$); ns = not significant (i.e. $p \geq \alpha$); *M* = mean R^2 value; *Mdn* = median R^2 value; Min = minimum R^2 value; Max = maximum R^2 value.

Table 5: Frequency of accurate and inaccurate effect size estimates by sample size and statistical significance.

N	Effect Size					
	Significant			Not significant		
	> 95% CI	= 95% CI	< 95% CI	> 95% CI	= 95% CI	< 95% CI
360	173	5686	140	0	0	1
180	556	4925	493	0	0	26
120	795	4430	522	0	0	253
90	1081	3887	321	0	0	711
60	1417	2855	0	0	421	1307
30	1873	341	0	0	2007	1779

The table shows, for each sample size, the frequency with which effect size estimates reached statistical significance (i.e. $p < 0.05$) and fell within (=) or outside the 95% credible interval (i.e. 0.06-0.18) of the best estimate of the "true" population effect (i.e. $R^2 = 0.11$). 95% CI = 95% credible interval; > = larger than the upper bound of 95% CI; < = smaller than the lower bound of 95% CI.

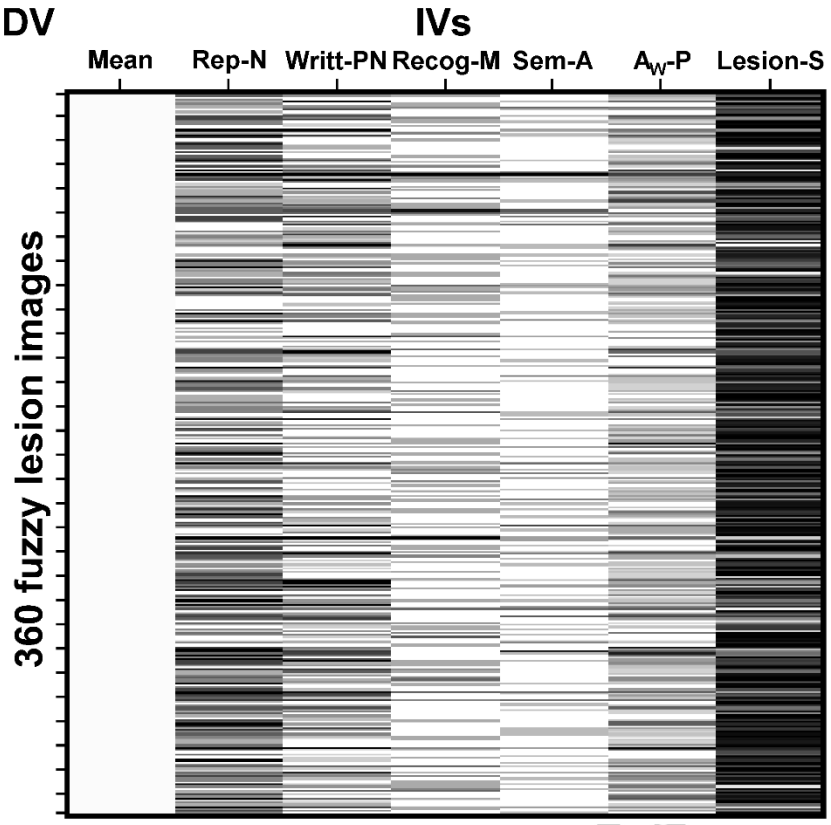


Fig. 1. Design matrix. The design matrix for Analysis 1 is shown, where the columns represent the subject-specific independent variables (IVs), with one value for each subject, and the rows correspond to the dependent variable (DV) indexing the degree of structural abnormality in the fuzzy lesion images.

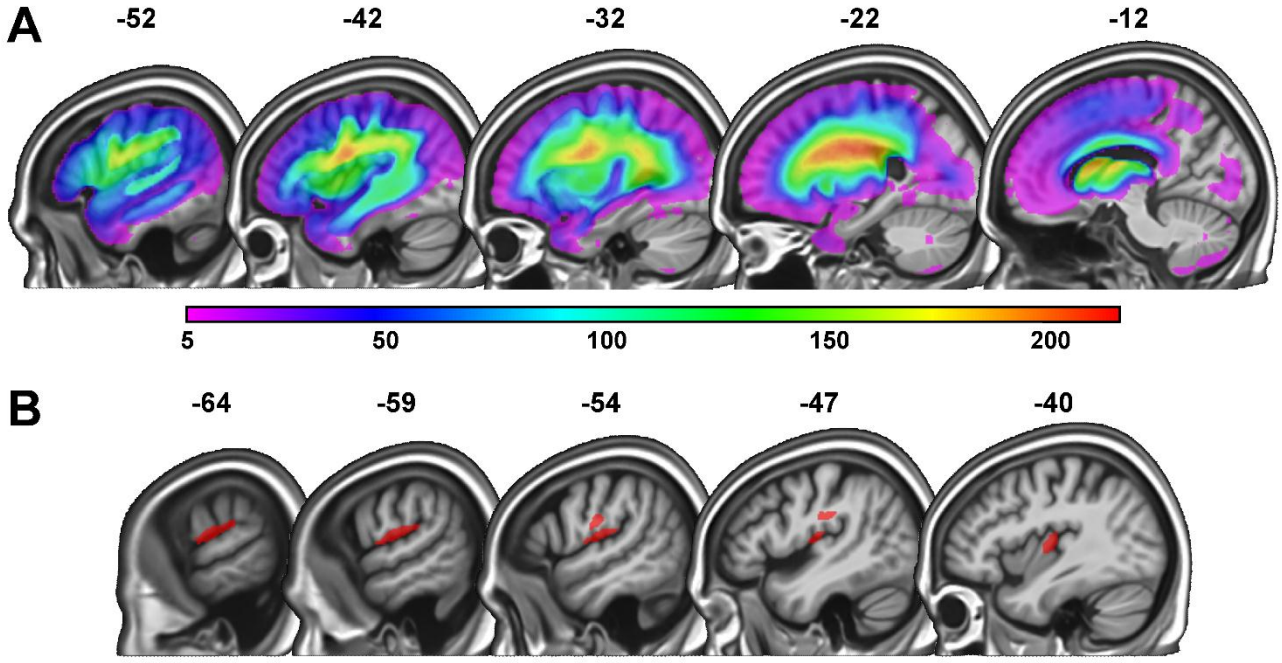


Fig. 2. Lesion overlap map and region of interest from Analysis 1. **(A)** Lesion overlap map for the full sample of 360 stroke patients, depicting voxels that were damaged in a minimum of 5 and a maximum of 215 patients. The colour scale indicates the number of patients with overlapping lesions at each given voxel. **(B)** In red, the region of interest identified in Analysis 1 (i.e. 549 voxels) where a significant association between lesion load and speech articulation abilities was found.

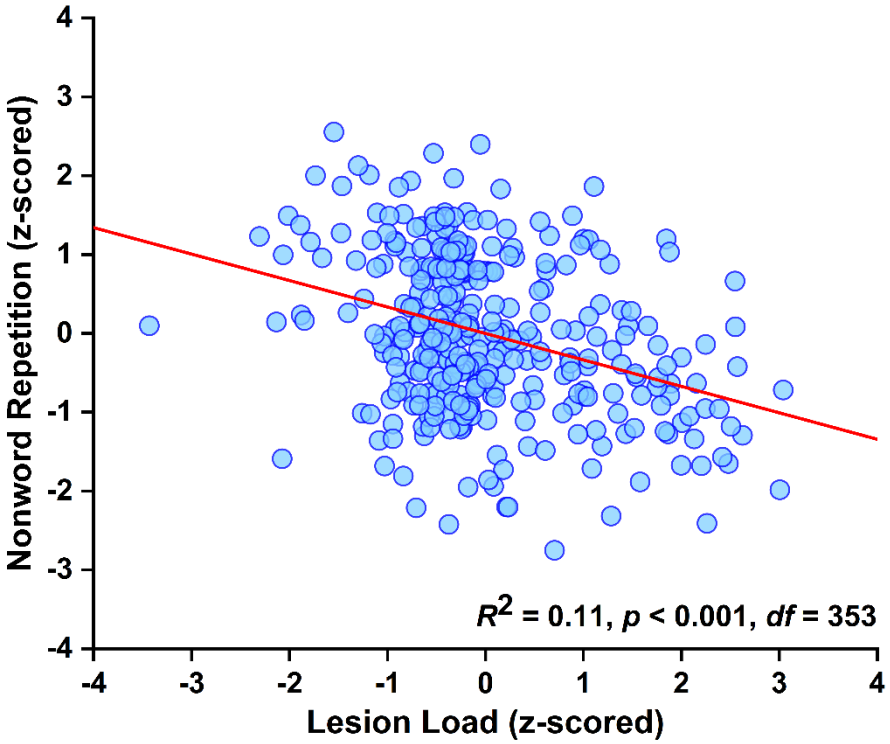


Fig. 3. Effect of interest. Visual illustration of the strength of the relationship between lesion load in the region of interest and nonword repetition scores, after factoring out variance explained by the covariates of no interest (i.e. a plot of the lesion load and nonword repetition residuals; Analysis 1).

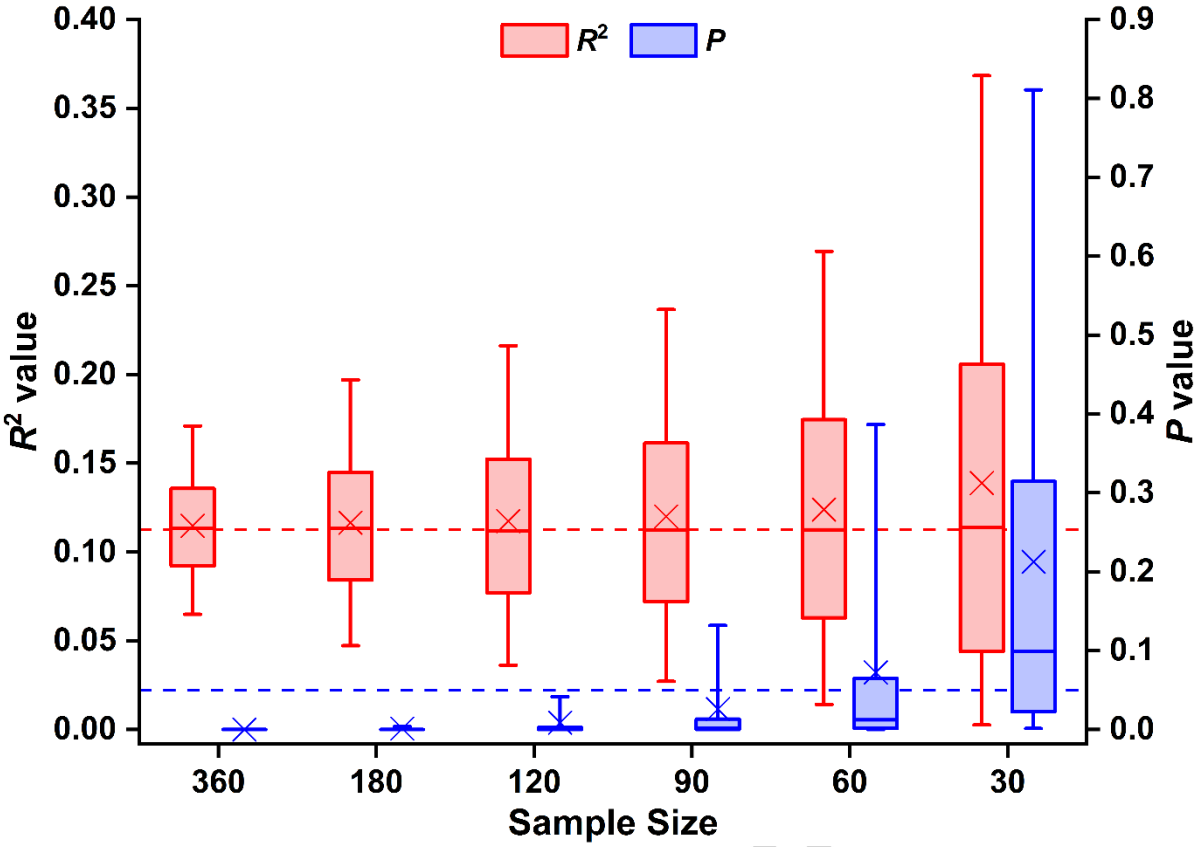


Fig. 4. Differential sensitivity of effect sizes and *p* values to sample size. The figure highlights that, while the mean and median of the effect size distributions remained relatively constant across the different sample sizes, the mean and median of the *p* value distributions exhibited substantial and systematic variability. Box plots depict medians with interquartile ranges and whiskers represent the 5th and 95th percentiles. The crosses indicate the mean for each sample size. The horizontal dashed line in red signals the R^2 value obtained in Analysis 1 (including data from all 360 patients), whereas the horizontal dashed line in blue shows the standard alpha level (i.e. 0.05).

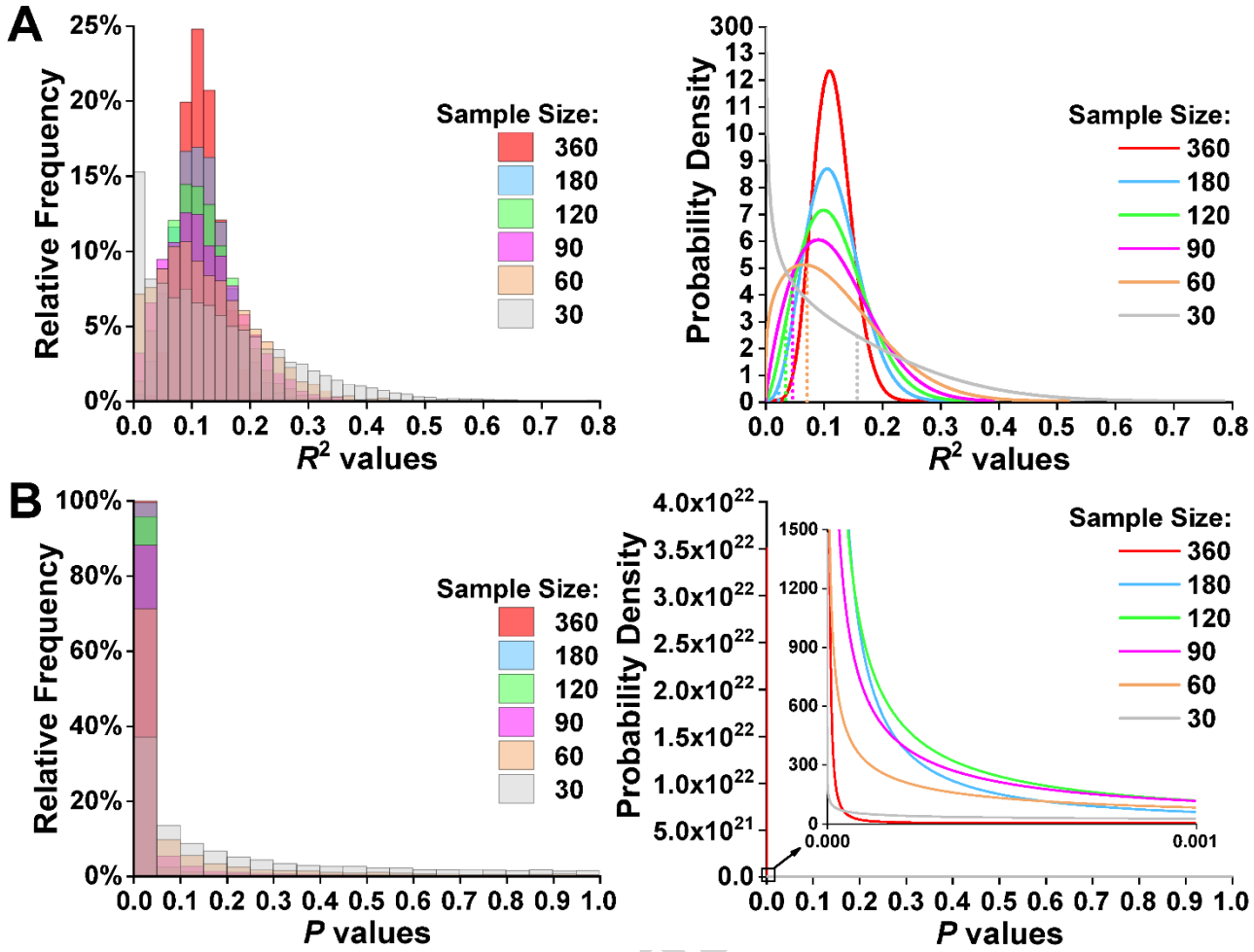


Fig. 5. Distribution of R^2 and p values. **(A)** From left to right, the frequency (in intervals of 0.02) and probability distributions of effect sizes for each sample size. The vertical dotted lines indicate the boundary between non-significant ($p \geq 0.05$; to the left) and significant ($p < 0.05$; to the right) R^2 values. **(B)** From left to right, the frequency (in intervals of 0.05) and probability distributions of p values for each sample size.

References

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-Size Planning for More Accurate Statistical Power: A Method Adjusting Sample Effect Sizes for Publication Bias and Uncertainty. *Psychol Sci, 28*, 1547-1562.

Ashburner, J., & Friston, K. J. (2000). Voxel-based morphometry-the methods. *NeuroImage, 11*, 805-821.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspect Psychol Sci, 7*, 543-554.

Baldo, J. V., Wilkins, D. P., Ogar, J., Willock, S., & Dronkers, N. F. (2011). Role of the precentral gyrus of the insula in complex articulation. *Cortex, 47*, 800-807.

Basilakos, A., Rorden, C., Bonilha, L., Moser, D., & Fridriksson, J. (2015). Patterns of poststroke brain damage that predict speech production errors in apraxia of speech and aphasia dissociate. *Stroke, 46*, 1561-1566.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., . . . Johnson, V. E. (2018). Redefine statistical significance. *Nat Hum Behav, 2*, 6-10.

Boekel, W., Wagenmakers, E.-J., Belay, L., Verhagen, J., Brown, S., & Forstmann, B. U. (2015). A purely confirmatory replication study of structural brain-behavior correlations. *Cortex, 66*, 115-133.

Brooks, J. L., Zoumpoulaki, A., & Bowman, H. (2017). Data-driven region-of-interest selection without inflating Type I error rate. *Psychophysiology, 54*, 100-113.

Butler, R. A., Lambon Ralph, M. A., & Woollams, A. M. (2014). Capturing multidimensionality in stroke aphasia: mapping principal behavioural components to neural structures. *Brain, 137*, 3248-3266.

Button, K. S. (2016). Statistical Rigor and the Perils of Chance. *eNeuro, 3*.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013a). Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci, 14*, 365-376.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013b). Confidence and precision increase with high statistical power. *Nat Rev Neurosci, 14*, 585-586.

Chen, G., Taylor, P. A., & Cox, R. W. (2017). Is the statistic value all we should care about in neuroimaging? *NeuroImage, 147*, 952-959.

Chen, X., Lu, B., & Yan, C.-G. (2018). Reproducibility of R-fMRI metrics on the impact of different strategies for multiple comparison correction and sample sizes. *Hum Brain Mapp, 39*, 300-318.

Corbetta, M., Ramsey, L., Callejas, A., Baldassarre, A., Hacker, C. D., Siegel, J. S., . . . Shulman, G. L. (2015). Common behavioral clusters and subcortical anatomy in stroke. *Neuron, 85*, 927-941.

Cremers, H. R., Wager, T. D., & Yarkoni, T. (2017). The relation between statistical power and inference in fMRI. *PLoS One, 12*, e0184923.

Cumming, G. (2008). Replication and p Intervals: p Values predict the future only vaguely, but confidence Intervals do much better. *Perspect Psychol Sci, 3*, 286-300.

Cumming, G. (2014). The new statistics: why and how. *Psychol Sci, 25*, 7-29.

David, S. P., Ware, J. J., Chu, I. M., Loftus, P. D., Fusar-Poli, P., Radua, J., . . . Ioannidis, J. P. A. (2013). Potential reporting bias in fMRI studies of the brain. *PLoS One, 8*, e70104.

de Haan, B., & Karnath, H.-O. (2018). A hitchhiker's guide to lesion-behaviour mapping. *Neuropsychologia*, *this issue.*

Deichmann, R., Schwarzbauer, C., & Turner, R. (2004). Optimisation of the 3D MDEFT sequence for anatomical brain imaging: technical implications at 1.5 and 3 T. *NeuroImage, 21*, 757-767.

Dressing, A., Nitschke, K., Kümmerer, D., Bormann, T., Beume, L., Schmidt, C. S. M., . . . Martin, M. (2018). Distinct contributions of dorsal and ventral streams to imitation of tool-use and communicative gestures. *Cereb Cortex, 28*, 474-492.

Du, H., Liu, F., & Wang, L. (2017). A Bayesian "fill-in" method for correcting for publication bias in meta-analysis. *Psychol Methods, 22*, 799-817.

Falk, E. B., Hyde, L. W., Mitchell, C., Faul, J., Gonzalez, R., Heitzeg, M. M., . . . Schulenberg, J. (2013). What is a representative brain? Neuroscience meets population science. *Proc Natl Acad Sci USA, 110*, 17615-17622.

Fanelli, D. (2010). "Positive" results increase down the Hierarchy of the Sciences. *PLoS One, 5*, e10068.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*, 891-904.

Flandin, G., & Friston, K. J. (2015). Topological Inference A2 - Toga, Arthur W. In *Brain Mapping* (pp. 495-500). Waltham: Academic Press.

Forstmeier, W., Wagenmakers, E.-J., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings-a practical guide. *Biol Rev Camb Philos Soc, 92*, 1941-1968.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Social science. Publication bias in the social sciences: unlocking the file drawer. *Science, 345*, 1502-1505.

Fridriksson, J., Guo, D., Fillmore, P., Holland, A., & Rorden, C. (2013). Damage to the anterior arcuate fasciculus predicts non-fluent speech production in aphasia. *Brain, 136*, 3451-3460.

Fridriksson, J., Yourganov, G., Bonilha, L., Basilakos, A., Den Ouden, D.-B., & Rorden, C. (2016). Revealing the dual streams of speech processing. *Proc Natl Acad Sci USA, 113*, 15108-15113.

Friston, K. (2012). Ten ironic rules for non-statistical reviewers. *NeuroImage, 61*, 1300-1310.

Fusar-Poli, P., Radua, J., Frascarelli, M., Mechelli, A., Borgwardt, S., Di Fabio, F., . . . David, S. P. (2014). Evidence of reporting biases in voxel-based morphometry (VBM) studies of psychiatric and neurological disorders. *Hum Brain Mapp, 35*, 3052-3065.

Gajardo-Vidal, A., Lorca-Puls, D.L., Crinion, J., White, J., Seghier, M.L., Leff, A.P., Hope, T.M.H., Ludersdorfer, P., Green, D.W., Bowman, H., Price, C.J.(2018) How distributed processing produces false negatives in voxel-based lesion-deficit analyses. *Neuropsychologia, this issue.*

Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspect Psychol Sci, 9*, 641-651.

Gelman, A., & Geurts, H. M. (2017). The statistical crisis in science: how is it relevant to clinical neuropsychology? *Clin Neuropsychol, 31*, 1000-1014.

Gronau, Q. F., Duizer, M., Bakker, M., & Wagenmakers, E.-J. (2017). Bayesian mixture modeling of significant p values: A meta-analytic method to estimate the degree of contamination from H_0. *J Exp Psychol Gen, 146*, 1223-1233.

Halai, A. D., Woollams, A. M., & Lambon Ralph, M. A. (2017). Using principal component analysis to capture individual differences within a unified neuropsychological model of chronic post-stroke aphasia: Revealing the unique neural correlates of speech fluency, phonology and semantics. *Cortex, 86*, 275-289.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nat Methods, 12*, 179-185.

Hentschke, H., & Stuttgen, M. C. (2011). Computation of measures of effect size for neuroscience data sets. *Eur J Neurosci, 34*, 1887-1894.

Higginson, A. D., & Munafò, M. R. (2016). Current Incentives for Scientists Lead to Underpowered Studies with Erroneous Conclusions. *PLoS Biol, 14*, e2000995.

Hillis, A. E., Work, M., Barker, P. B., Jacobs, M. A., Breese, E. L., & Maurer, K. (2004). Re-examining the brain regions crucial for orchestrating speech articulation. *Brain, 127*, 1479-1487.

Hope, T. M., Parker, J., Grogan, A., Crinion, J., Rae, J., Ruffle, L., . . . Green, D. W. (2015). Comparing language outcomes in monolingual and bilingual stroke patients. *Brain, 138*, 1070-1083.

Ingre, M. (2013). Why small low-powered studies are worse than large high-powered studies and how to protect against "trivial" findings in research: comment on Friston (2012). *NeuroImage, 81*, 496-498.

Inoue, K., Madhyastha, T., Rudrauf, D., Mehta, S., & Grabowski, T. (2014). What affects detectability of lesion-deficit relationships in lesion studies? *NeuroImage Clin, 6*, 388-397.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med, 2*, e124.

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

Ioannidis, J. P. A. (2011). Excess significance bias in the literature on brain volume abnormalities. *Arch Gen Psychiatry, 68*, 773-780.

Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends Cogn Sci, 18*, 235-241.

Ioannidis, J. P. A., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clin Trials, 4*, 245-253.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proc Natl Acad Sci USA, 110*, 19313-19317.

Johnson, V. E., Payne, R. D., Wang, T. Y., Asher, A., & Mandal, S. (2017). On the Reproducibility of Psychological Science. *J Am Stat Assoc, 112*, 1-10.

Karnath, H.-O., Sperber, C., & Rorden, C. (2018). Mapping human brain lesions and their functional consequences. *NeuroImage, 165*, 180-189.

Kimberg, D. Y., Coslett, H. B., & Schwartz, M. F. (2007). Power in Voxel-based lesion-symptom mapping. *J Cogn Neurosci, 19*, 1067-1080.

Lacey, E. H., Skipper-Kallal, L. M., Xing, S., Fama, M. E., & Turkeltaub, P. E. (2017). Mapping Common Aphasia Assessments to Underlying Cognitive Processes and Their Neural Substrates. *Neurorehabil Neural Repair, 31*, 442-450.

Lakens, D., & Evers, E. R. K. (2014). Sailing From the Seas of Chaos Into the Corridor of Stability: Practical Recommendations to Increase the Informational Value of Studies. *Perspect Psychol Sci, 9*, 278-292.

LeWinn, K. Z., Sheridan, M. A., Keyes, K. M., Hamilton, A., & McLaughlin, K. A. (2017). Sample composition alters associations between age and brain structure. *Nat Commun, 8*, 874.

Lindquist, M. A., Caffo, B., & Crainiceanu, C. (2013). Ironing out the statistical wrinkles in "ten ironic rules". *NeuroImage, 81*, 499-502.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science, 355*, 584-585.

Mah, Y.-H., Husain, M., Rees, G., & Nachev, P. (2014). Human brain lesion-deficit inference remapped. *Brain, 137*, 2522-2531.

Mechelli, A., Price, C. J., Friston, K. J., & Ashburner, J. (2005). Voxel-based morphometry of the human brain: Methods and applications. *Curr Med Imaging Rev, 1*, 105-113.

Mehta, S., Grabowski, T. J., Trivedi, Y., & Damasio, H. (2003). Evaluation of voxel-based morphometry for focal lesion detection in individuals. *NeuroImage, 20*, 1438-1454.

Mirman, D., Chen, Q., Zhang, Y., Wang, Z., Faseyitan, O. K., Coslett, H. B., & Schwartz, M. F. (2015a). Neural organization of spoken language revealed by lesion-symptom mapping. *Nat Commun, 6*, 6762.

Mirman, D., Zhang, Y., Wang, Z., Coslett, H. B., & Schwartz, M. F. (2015b). The ins and outs of meaning: Behavioral and neuroanatomical dissociation of semantically-driven word retrieval and multimodal semantic recognition in aphasia. *Neuropsychologia, 76*, 208-219.

Mirman, D., Landrigan, J.-F., Kokolis, S., Verillo, S., Ferrara, C., & Pustina, D. (2018). Corrections for multiple comparisons in voxel-based lesion-symptom mapping. *Neuropsychologia*, *this issue*.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic bulletin & review, 23*, 103-123.

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: a comment on Cumming (2014). *Psychol Sci, 25*, 1289-1290.

Müller, V. I., Cieslik, E. C., Laird, A. R., Fox, P. T., Radua, J., Mataix-Cols, D., . . . Eickhoff, S. B. (2018). Ten simple rules for neuroimaging meta-analysis. *Neurosci Biobehav Rev, 84*, 151-161.

Mummery, C. J., Patterson, K., Price, C. J., Ashburner, J., Frackowiak, R. S., & Hodges, J. R. (2000). A voxel-based morphometry study of semantic dementia: relationship between temporal lobe atrophy and semantic memory. *Ann Neurol, 47*, 36-45.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., . . . Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nat Hum Behav, 1,* 0021.

Nissen, S. B., Magidson, T., Gross, K., & Bergstrom, C. T. (2016). Publication bias and the canonization of false facts. *eLife, 5*.

Ogar, J., Willock, S., Baldo, J., Wilkins, D., Ludy, C., & Dronkers, N. (2006). Clinical and anatomical correlates of apraxia of speech. *Brain Lang, 97*, 343-350.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspect Psychol Sci, 7*, 528-530.

Paus, T. (2010). Population neuroscience: why and how. *Hum Brain Mapp, 31*, 891-903.

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard Power as a Protection Against Imprecise Power Estimates. *Perspect Psychol Sci, 9*, 319-332.

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., . . . Yarkoni, T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat Rev Neurosci, 18*, 115-126.

Price, C. J., Crinion, J. T., Leff, A. P., Richardson, F. M., Schofield, T. M., Prejawa, S., . . . Seghier, M. L. (2010). Lesion sites that predict the ability to gesture how an object is used. *Arch Ital Biol, 148*, 243-258.

Price, C. J., Hope, T. M., & Seghier, M. L. (2017). Ten problems and solutions when predicting individual outcome from lesion site after stroke. *NeuroImage, 145, Part B*, 200-208.

Price, C. J., Seghier, M. L., & Leff, A. P. (2010). Predicting language outcome and recovery after stroke: the PLORAS system. *Nat Rev Neurol, 6*, 202-210.

Pustina, D., Avants, B., Faseyitan, O. K., Medaglia, J. D., & Coslett, H. B. (2018). Improved accuracy of lesion to symptom mapping with multivariate sparse canonical correlations. *Neuropsychologia, this issue*.

Ramsey, L. E., Siegel, J. S., Lang, C. E., Strube, M., Shulman, G. L., & Corbetta, M. (2017). Behavioural clusters and predictors of performance during recovery from stroke. *Nat Hum Behav, 1, 0038* .

Reddan, M. C., Lindquist, M. A., & Wager, T. D. (2017). Effect Size Estimation in Neuroimaging. *JAMA Psychiatry, 74*, 207-208.

Rorden, C., Karnath, H.-O., & Bonilha, L. (2007). Improving lesion-symptom mapping. *J Cogn Neurosci, 19*, 1081-1088.

Rousselet, G. A., & Pernet, C. R. (2012). Improving standards in brain-behavior correlation analyses. *Front Hum Neurosci, 6*, 119.

Schwartz, M. F., Faseyitan, O., Kim, J., & Coslett, H. B. (2012). The dorsal stream contribution to phonological retrieval in object naming. *Brain, 135*, 3799-3814.

Schwartz, M. F., Kimberg, D. Y., Walker, G. M., Brecher, A., Faseyitan, O. K., Dell, G. S., . . . Coslett, H. B. (2011). Neuroanatomical dissociation for taxonomic and thematic knowledge in the human brain. *Proc Natl Acad Sci USA,108*, 8520-8524.

Seghier, M. L., Friston, K. J., & Price, C. J. (2007). Detecting subject-specific activations using fuzzy clustering. *NeuroImage, 36*, 594-605.

Seghier, M. L., Patel, E., Prejawa, S., Ramsden, S., Selmer, A., Lim, L., . . . Price, C. J. (2016). The PLORAS Database: A data repository for Predicting Language Outcome and Recovery After Stroke. *NeuroImage, 124*, 1208-1212.

Seghier, M. L., Ramlackhansingh, A., Crinion, J., Leff, A. P., & Price, C. J. (2008). Lesion identification using unified segmentation-normalisation models and fuzzy clustering. *NeuroImage, 41*, 1253-1266.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). p-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results. *Perspect Psychol Sci, 9*, 666-681.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). P-curve: a key to the file-drawer. *J Exp Psychol Gen, 143*, 534-547.

Smith, S. M., & Nichols, T. E. (2018). Statistical Challenges in "Big Data" Human Neuroimaging. *Neuron, 97*, 263-268.

Sperber, C., & Karnath, H.-O. (2017). Impact of correction factors in human brain lesion-behavior inference. *Hum Brain Mapp, 38*, 1692-1701.

Sperber, C., & Karnath, H.-O. (2018). On the validity of lesion-behaviour mapping methods. *Neuropsychologia, this issue.*

Stamatakis, E. A., & Tyler, L. K. (2005). Identifying lesions on structural brain images-validation of the method and application to neuropsychological patients. *Brain Lang, 94*, 167-177.

Stanley, D. J., & Spence, J. R. (2014). Expectations for Replications: Are Yours Realistic? *Perspect Psychol Sci, 9*, 305-318.

Steward, O. (2016). A Rhumba of "R's": Replication, Reproducibility, Rigor, Robustness: What Does a Failure to Replicate Mean? *eNeuro, 3.*

Swinburn, K., Porter, G., & Howard, D. (2004). *Comprehensive Aphasia Test*. Hove: Psychology Press.

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS biology, 15*, e2000797.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic Appl Soc Psych, 37*.

Tyler, L. K., Marslen-Wilson, W., & Stamatakis, E. A. (2005). Dissociating neuro-cognitive component processes: voxel-based correlational methodology. *Neuropsychologia, 43*, 771-778.

Vsevolozhskaya, O., Ruiz, G., & Zaykin, D. (2017). Bayesian prediction intervals for assessing P-value variability in prospective replication studies. *Transl Psychiatry, 7*, 1271.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on p-Values: Context, Process, and Purpose. *Am Stat, 70*, 129-131.

Wende, K. C., Thiel, C., Sommer, J., Paulus, F. M., Krach, S., & Jansen, A. (2017). Mechanisms of hemispheric lateralization: A replication study. *Cortex, 94*, 182-192.

Yarkoni, T. (2009). Big Correlations in Little Studies: Inflated fMRI Correlations Reflect Low Statistical Power-Commentary on Vul et al. (2009). *Perspect Psychol Sci, 4*, 294-298.

Yourganov, G., Fridriksson, J., Rorden, C., Gleichgerrcht, E., & Bonilha, L. (2016). Multivariate Connectome-Based Symptom Mapping in Post-Stroke Patients: Networks Supporting Language and Speech. *J Neurosci, 36*, 6668-6679.

Zhang, Y., Kimberg, D. Y., Coslett, H. B., Schwartz, M. F., & Wang, Z. (2014). Multivariate lesion-symptom mapping using support vector regression. *Hum Brain Mapp, 35*, 5861-5876.

Highlights

- The same lesion-deficit analysis was repeated on thousands of bootstrap samples.
- Replicability of the original effect was contingent upon the size of the sample.
- With smaller samples, only inflated effect size estimates reached significance.
- With larger samples, even trivial effect sizes yielded significant *p* values.