NeuroImage 189 (2019) 248–257

Contents lists available at ScienceDirect

NeuroImage

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

Estimating effects of graded white matter damage and binary tract disconnection on post-stroke language impairment

Jason Geller *, Melissa Thye, Daniel Mirman

University of Alabama at Birmingham, USA

ARTICLE INFO

Keywords:
Aphasia
White matter
Replication
Lesion load
Disconnection
Lesion-symptom mapping

ABSTRACT

Despite the critical importance of close replications in strengthening and advancing scientific knowledge, there are inherent challenges to conducting replications of lesion-based studies. In the present study, we conducted a close conceptual replication of a study (i.e., Hope et al., 2016) that found that fluency and naming scores in post-stoke aphasia were more strongly associated with a binary measure of structural white matter integrity (tract disconnection) than a graded measure (lesion load). Using a different sample of stroke patients ($N = 128$) and four language deficit measures (aphasia severity, picture naming, and composite scores for speech production and semantic cognition), we examined tract disconnection and lesion load in three white matter tracts that have been implicated in language processing: arcuate fasciculus, uncinate fasciculus, and inferior fronto-occipital fasciculus. We did not find any consistent evidence that binary tract disconnection was more strongly associated with language impairment over and above lesion load, though individual deficit measures differed with respect to whether lesion load or tract disconnection was the stronger predictor. Given the mixed findings, we suggest caution when using such indirect estimates of structural white matter integrity, and direct individual measurements (for example, using diffusion weighted imaging) should be preferred when they are available. We end by highlighting the complex nature of replication in lesion-based studies and offer some potential solutions.

1. Introduction

1.1. Tract disconnection and language deficits

Neural models of language processing and studies examining the impact of focal brain damage on language functioning have consistently emphasized the importance of white matter tracts, often focusing on tract disconnection as an index of damage severity (Catani and Mesulam, 2008; Lichteim, 1885). White matter disruption contributes to the severity of language deficits after stroke (e.g., Forkel et al., 2014; Gleichgerrcht et al., 2017; Marebwa et al., 2017), and specific tracts appear to be important for particular language functions. For example, the arcuate fasciculus (AF) is important for speech production whereas semantic processing appears to more strongly rely on the inferior fronto-occipital fasciculus (IFOF) and uncinate fasciculus (UF; Acosta-Cabronero et al., 2011; Almairac et al., 2015; de Zubicaray et al., 2011; Han et al., 2013; Harvey et al., 2013). Most studies on this topic have either directly measured the integrity of white matter connections using diffusion-weighted imaging (DWI) or estimated white matter damage by using a probabilistic white matter atlas and calculating the amount of overlap between a critical region and the likely location of white matter tracts. Although DWI provides a more reliable measure of white matter damage after stroke, the acquisition of diffusion imaging scans is not standard practice in clinical settings, so researchers often rely on more indirect measures such as atlas-based estimates of tract damage and disconnection to study white matter damage.

In one such study, Marchina et al. (2011) examined an indirect measure of white matter integrity—"lesion load" (i.e., percent overlap between the lesion and the tract of interest)—to assess the relationship between three white matter tracts (i.e., arcuate fasciculus, uncinate fasciculus, and extreme capsule) and speech production. Greater lesion load in the arcuate fasciculus was associated with deficits in fluency measures of speech production, and this effect was not seen for either the uncinate or the extreme capsule. In a replication and extension of this research, Hope et al. (2016) examined whether a different indirect measure of tract damage, binary tract disconnection, was more informative than continuous lesion load in predicting deficit severity (see Fig. 1 for an illustration of lesion load versus tract disconnection). Replicating the work of Marchina et al. (2011), Hope et al. (2016) found that lesion load in the arcuate, but not the uncinate, predicted deficits in

* Corresponding author.415 Campbell Hall, 1530 3rd Avenue South, Birmingham, AL, 35294-1170, USA.
E-mail address: jgeller1@uab.edu (J. Geller).

https://doi.org/10.1016/j.neuroimage.2019.01.020
Received 10 October 2018; Received in revised form 5 January 2019; Accepted 8 January 2019
Available online 14 January 2019

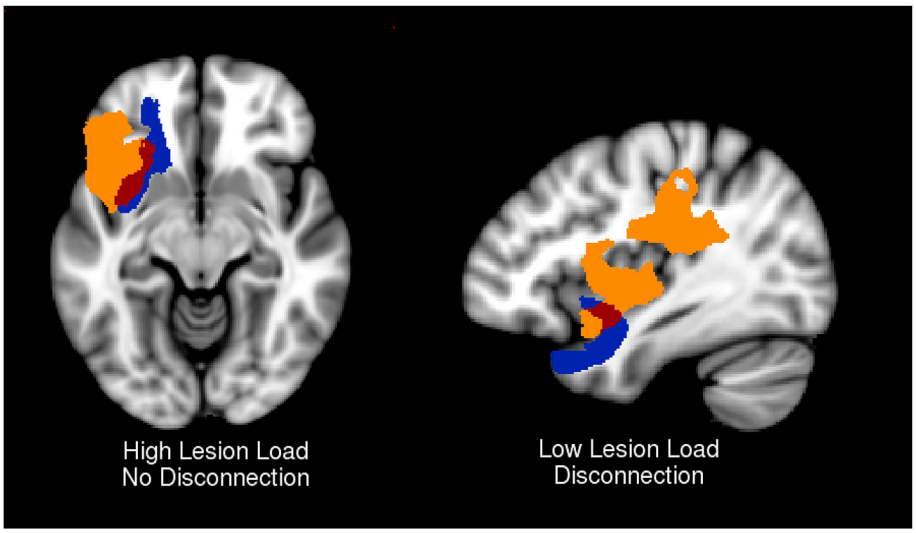

Fig. 1. Relationship between lesion load and tract disconnection. Data from two participants illustrating high lesion load (31%) with no tract disconnection (left) and low lesion load (3%) with tract disconnection (right). The uncinate is shown in blue, the lesion is shown in orange, and the overlap between the lesion and tract is shown in red.

fluency and naming. In addition, disconnection of both the AF and UF were associated with deficits in fluency and object naming, suggesting that tract disconnection may be a more sensitive and effective indirect measure of white matter integrity compared to lesion load. Hope et al. were primarily interested in the general implication that binary tract disconnection makes a unique contribution, with their specific measure being a promising measure of tract disconnection, though not necessarily the ideal one. Nevertheless, their results have substantial methodological implications. A simple, reliable, and meaningful way to estimate tract disconnection from structural scans would be a valuable tool for basic and clinical research, and perhaps even clinical practice. Given the important theoretical and applied implications, the present study sought to further test this approach to measuring tract disconnection in a replication using a broader range of language deficit measures and language-relevant white matter tracts.

1.2. Replication in lesion-based studies

There is a growing concern about reproducibility, particularly in the psychological and neurological sciences (see Boekel et al., 2015; Pashler and Harris, 2012), with one report claiming that more than half of the findings in the literature are spurious (Ioannidis, 2005). Bolstering this, a large-scale attempt to replicate 100 findings across cognitive and social psychology found that only 36% of findings replicated and that effects were, on average, half the size of the originally reported effect (Open Science Collaboration, 2015). In the field of neuroscience, Boekel et al. (2015) attempted to replicate 17 structural brain-behavior findings, but was only able to replicate 1 of the effects (6%). An obvious solution to the issue of reproducibility is to make replications more mainstream (Zwaan et al., 2018). Increasing the visibility of replications in journals will have the desirable effect of improving the credibility of research findings in the literature.

Replication attempts fall into two categories: close replications and conceptual replications (Zwaan et al., 2018). Close replications aim to reproduce the results of a study using the same methodological and analytic decisions. Conceptual replications, on the other hand, are designed to test the same theoretical ideas presented in a study, but with different methodological and analytic choices. With this type of replication, the result is greater insight through both replication and generation of new knowledge. Although close replication is the bedrock of scientific inquiry, financial and practical considerations can severely limit the feasibility of these studies. This is especially true in lesion-based research, which typically requires a large sample of participants (see Lorca-Puls et al., 2018) with particular neurocognitive profiles (e.g., cognitive performance, lesion characteristics) and expensive imaging procedures (e.g., MRI and/or CT scans). Recent recommendations indicate that a strongly powered replication should have a sample twice the size of the original study (Brandt et al., 2014; Simonsohn, 2015). By that standard, a close replication of a lesion-based study would cost an exorbitant amount of money and would require a massive recruitment effort. In general, a direct planned replication of a lesion study is practically and financially infeasible.

Another challenge concerns methodological reproducibility. Due to the lack of consensus regarding a standard lesion analysis method and rapid improvement of methods, a direct replication may involve utilizing sub-optimal methods in relation to imaging type, lesion segmentation, and lesion normalization. For example, 3T MRI provides the highest quality images for lesion analyses, but including 1.5T MRI and CT images may align with the methods adopted in an earlier paper and would allow for a larger sample size (because many individuals are either unable or unwilling to undergo 3T imaging). After the images are acquired, the lesioned regions need to be identified, which can be done manually, automatically (Griffis et al., 2016; Pustina et al., 2016), or semi-automatically (de Haan et al., 2015; de Haan and Karnath, 2018). Manual segmentation is generally considered the “gold standard”, but it is labor-intensive and inherently less replicable than automated methods (because it relies on subjective judgments that are based on in-house training protocols and/or a neurologist's expertise), making it unclear whether exactly replicating an original study's lesion segmentation method is ideal, or even possible. After segmentation, the native-space images need to be normalized to a common template and some registration methods are more robust than others (e.g., a rigid registration is unlikely to work well). An exact replication of a poor registration method would have limited scientific value.

For lesion-based research, it may be helpful to consider a continuum between close and conceptual replication rather than a strict binary distinction. Since direct replications of lesion studies are essentially impossible, all replications will be conceptual, though they may vary in their respective degree of separation from the original study. Given that many lesion studies are conducted by research groups with access to large patient databases, conceptual replication is viable and might be a better strategy. This is the approach we have taken in the present study.

1.3. The present study

With these issues in mind, the present study is a conceptual replication of Hope et al. (2016). In addition to replicating that study, the scope was expanded, both in terms of the tracts considered (arcuate, uncinate, and inferior fronto-occipital fasciculi) and the language deficit measures (aphasia severity, picture naming, and composite scores for speech production and semantic cognition). It is important to note that while Hope et al. (2016) did not include the IFOF, the study they replicated (Marchina et al., 2011) used the extreme capsule which is partially captured by the anterior portion of the IFOF and, like IFOF, is part of a white matter "bottleneck" that is particularly important for semantic processing (e.g., Mirman et al., 2015a, 2015b; Griffis et al., 2017). The inclusion of the IFOF extended the scope of the current paper to examine how damage to both IFOF and uncinate (critical components of the semantic bottleneck) affects semantic processing. Consistent with Hope et al. (2016) and Marchina et al. (2011), white matter damage was estimated based on overlap between individual lesion maps and probabilistic tractography maps. We re-examined a graded effect of tract lesion load (i.e., proportion overlap) and a binary tract disconnection measure for each language deficit measure.

A close replication of Hope et al. was not feasible for several of the practical reasons outlined in section 1.2. However, our data were analogous in the following respects: (1) we utilized a large data set of participants with left hemisphere stroke, (2) participants completed a battery of language tasks which tapped into the same language processes examined in the original paper (e.g., naming, speech production), and (3) lesion images were processed following segmentation and spatial normalization practices that are commonly utilized in lesion-based studies. As a result, the current study represents a close conceptual replication of Hope et al. using a larger set of tracts and language deficit measures, thereby further testing the robustness and generality of the relationship between tract disconnection and language deficits. See Table 1 for details regarding how the current conceptual replication deviated from the Marchina et al. (2011) and Hope et al. (2016) studies.

2. Methods

2.1. Data

The data were drawn from a large-scale, ongoing study of language processing following left hemisphere stroke conducted at the Moss Rehabilitation Research Institute (MRRI). Analyses of other language deficits in earlier subsets of the participants have been reported in several previous articles (Mirman et al., 2015; Mirman and Graziano, 2013; Mirman et al., 2015; Schwartz et al., 2009, 2011; Schwartz et al., 2012; Thothathiri et al., 2012; Walker et al., 2011), which also provide more detailed descriptions of the participants and imaging methods. The study was carried out in accordance with protocols approved by the Institutional Review Boards at the Einstein Healthcare Network and University of Pennsylvania School of Medicine.

The participants were 128 individuals with aphasia secondary to left hemisphere stroke (not bilateral or solely subcortical). To be included in this study, participants had to be at least 1 month post onset of aphasia

Table 1
Replication Details.

	Marchina et al. (2011) Study	Hope et al. (2016) Study	Current Study
Participant Data			
Sample Size (M:F)	30 (24:6)	142 (84:58)	128 (71:57)
Mean Age	58.50 years	52.10 years	58.20 years
Mean Time Post-Stroke	35.00 months	74.10 months	100.97 months
Scan Data (MRI:CT)	30:0	142:0	75:53
Exclusion Criteria	left-handedness; previous stroke; <11 months post-onset; other neurological conditions; right or bi-hemispheric stroke; severe comprehension or cognitive deficits (BDAE; RCPM)	left-handedness; <12 months post-onset; other neurological conditions; right or bi-hemispheric stroke; dispersed (not focal) damage; non-native English speaker; severe comprehension or cognitive deficits (CAT)	left-handedness; previous stroke; <1 month post-onset; other neurological conditions; right or bi-hemispheric stroke; vision or hearing difficulties; non-native English speaker
Language Scores			
Fluency/Speech Production	speech rate, informativeness, & efficiency	category & letter fluency (CAT)	PCA factor scores*
Naming	object naming (BNT)	object & action naming (CAT)	object naming (PNT)
Semantics	-	-	PCA factor scores*
Overall Language Impairment	-	-	WAB AQ
Tracts			
Probabilistic Atlas	DWI from 10 control participants	Thiebaut de Schotten et al., 2011	Rojkova et al., 2016
Tract Bookends	-	✓	✓
Results: Lesion Load			
Lesion Volume	not a significant predictor	not a significant predictor	***WAB AQ; semantics*** **
Arcuate	***fluency; naming***	***fluency; naming***	***WAB AQ; speech production; semantics***
Uncinate	not a significant predictor	not a significant predictor	***speech production***
Extreme Capsule	not a significant predictor	-	-
IFOF	-	-	***speech production*** **
Results: Tract Disconnection			
Lesion Volume	-	***naming***	***WAB AQ; naming; semantics***
Arcuate	-	***fluency; naming***	***WAB AQ; naming; speech production***
Uncinate	-	***fluency; naming***	not a significant predictor
IFOF	-	-	not a significant predictor

Note. BDAE, Boston Diagnostic Aphasia Evaluation; RCPM, Raven's Coloured Progressive Matrices; CAT, Comprehensive Aphasia Test; BNT, Boston Naming Test; PCA, principle component analysis; PNT, Picture Naming Test; WAB AQ, Western Aphasia Battery Aphasia Quotient; IFOF, inferior fronto-occipital fasciculus. Dashes indicate variables that were not included in the study. *The tests which loaded strongly onto each of these factors are described in greater detail in the manuscript and in the supplementary materials. **Positive relationship with increased white matter damage associated with better performance. Shading: White indicates very little deviation from the other studies with darker shades of green indicating greater deviations from the other studies. Where necessary additional details which differ between the studies are underlined to clarify the nature of the differences.

secondary to stroke, living at home, medically stable without major psychiatric or neurological co-morbidities, no previous history of stroke, and premorbidly right handed. Participants were also required to have English as the primary language, adequate vision and hearing (with or without correction) and computed tomography (CT) or magnetic resonance imaging (MRI) confirmed left hemisphere cortical lesion. Participants completed a detailed battery of psycholinguistic tests which have been described in previous studies (Mirman et al., 2010). Participant demographic and language assessment information is presented in Table 2.

2.2. Lesion location

Lesion location was assessed based on MRI ($n = 75$) or CT ($n = 53$) brain scans, following the same procedures as previous studies of this data set (or sub-sets of these data) (Mirman et al., 2015; Mirman and Graziano, 2013; Mirman et al., 2015; Schwartz et al., 2009, 2011; 2012; Thothathiri et al., 2012; Thye and Mirman, 2018; Walker et al., 2011). For the MRI scans, lesions were manually segmented on each participant's T1-weighted structural image, then the structural scans and lesion maps were normalized to the Montreal Neurological Institute (MNI) space Colin27 template by an automated process (Avants et al., 2006). The MRI lesion drawing was done by a trained technician. The CT scans were drawn by an expert neurologist. In both cases, the person doing the lesion drawing was blind to the behavioral performance of the participant. For the CT scans, the lesion was drawn directly onto the Colin27 template after rotating it (pitch only) to match the approximate slice plane of the participant's scan. The lesion overlap map for the full sample of participants is shown in Fig. 2.

2.3. Language scores

Picture naming ability (Philadelphia Naming Test; PNT) was used as an approximate replication of the naming score used by Hope et al. (2016). We also included overall aphasia severity (Western Aphasia Battery Aphasia Quotient; WAB AQ) as a general measure of language impairment. In addition, speech production and semantic cognition scores were calculated using a principal component analysis (PCA) that we have used in previous lesion symptom mapping studies of language sub-systems (Mirman et al., 2015; Mirman et al., 2015): participant scores on 17 psycholinguistic measures were entered into a principle component analysis with varimax rotation to obtain four factors (Semantic Cognition, Speech Production, Speech Recognition, and Semantic Errors) that accounted for 27%, 24%, 19%, and 7% of the variance respectively. In the current study, only factor scores for Semantic Cognition (e.g., Camel and Cactus Test, Pyramids and Palm Trees Test, synonym judgments, semantic category discrimination, Peabody Picture Vocabulary Test) and Speech Production (e.g., word and nonword repetition, phonological errors in picture naming, immediate serial recall span) were used for three reasons. First, the speech production factor was chosen because it roughly aligns with the fluency composite score used by Hope et al. (2016), so it contributes to the replication goal of this study. The speech production factor aligns with fluency in capturing the critical phonological aspects of speech planning and production, although the fluency measure used by Hope et al. also reflects word retrieval, memory, and cognitive control processes. Second, damage to the white matter tracts included in the current study has been consistently associated with speech production deficits, such as fluency and picture naming (Fridriksson et al., 2013; Wang et al., 2013), and semantic deficits (Han et al., 2013). Thus, the semantic cognition factor score was also included in this study as an extension of the Hope et al. (2016; also see Marchina et al., 2011) study. Damage to these tracts (and, to our knowledge, any other tracts), is not associated with speech recognition deficits; thus, there was no *a priori* reason to expect that damage to the white matter tracts of interest would meaningfully relate to deficits in speech recognition. Third, the Semantic Errors factor was characterized by a single high loading on semantic errors in picture naming and had an eigenvalue below 1.0 (0.915). Although studies of semantic errors are certainly valuable, this measure does not appear to capture a language deficit sub-domain and did not seem to be a good candidate measure for this study. See https://osf.io/3r7qn/for the correlation matrix and factor loadings for our speech production and semantic factors.

Table 2
Participant demographics.

	N	Mean (SD)	Range
Age	128	58.20 (11.68)	26–79
Years of Education	128	14.26 (2.97)	6–21
Lesion Size (mm^3)	128	100.97 (82.76)	5.38–376.12
Time Since Stroke (months)	128	51.59 (65.71)	1–381
WAB Aphasia Quotient	128	73.66 (19.38)	25.20–99.30
Philadelphia Naming Test (% correct)	128	64.92 (28.87)	1.10–97.70
Speech Production*	128	0 (1)	−3.44–1.56
Semantics*	128	0 (1)	−2.89–1.82
Gender (M:F)	71:57		
Aphasia subtype			
Anomic Aphasia	55		
Broca's Aphasia	31		
Conduction Aphasia	18		
Wernicke's Aphasia	10		
Transcortical Motor Aphasia	3		
Transcortical Sensory Aphasia	2		
Global Aphasia	1		
Other	8		

Note. N, number of participants; SD, standard deviation of the mean; WAB, Western Aphasia Battery; M, male; F, female; *factor scores from principle component analysis, which produces scores that are constrained to have Mean = 0 and SD = 1.0.

2.4. White matter tracts

The white matter tracts of interest—uncinate fasciculus (UF), arcuate fasciculus (AF), and inferior fronto-occipital fasciculus (IFOF)—were derived from a probabilistic white matter atlas (Rojkova et al., 2016). This atlas was constructed using an advanced spherical deconvolution diffusion tractography procedure on data from 47 participants to model the orientation of different fibers within a single voxel in order to capture the presence of crossing fibers, a common limitation of other diffusion tractography methods (Seunarine and Alexander, 2014). The final volume of each tract was constrained to the area where the tract was observed in at least 75% of the atlas sample. The lesion files were binarized and spatially normalized to the same stereotaxic space as the white matter tracts (MNI152) using symmetric normalization with cross-correlation (Avants et al., 2008) prior to calculation of lesion load and tract disconnection.

A potential problem that may arise when calculating tract disconnection (see section 2.5) is that a lesion may destroy the end of a tract while leaving the rest of the tract preserved. In this scenario, the calculation of the remaining tract clusters would return one cluster which would falsely suggest that the tract was preserved. To address this problem, Hope et al. (2016) manually created "bookends" (perpendicular planes placed at the extreme portions of the tracts) to create an extended boundary at the termination points of the tracts where disconnection could be calculated by examining whether any bookends were separated from the tract. The Hope et al. bookends were not publicly available and could not be obtained from the authors, so we created approximations of the bookends used in the original paper. Briefly, each bookend is a 50 × 2mm plane placed perpendicularly to the tract. In order to account for the variable neuroanatomy of the tracts and to ensure that the bookends were contained within the cortex, some of these bookends were placed *near* the edges of the tract rather than at the most extreme portion of the tract. In addition, to detect disconnection at the posterior portion and the two anterior extensions of the arcuate fasciculus, three

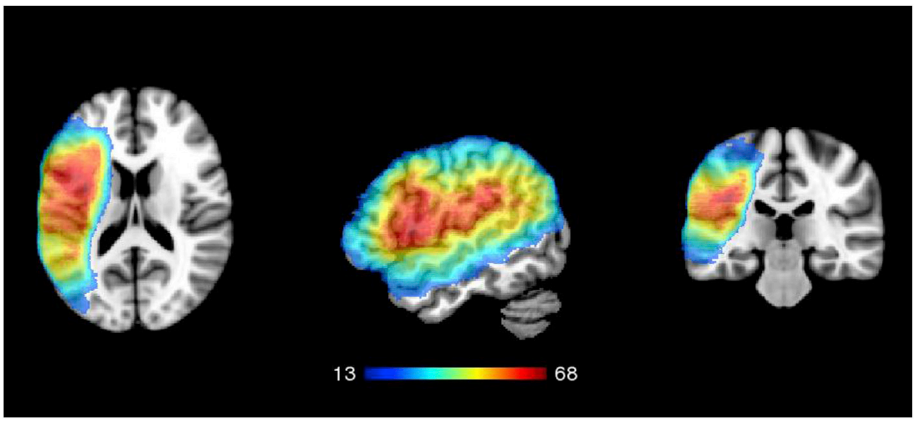

Fig. 2. Lesion overlap for full sample of participants (N = 128). Hotter colors indicate voxels where a larger number of participants had lesions. Only voxels where at least 10% of participants had lesions are shown in the figure and were included in the analyses.

bookends were used for this tract by Hope et al. as well as in the current study (Fig. 3). Additional anatomical information for each bookend is provided in Table 3 and the bookend files are publically available on OSF.

2.5. Lesion load and tract disconnection calculation

Our analyses focused on the distinction between lesion load in each tract and disconnection in those same tracts. Similar to Hope et al. (2016), lesion load was defined as the proportion of each tract image that is destroyed by (i.e., overlaps with) a given participant's binarized lesion image, ranging from 0% if the tract is completely unaffected by a lesion, to 100% when the tract is completely destroyed. Lesion load was calculated with the Lesionload function distributed as part of the LESYMAP package (Pustina et al., 2017).

To calculate tract disconnection, each participant's lesion image was subtracted from each tract, the bookends were added, and the labelClusters and labelStats functions from the ANTsR package (Avants et al., 2008) were used to count the number of clusters in the resulting three-dimensional image. The tract was considered to be disconnected if more than one cluster was identified in the subtracted image (e.g., if the tract had been divided into multiple distinct sections). See Table 4 for the number of connection and disconnection cases for each tract.

Table 3
MNI coordinates for the center of mass for each bookend.

	x	*y*	*z*
Arcuate			
Superior Anterior Bookend	40	−5	9
Inferior Anterior Bookend	41	27	1
Posterior Bookend	40	62	31
Uncinate			
Anterior Bookend	12	−53	−18
Posterior Bookend	27	−21	−30
IFOF			
Anterior Bookend	18	−55	−11
Posterior Bookend	14	93	0

Table 4
Number of connection versus disconnection cases for each tract.

	Connected	Disconnected	Lesion Load: Mean (range)
AF	42	86	32% (0–99)
UF	82	46	12% (0–68)
IFOF	98	30	13% (0–58)

3. Results

We examined the effect of lesion load and tract disconnection on overall aphasia severity, picture naming, and composite measures of speech production and semantic cognition. Lesion load and tract disconnection were tested separately as predictors across four stepwise regression analyses (one for each language measure). Overall lesion volume was included as a control variable. Stepwise selection alternates between forward and backward selection, adding variables that meet a statistical threshold for inclusion and removing variables that do not meet those criteria, until a stable set of variables is attained. Forward and backward selection based on Akaike information criterion (AIC) was used to determine the best fitting models. To determine statistical significance of predictors in the final model, a Bonferroni-corrected alpha of 0.02 was used. The analysis script used to perform each of the analyses reported below is located on OSF (https://osf.io/3r7qn/).

3.1. Best-fitting models

The best-fitting models determined by stepwise regression are summarized in Table 5. Lesion load and tract disconnection measures of white matter damage produced similar results, but there were some

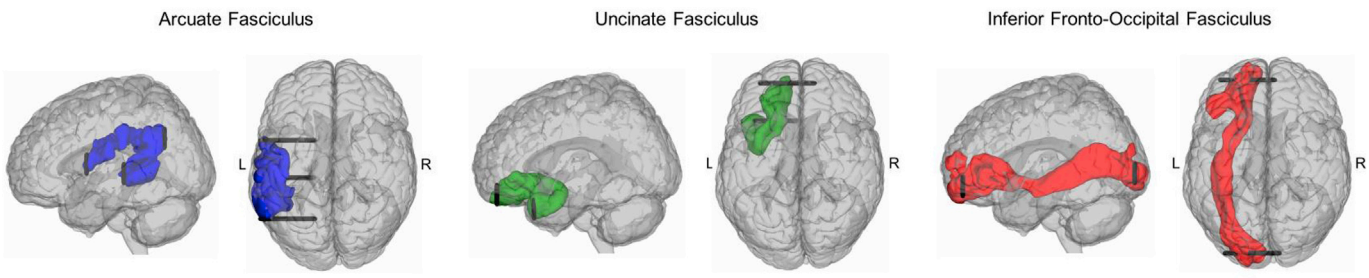

Fig. 3. White matter tracts. The arcuate fasciculus (blue), uncinate fasciculus (green), and inferior fronto-occipital fasciculus (red). The bookends for each tract are shown in black. L, left; R, right.

Table 5
Parameter estimates for the best-fitting models.

	WAB AQ	PNT Accuracy	Speech Production	Semantic Cognition
Lesion Load				
Lesion Size	0.00 (0.00) ***	n.s.	n.s.	−0.00 (0.00) ***
Arcuate Fasciculus	−16.64 (6.81) *	n.s.	−1.61 (0.32) ***	1.00 (0.40) *
Uncinate Fasciculus	n.s.	n.s.	−2.042 (0.76) **	n.s.
Inferior Fronto-Occipital Fasciculus	n.s.	n.s.	2.98 (0.98) **	n.s.
Disconnection				
Lesion Size	−0.00 (0.00) ***	−0.00 (0.00) ***	n.s.	−0.00 (0.00) **
Arcuate Fasciculus	−11.59 (3.46) **	−0.16 (0.056) ***	−0.98 (0.17) ***	n.s.
Uncinate Fasciculus	n.s.	n.s.	n.s.	n.s.
Inferior Fronto-Occipital Fasciculus	n.s.	n.s.	n.s.	n.s.

Note. Standard error of the mean is shown in parentheses next to the parameter estimates. Full model results can be found at our OSF page: https://osf.io/3r7qn/
.
*$p < .05$; **$p < .01$; ***$p < .001$.

notable differences. For overall aphasia severity (WAB AQ), lesion size and damage to the arcuate fasciculus were significant predictors in both the lesion load and tract disconnection models. For picture naming (PNT accuracy), overall lesion size and damage (disconnection) in the arcuate fasciculus were significant predictors only in the disconnection model. There were no significant predictors of picture naming in the lesion load model. For speech production, damage to the arcuate fasciculus was a significant predictor in both the lesion load and tract disconnection models. In addition, for the lesion load model, percent damage in the uncinate fasciculus was associated with speech production deficits whereas greater lesion load in the inferior fronto-occipital fasciculus was associated with *less severe* speech production deficits. For semantic cognition, overall lesion size was a significant predictor in both the lesion load and tract disconnection models, and, for the lesion load model, damage in the arcuate fasciculus was associated with *less severe* semantic deficits.

3.2. Lesion load vs. binary tract disconnection

Following Hope et al. (2016), we quantified the relative evidence for the lesion load and disconnection measures using hierarchal regression and Bayesian regression (Bayes Factor package in R; Morey and Rouder, 2011) analyses for each language test. For each test, we compared the full lesion load model to a full disconnection model. In the stepwise regression analyses, significant changes in R-squared were based on F tests. Following recommendations by Jefferys (1961), a Bayes factor (BF; i.e., marginal likelihood of one model against another model) between 1 and 3 was interpreted as equivalency between the lesion load and disconnection models; a BF between 3 and 10 was considered as substantial evidence for one model over the other; and a BF > 10 was considered strong evidence for one model over the other.

3.2.1. Hierarchal regression analyses

The relative contribution of tract disconnection was assessed by fitting the full lesion load model and then adding the disconnection measures. There was a marginal increase in the variance explained in overall language severity ($\Delta R^2 = 0.038$, $p = .059$) and picture naming ($\Delta R^2 = 0.038$, $p = .064$) and a significant increase in the amount of variance explained in speech production ($\Delta R^2 = 0.070$, $p = .009$) when disconnection measures were added to the lesion load model. Disconnection measures did not explain a significant amount of the variance in semantic cognition when added to the lesion load model ($\Delta R^2 = 0.028$, $p = .254$).

The relative contribution of lesion load was assessed by fitting the disconnection model and then adding the lesion load measures. There was a marginal increase in the amount of variance explained in language severity when the lesion load measures were added to the disconnection model ($\Delta R^2 = 0.034$, $p = .06$). The variance explained in speech production ($\Delta R^2 = 0.100$, $p = .001$) and semantic cognition ($\Delta R^2 = 0.063$, $p = .03$) increased when lesion load was added to the disconnection model, and the variance explained for speech production was greater than that observed for the lesion load measures. Lesion load measures did not explain a significant amount of the variance in picture naming when added to the disconnection model ($\Delta R^2 = 0.024$, $p = .244$).

3.2.2. Bayesian regression analyses

The Bayesian analyses converged with the hierarchical regression analyses. The tract disconnection model was preferred over the lesion load model for overall aphasia severity ($BF = 3.84$) and strongly preferred for picture naming ($BF > 10$). The lesion load model was preferred over the tract disconnection model for speech production ($BF = 6.44$) and strongly preferred for semantic cognition ($BF = 10.70$).

4. Discussion

4.1. Summary of results

Contemporary language models have emphasized the critical role of white-matter tracts in language processing (Catani et al., 2005). In most studies (e.g., Marchina et al., 2011), white matter damage is quantified as the continuous proportion of the white matter tract that is affected by an individual's lesion. Hope et al. (2016) suggested that binary tract disconnection captures an additional dimension of white matter damage and proposed a measure of such disconnection. With a sample of 128 participants with aphasia following left hemisphere stroke, we conducted an independent close conceptual replication and extension of Hope et al. (2016). We examined how two proxy measures of white matter integrity (lesion load and tract disconnection) in three key tracts (AF, UF, and IFOF) were related to four different language deficit measures: aphasia severity (WAB AQ), picture naming accuracy, speech production factor score, and semantic cognition factor score. Unsurprisingly, aphasia severity was associated with overall lesion size and with damage to the arcuate fasciculus. Damage to the AF was also associated with impaired speech production, consistent with previous work highlighting its role in fluent speech production (Hickok and Poeppel, 2004; Hope et al., 2016; Marchina et al., 2011). Lesion load in the UF was also associated with speech production deficits, although the UF is not commonly implicated in phonological aspects of speech production (but see Griffis et al., 2017; Hope et al., 2016). Additionally, we found that damage to the IFOF was associated with *better* speech production scores. These unexpected effects of UF and IFOF damage may be indirectly reflecting consequences of damage to nearby grey matter regions. Specifically, damage to inferior frontal cortex and anterior insula are associated with speech production deficits (Baldo et al., 2011; Dronkers, 1996; Ogar et al., 2006), and lesions affecting these regions may also be damaging the frontal portion of the uncinate fasciculus, thus producing an association between UF damage and speech production deficits. Damage to IFOF may reflect comparatively ventral lesions that spare the dorsal (parietal-frontal) stream system that is critical for speech production, thus producing an association between IFOF damage and better speech production scores. Similarly, AF lesion load was positively associated with semantic cognition, possibly because AF lesion load reflects parietal lesions that tend to spare the anterior temporal and bottleneck regions that are critical for semantic cognition. The IFOF and UF have both been implicated in semantic processing (Han et al., 2013), but we did not find an

association between damage to either of these tracts and semantic deficits. This could be because we did not have adequate coverage in those areas. Shahid et al. (2017) noted that adequate lesion coverage is crucial for detecting effects of interest.

The main question of interest in both Hope et al. (2016) and in this study was whether tract disconnection is a better predictor of language deficits than lesion load. For this question, the present results are mixed. Hope et al. (2016) found that regression models with tract disconnection measures consistently accounted for more variance in fluency and naming deficits than models with lesion load measures did. This was further bolstered by a Bayesian analysis. In the present study, Bayes factors indicated that tract disconnection models were preferred over lesion load models for picture naming and overall aphasia severity, and lesion load models were preferred over tract disconnection models for speech production and semantic factor scores. The hierarchical regression analyses also indicated that picture naming was better predicted by tract disconnection than lesion load whereas semantic cognition was better predicted by lesion load than tract disconnection. It is possible that tract disconnection is particularly useful when the deficit measure reflects a broad behavioral deficit based on multiple sub-systems (i.e., aphasia severity, fluency, and picture naming each rely on multiple distinct cognitive sub-systems). In contrast, lesion load may be a better predictor for more narrowly-defined deficits within a single cognitive sub-system (i.e., speech production and semantic scores that are based on a factor analysis designed to isolate functionally distinct sub-systems). We acknowledge that this is a *post hoc* speculation based on the observed pattern of results in the two studies on this topic and should be considered a hypothesis for further testing rather than a conclusion.

In summary, Hope et al. (2016) replicated an earlier study by Marchina et al. (2011), finding that AF lesion load was significantly related to naming and fluency, but UF lesion load and lesion volume were not. Additionally, Hope et al. found that disconnection of the AF or UF was also associated with naming and fluency deficits. The present results largely replicate the association of AF damage with aphasia severity, impaired picture naming (disconnection only), and speech production deficits. Where this study departs from previous work (Hope et al., 2016; Marchina et al., 2011) is that naming deficits were not related to lesion load in any of the tracts, but both AF and UF lesion load were associated with our fluency proxy. The latter difference may be related to differences between fluency measures: our speech production measure more specifically reflects phonological aspects of speech planning and production with minimal contributions of word retrieval or cognitive control processes.

4.2. Estimating white matter damage

These mixed results highlight the difficulty of estimating white matter integrity from indirect measures. It is important to recognize that neither tract disconnection nor lesion load are direct measures of white matter integrity – both of these are *estimates* of white matter damage based on aligning a normalized lesion map with a probabilistic white matter atlas. Tract disconnection is binary, so mis-estimations that result from individual differences in tract morphology and small errors during image registration can flip an individual's score to the opposite value (a connected tract may be estimated as disconnected and vice versa). Note that this is a measurement or estimation issue and does not rule out the theoretical claim that full disconnection of a white matter tract would have a unique effect on language performance that is not captured by overall amount of tract damage (Hope et al., 2018). Further, white matter damage is closely related to damage to the surrounding grey matter in this patient population, and the presentation and severity of deficits almost certainly reflects the consequences of a combination of grey matter and white matter damage. Measures such as lesion load and tract disconnection do not take into account surrounding grey matter damage. Examining the grey matter damage in conjunction with measures such as lesion load and tract disconnection may improve deficit prediction by constraining the analysis to individuals who have damage (e.g., high lesion load or disconnection) at a particular point along a tract underlying cortical damage.

One approach that may overcome this measurement problem is to leverage diffusion data to better localize white matter damage and directly quantify tract integrity, as several recent studies have done. Of particular interest is the recently-developed connectome-based lesion symptom mapping (CLSM) approach (Del Gaizo et al., 2017; Fridriksson et al., 2018; Gleichgerrcht et al., 2017; Yourganov et al., 2016), which generates a network of structural connections across the brain and relates behavioral or cognitive deficits to those portions of the network with damage (e.g. white matter underlying lesioned tissue). This method may provide a more comprehensive and detailed examination of how structural connections relate to language functioning, and it provides a complete view of white matter connections rather than relying on *a priori* tract selection. However, we are not aware of any direct comparisons showing that such diffusion-based measures are stronger or more reliable predictors than the template-based lesion load calculations used in the present study (and many other studies). Further, diffusion-based measures are a valuable research tool, but their clinical application will be limited by the challenge of collecting diffusion MRI data (and other advanced neuroimaging modalities) in clinical settings. Therefore, it would also be useful to find ways of robustly estimating white matter damage or dysfunction from routine clinical scans. The present results (see also Hope et al., 2018) suggest that indirect measures of white matter structural integrity may be of limited utility.

4.3. Replication in lesion-based research

As summarized in the introduction (section 1.2), lesion-based studies typically require large-scale collection of behavioral and neuroimaging data from a specific neurological population, which makes a planned direct replication essentially impossible for both practical and financial reasons. The flipside of the large-scale data collection requirement is that most research of this type is being carried out by research groups with large data sets, making conceptual replications generally easy to run using existing data.

The present study is a representative example of this point: a direct replication of Hope et al. (2016) would have required collecting behavioral and neuroimaging data from 150 to 300 individuals with aphasia following left hemisphere stroke (Hope et al. had $N = 146$ and it has been suggested that the sample size should be twice the size of the original study to ensure adequate power; e.g., Brandt et al., 2014; Simonsohn, 2015). For a single large medical research institution, this could take a decade and millions of dollars. However, we had a relatively large data set ($N = 128$) that contained behavioral and neuroimaging data that, although not identical to the Hope et al. measures, were appropriate for conducting a replication of their study. There are at least 2–4 other research groups that could similarly readily carry out close replications of lesion-based studies in the domain of post-stroke aphasia.

There are no hard and fast rules when it comes to replication. When adopting a close conceptual (rather than direct) replication there is some ambiguity about how close the replication should be and what the downstream effects of this are as one deviates further away from the original study. In the present study, we made a number of decisions that may have influenced our findings. One deviation from the original study was the use of different measures. The Comprehensive Aphasia Test was not administered as part of our battery, so it was not possible to use the same fluency and naming measures. Instead, we chose measures that capture analogous aspects of language processing – picture naming, speech production – and extended the analyses to include measures of aphasia severity and semantic cognition. As alluded to in the introduction, researchers conducting lesion studies are often limited to the resources available to them, and although this is an inherent aspect of conceptually replicating a previous study, it is possible that the use of different measures impacts replication. For instance, the naming task in

the Comprehensive Aphasia Test (CAT) includes both object and action naming whereas the PNT requires only object naming. Both are measures of picture naming and semantically-driven single word production, so the two measures should be highly correlated with one another, but the differences between them could affect the results. Similarly, the CAT fluency measure is broader than the speech production composite score used in the present study (which primarily captures phonological aspects of speech production), but these are closely related and should rely on similar neural substrates.

Differences regarding imaging type, time of assessment, and choice of atlas also affect closeness of replication. First, while Hope et al. (2016; also see Marchina et al., 2011) restricted their analyses to participants who had undergone MRI scans, we included participants who had undergone MRI or CT scans. The inclusion of either MRI or CT scans allowed us to use more data in the current study and increased our power.[1] Second, our study included behavioral assessments from a wide-range of times post-onset (1–384 months) whereas Marchina et al. (2011) and Hope et al. (2016) excluded patients who were less than 11 months and 12 months post-onset, respectively. Language abilities can change drastically over time and timing of assessment can be a critical factor to take into consideration (Shahid et al., 2017). In other analyses, we have found that excluding participants with sub-acute assessments (e.g., <6 months post-onset) does not affect the results, so any systematic influence (if any) of timing of assessment on the results remains unclear. Third, the tracts of interest in the current study were derived from an updated white matter atlas (Rojkova et al., 2016) that provides (seemingly) more accurate localization of the tracts compared to the atlas used by Hope et al. (2016).[2] We did not feel that closeness of replication was a sufficiently good reason to use an older and (seemingly) less precise white matter atlas. In addition, if the obtained results are atlas dependent, then this significantly undermines the clinical utility and robustness of the reported findings.

More generally, both original and replication lesion-based studies need to consider "best practices" in lesion-based research (see Sperber and Karnath, 2017), including controlling for lesion volume, only testing voxels or regions with sufficient lesion involvement, and correcting for multiple comparisons (e.g., Mirman et al., 2018). Another key best practice is sharing analysis methods. For example, the "bookends" used by Hope et al. and replicated (to the best of our ability) in the present study are not a standard aspect of lesion-based research and should be shared for replication purposes. To this end, the thresholded white matter tracts and bookends along with the preprocessing and analysis pipelines used in the present study are available on our OSF page. Although we cannot make the lesion files and the behavioral data that went into the analyses publicly available at this time, what is posted on our OSF page will help other researchers to replicate the present analyses, with new samples and measures.

4.4. Overcoming barriers

In the current study, we found inconsistent evidence for the conclusions of Hope et al. (2016). Traditionally, it would be highly unlikely for these findings to be publishable. Manuscripts that report mixed or null findings from a close or conceptual replication of a previously published paper face a double bias against publication. First, many journals and reviewers regard "novelty" as a key criterion for publication, which creates an inherent bias against replication studies of any sort. The word *novelty* appears in quotes in the previous sentence because it tends to be defined in a very specific way. Using the tract disconnection example from this report, the Hope et al. (2016) study was "novel" because there had not been a previous report of a tract disconnection analysis, whereas the present replication study does not fit this narrow definition of novelty and would be considered less impactful as a result. However, it is the first replication study examining the effect described initially by Hope et al. which could be considered a different kind of novelty. Because replicability/reproducibility is a hallmark of science, one way to overcome this form of publication bias is for journals and reviewers to consider replication to be an important contribution, possibly by broadening their definition of novelty to include first and/or strong replication studies.

Second, there is a bias in favor of publishing clear and conclusive results, and against mixed or null findings. This general bias affects all studies, not just replication studies, and creates an incentive for selective reporting of results (e.g., p-hacking). Pre-registration has been offered as a possible (partial) solution to this problem and, indeed, many journals have added a "registered reports" article format specifically to encourage researchers to pre-register their research (e.g., Cortex, eNeuro, European Journal of Neuroscience). Some journals even have a registered replications format to encourage planned replication studies. Study registration typically involves specifying the full study design before the data are collected. This is unlikely to work for lesion-based research because (as discussed above) the only feasible way to run lesion-based replication studies is to use existing data that have already been collected. Nevertheless, a study based on existing data can still be registered by specifying the hypotheses and critical replication targets, the data set to be used, and the analysis plan. Journals' guidelines for registered reports may need to be adjusted slightly to allow for this kind of study.

The final barrier is lack of incentives for running and publishing replication studies. Reducing publication bias would be an important step that would remove the disincentives, but this may not be sufficient. We suggest two additional strategies for making replication studies mainstream (see also Zwaan et al., 2018). First, a replication-and-extension approach (as in the present study, and in Hope et al., 2016) provides a way to include replication analyses along with new analyses. When following up on a study from another research group, researchers can begin with a replication of that previous study and include that replication analysis along with their follow-up when writing up the results for publication. This research approach is already quite common in the field, but the replication portion is often not included in the report because it is perceived as lacking novelty and importance. Overcoming these biases and including the replication portion in the report (possibly as an Appendix or Supplemental Materials if space in the main text is limited) would increase replication in lesion-based research. Second, replication studies provide a clear training opportunity for new researchers (see also Frank and Saxe, 2012; Hawkins et al., 2018). For example, when a student, post-doctoral fellow, or other trainee joins a lab and is planning to conduct lesion-based research, they could start by conducting a replication study. Because the hypotheses and design are (mostly) specified by the replication target (the original study), this is an opportunity to focus on learning the technical details of running lesion-based analyses and interpreting the results. The trainee can then apply these skills to new lesion-based studies.

5. Conclusion

In sum, the constraints of large-scale lesion-based research make planned direct replications essentially impossible, but close conceptual replications relatively easy. The bias toward selectively reporting only those studies that investigate novel hypotheses and report positive findings has a significant detrimental impact on the state of science. This "file drawer" problem skews the information available to researchers and clinicians attempting to synthesize the reported results into a converging theory. This is particularly problematic considering that the measures investigated here (e.g., indirect measures of white matter integrity) are commonly used in research studies examining white matter involvement in language functioning after stroke. The absence of non-confirmatory

[1] We ran multiple regressions for each measure, excluding CTs scans, and the effects were in the same direction, but some were not significant due to decreased power. Analysis results are available on our OSF page.

[2] We also conducted analyses using the older atlas and found that the results were consistent with the reported findings.

results may lead to the false impression that these measures are consistently useful to both researchers and clinicians in understanding the neural basis of language functioning. We discussed strategies for reducing biases against publication of replication studies, especially when the replication results are mixed or negative, and making replication research part of standard practice. These are important steps toward increasing replication in lesion-based research.

Author note

This research was supported by University of Alabama at Birmingham. The authors declare that they have no competing interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2019.01.020.

References

Acosta-Cabronero, J., Patterson, K., Fryer, T.D., Hodges, J.R., Pengas, G., Williams, G.B., Nestor, P.J., 2011. Atrophy, hypometabolism and white matter abnormalities in semantic dementia tell a coherent story. Brain 134 (7), 2025–2035. https://doi.org/10.1093/brain/awr119.

Almairac, F., Herbet, G., Moritz-Gasser, S., de Champfleur, N.M., Duffau, H., 2015. The left inferior fronto-occipital fasciculus subserves language semantics: a multilevel lesion study. Brain Struct. Funct. 220 (4), 1983–1995. https://doi.org/10.1007/s00429-014-0773-1.

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. 12 (1), 26–41. https://doi.org/10.1016/j.media.2007.06.004.

Avants, B.B., Schoenemann, P.T., Gee, J.C., 2006. Lagrangian frame diffeomorphic image registration: morphometric comparison of human and chimpanzee cortex. Med. Image Anal. 10, 397–412.

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 (7), 800–807. https://doi.org/10.1016/j.cortex.2010.07.001.

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. https://doi.org/10.1016/J.CORTEX.2014.11.019.

Brandt, M.J., IJzerman, H., Dijksterhuis, A., Farach, F.J., Geller, J., Giner-Sorolla, R., et al., 2014. The Replication Recipe: what makes for a convincing replication? J. Exp. Soc. Psychol. 50, 217–224. https://doi.org/10.1016/J.JESP.2013.10.005.

Catani, M., Jones, D.K., ffytche, D.H., 2005. Perisylvian language networks of the human brain. Ann. Neurol. 57 (1), 8–16. https://doi.org/10.1002/ana.20319.

Catani, M., Mesulam, M., 2008. The arcuate fasciculus and the disconnection theme in language and aphasia: history and current state. Cortex 44 (8), 953–961. https://doi.org/10.1016/j.cortex.2008.04.002.

de Haan, B., Clas, P., Juenger, H., Wilke, M., Karnath, H.-O., 2015. Fast semi-automated lesion demarcation in stroke. Neuroimage: Clinic 9, 69–74. https://doi.org/10.1016/j.nicl.2015.06.013.

de Haan, B., Karnath, H.-O., 2018. A hitchhiker's guide to lesion-behaviour mapping. Neuropsychologia 115, 5–16. https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2017.10.021.

de Zubicaray, G.I., Rose, S.E., McMahon, K.L., 2011. The structure and connectivity of semantic memory in the healthy older adult brain. Neuroimage 54 (2), 1488–1494. https://doi.org/10.1016/j.neuroimage.2010.08.058.

Del Gaizo, J., Fridriksson, J., Yourganov, G., Hillis, A.E., Hickok, G., Misic, B., et al., 2017. Mapping language networks using the structural and dynamic brain connectomes. ENeuro 4 (5), 2017. ENEURO.0204-17. https://doi.org/10.1523/ENEURO.0204-17.2017.

Dronkers, N.F., 1996. A new brain region for coordinating speech articulation. Nature 384 (6605), 159–161. https://doi.org/10.1038/384159a0.

Forkel, S.J., Thiebaut de Schotten, M., Dell'Acqua, F., Kalra, L., Murphy, D.G.M., Williams, S.C.R., Catani, M., 2014. Anatomical predictors of aphasia recovery: a tractography study of bilateral perisylvian language networks. Brain 137 (7), 2027–2039. https://doi.org/10.1093/brain/awu113.

Frank, M.C., Saxe, R., 2012. Teaching replication. Perspect. Psychol. Sci. 7 (6), 600–604. https://doi.org/10.1177/1745691612460686.

Fridriksson, J., den Ouden, D.-B., Hillis, A.E., Hickok, G., Rorden, C., Basilakos, A., et al., 2018. Anatomy of aphasia revisited. Brain 141 (3), 848–862. https://doi.org/10.1093/brain/awx363.

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 (Pt 11), 3451–3460. https://doi.org/10.1093/brain/awt267.

Gleichgerrcht, E., Fridriksson, J., Rorden, C., Bonilha, L., 2017. Connectome-based lesion-symptom mapping (CLSM): a novel approach to map neurological function. Neuroimage: Clinic 16, 461–467. July. https://doi.org/10.1016/j.nicl.2017.08.018.

Griffis, J.C., Allendorfer, J.B., Szaflarski, J.P., 2016. Voxel-based Gaussian naïve Bayes classification of ischemic stroke lesions in individual T1-weighted MRI scans. J. Neurosci. Methods 257, 97–108. https://doi.org/10.1016/j.jneumeth.2015.09.019.

Griffis, J.C., Nenert, R., Allendorfer, J.B., Szaflarski, J.P., 2017. Damage to white matter bottlenecks contributes to language impairments after left hemispheric stroke. Neuroimage: Clinic 14, 552–565. https://doi.org/10.1016/j.nicl.2017.02.019.

Han, Z., Ma, Y., Gong, G., He, Y., Caramazza, A., Bi, Y., 2013. White matter structural connectivity underlying semantic processing: evidence from brain damaged patients. Brain 136 (10), 2952–2965. https://doi.org/10.1093/brain/awt205.

Harvey, D.Y., Wei, T., Ellmore, T.M., Hamilton, A.C., Schnur, T.T., 2013. Neuropsychological evidence for the functional role of the uncinate fasciculus in semantic control. Neuropsychologia 51 (5), 789–801. https://doi.org/10.1016/j.neuropsychologia.2013.01.028.

Hawkins, R.X.D., Smith, E.N., Au, C., Arias, J.M., Catapano, R., Hermann, E., et al., 2018. Improving the Replicability of Psychological Science through Pedagogy. https://doi.org/10.1177/2515245917740427.

Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92 (1–2), 67–99. https://doi.org/10.1016/J.COGNITION.2003.10.011.

Hope, T.M.H., Leff, A.P., Price, C.J., 2018. Predicting language outcomes after stroke: is structural disconnection a useful predictor? Neuroimage: Clinic 19, 22–29. https://doi.org/10.1016/J.NICL.2018.03.037.

Hope, T.M.H., Seghier, M.L., Prejawa, S., Leff, A.P., Price, C.J., 2016. Distinguishing the effect of lesion load from tract disconnection in the arcuate and uncinate fasciculi. Neuroimage 125, 1169–1173. https://doi.org/10.1016/j.neuroimage.2015.09.025.

Ioannidis, J.P.A., 2005. Why most published research findings are false. PLoS Med. 2 (8), e124. https://doi.org/10.1371/journal.pmed.0020124.

Jefferys, H., 1961. Theory of Probability, third ed. Clarendon Press, Oxford.

Lichteim, L., 1885. On aphasia. Brain 7 (4), 433–484. https://doi.org/10.1093/brain/7.4.433.

Lorca-Puls, D.L., Gajardo-Vidal, A., White, J., Seghier, M.L., Leff, A.P., Green, D.W., et al., 2018. The impact of sample size on the reproducibility of voxel-based lesion-deficit mappings. Neuropsychologia 115, 101–111. https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2018.03.014.

Marchina, S., Zhu, L.L., Norton, A., Zipse, L., Wan, C.Y., Schlaug, G., 2011. Impairment of speech production predicted by lesion load of the left arcuate fasciculus. Stroke 42 (8), 2251–2256. https://doi.org/10.1161/STROKEAHA.110.606103.

Marebwa, B.K., Fridriksson, J., Yourganov, G., Feenaughty, L., Rorden, C., Bonilha, L., 2017. Chronic post-stroke aphasia severity is determined by fragmentation of residual white matter networks. Sci. Rep. 7 (1), 8188. https://doi.org/10.1038/s41598-017-07607-9.

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. https://doi.org/10.1038/ncomms7762.

Mirman, D., Graziano, K.M., 2013. The neural basis of inhibitory effects of semantic and phonological neighbors in spoken word production. J. Cognit. Neurosci. 25 (9), 1504–1516. https://doi.org/10.1162/jocn.

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 115, 112–123. https://doi.org/10.1016/j.neuropsychologia.2017.08.025.

Mirman, D., Strauss, T.J., Brecher, A., Walker, G.M., Sobel, P., Dell, G.S., Schwartz, M.F., 2010. A large, searchable, web-based database of aphasic performance on picture naming and other tests of cognitive function. Cogn. Neuropsychol. 27 (6), 495–504. https://doi.org/10.1080/02643294.2011.574112.

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. https://doi.org/10.1016/j.neuropsychologia.2015.02.014.

Morey, R.D., Rouder, J.N., 2011. Bayes factor approaches for testing interval null hypotheses. Psychol. Methods 16 (4), 406–419. https://doi.org/10.1037/a0024377.

Ogar, J., Willock, S., Baldo, J., Wilkins, D., Ludy, C., Dronkers, N., 2006. Clinical and anatomical correlates of apraxia of speech. Brain Lang. 97 (3), 343–350. https://doi.org/10.1016/j.bandl.2006.01.008.

Open Science Collaboration, O.S., 2015. PSYCHOLOGY. Estimating the reproducibility of psychological science. Science 349 (6251), aac4716. https://doi.org/10.1126/science.aac4716.

Pashler, H., Harris, C.R., 2012. Is the replicability crisis overblown? Three arguments examined. Perspect. Psychol. Sci. 7 (6), 531–536. https://doi.org/10.1177/1745691612463401.

Pustina, D., Avants, B., Faseyitan, O.K., Medaglia, J.D., Coslett, H.B., 2017. Improved accuracy of lesion to symptom mapping with multivariate sparse canonical correlations. Neuropsychologia. https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2017.08.027.

Pustina, D., Coslett, H.B., Turkeltaub, P.E., Tustison, N., Schwartz, M.F., Avants, B., 2016. Automated segmentation of chronic stroke lesions using LINDA: lesion identification with neighborhood data analysis. Hum. Brain Mapp. 37 (4), 1405–1421. https://doi.org/10.1002/hbm.23110.

Rojkova, K., Volle, E., Urbanski, M., Humbert, F., Dell'Acqua, F., Thiebaut de Schotten, M., 2016. Atlasing the frontal lobe connections and their variability due to age and education: a spherical deconvolution tractography study. Brain Struct. Funct. 221 (3), 1751–1766. https://doi.org/10.1007/s00429-015-1001-3.

Schwartz, M.F., Faseyitan, O.K., Kim, J., Coslett, H.B., 2012. The dorsal stream contribution to phonological retrieval in object naming. Brain 135 (12), 3799–3814. https://doi.org/10.1093/brain/aws300.

Schwartz, M.F., Kimberg, D.Y., Walker, G.M., Brecher, A., Faseyitan, O.K., Dell, G.S., et al., 2011. Neuroanatomical dissociation for taxonomic and thematic knowledge in the human brain. Proc. Natl. Acad. Sci. U. S. A. 108 (20), 8520–8524. https://doi.org/10.1073/pnas.1014935108.

Schwartz, M.F., Kimberg, D.Y., Walker, G.M., Faseyitan, O.K., Brecher, A.R., Dell, G.S., Coslett, H.B., 2009. Anterior temporal involvement in semantic word retrieval: voxel-based lesion-symptom mapping evidence from aphasia. Brain 132 (12), 3411–3427. https://doi.org/10.1093/brain/awp284.

Seunarine, K.K., Alexander, D.C., 2014. Multiple fibers: beyond the diffusion tensor. In: Diffusion MRI. Elsevier, pp. 105–123. https://doi.org/10.1016/B978-0-12-396460-1.00006-8.

Shahid, H., Sebastian, R., Schnur, T.T., Hanayik, T., Wright, A., Tippett, D.C., et al., 2017. Important considerations in lesion-symptom mapping: illustrations from studies of word comprehension. Hum. Brain Mapp. 38 (6), 2990–3000. https://doi.org/10.1002/hbm.23567.

Simonsohn, U., 2015. Small telescopes. Psychol. Sci. 26 (5), 559–569. https://doi.org/10.1177/0956797614567341.

Sperber, C., Karnath, H.-O., 2017. Impact of correction factors in human brain lesion-behavior inference. Hum. Brain Mapp. 38 (3), 1692–1701. https://doi.org/10.1002/hbm.23490.

Thothathiri, M., Kimberg, D.Y., Schwartz, M.F., 2012. The neural basis of reversible sentence comprehension: evidence from voxel-based lesion symptom mapping in aphasia. J. Cognit. Neurosci. 24 (1), 212–222. https://doi.org/10.1162/jocn_a_00118.

Thye, M., Mirman, D., 2018. Relative contributions of lesion location and lesion size to predictions of varied language deficits in post-stroke aphasia. Neuroimage: Clinic 20, 1129–1138. https://doi.org/10.1016/j.nicl.2018.10.017.

Walker, G.M., Schwartz, M.F., Kimberg, D.Y., Faseyitan, O.K., Brecher, A.R., Dell, G.S., Coslett, H.B., 2011. Support for anterior temporal involvement in semantic error production in aphasia: new evidence from VLSM. Brain Lang. 117 (3), 110–122.

Wang, J., Marchina, S., Norton, A.C., Wan, C.Y., Schlaug, G., 2013. Predicting speech fluency and naming abilities in aphasic patients. Front. Hum. Neurosci. 7, 831. https://doi.org/10.3389/fnhum.2013.00831.

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 (25), 6668–6679. https://doi.org/10.1523/JNEUROSCI.4396-15.2016.

Zwaan, R.A., Etz, A., Lucas, R.E., Donnellan, M.B., 2018. Making replication mainstream. Behav. Brain Sci. 41, e120. https://doi.org/10.1017/S0140525X17001972.