NeuroImage: Clinical 24 (2019) 102005

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

Generalizing post-stroke prognoses from research data to clinical data

Robert Loughnan[a], Diego L. Lorca-Puls[b,c], Andrea Gajardo-Vidal[b,c,d], Valeria Espejo-Videla[c], Céline R. Gillebert[e,f], Dante Mantini[e,g,h], Cathy J. Price[b], Thomas M.H. Hope[b,*]

[a] Department of Cognitive Science, University of California, San Diego, USA
[b] Wellcome Centre for Human Neuroimaging, UCL Queen Square Institute of Neurology, 12 Queen Square, London WC1N 3AR, UK
[c] Department of Speech, Language and Hearing Sciences, Faculty of Medicine, Universidad de Concepcion, Concepcion, Chile
[d] Faculty of Health Sciences, Universidad del Desarrollo, Concepcion, Chile
[e] Department of Experimental Psychology, University of Oxford, Oxford, UK
[f] Department of Brain and Cognition, University of Leuven, Leuven, Belgium
[g] Research Center for Movement Control and Neuroplasticity, University of Leuven, Leuven, Belgium
[h] Functional Neuroimaging Laboratory, IRCCS San Camillo Hospital Foundation, Venice, Italy

ARTICLE INFO

Keywords:
Stroke
Aphasia
Prognosis
Plasticity
Lesion growth

ABSTRACT

Around a third of stroke survivors suffer from acquired language disorders (aphasia), but current medicine cannot predict whether or when they might recover. Prognostic research in this area increasingly draws on datasets associating structural brain imaging data with outcome scores for ever-larger samples of stroke patients. The aim is to learn brain-behaviour trends from these data, and generalize those trends to predict outcomes for new patients. The practical significance of this work depends on the expected breadth of that generalization. Here, we show that these models can generalize across countries and native languages (from British patients tested in English to Chilean patients tested in Spanish), across neuroimaging technology (from MRI to CT), and from scans collected months or years after stroke for research purposes, to scans collected days or weeks after stroke for clinical purposes.

Our results suggest one important confound, in attempting to generalize from research data to clinical data, is the delay between scan acquisition and language assessment. This delay is typically small for research data, where scans and assessments are often acquired contemporaneously. But the most natural, clinical application of these predictions will employ acute prognostic factors to predict much longer-term outcomes. We mitigated this confound by projecting the clinical patients' lesions from the time when their scans were acquired, to the time when their language abilities were assessed; with this projection in place, there was strong evidence that prognoses derived from research data generalized equally well to research and clinical data. These results encourage attention to the confounding role that lesion growth may play in other types of lesion-symptom analysis.

1. Introduction

Around a third of stroke survivors suffer from acquired language disorders (Mozaffarian et al., 2015), or aphasia, yet current medicine cannot tell these people whether and when they might recover (Watila and Balarabe, 2015). To try to bridge this gap, researchers have begun to build large databases encoding structural brain imaging and demographic data on patients whose language outcomes are known, aiming to generalize trends learned from these data to new patients (Fridriksson et al., 2013; Halai et al., 2018; Mah et al., 2014; Nachev, 2015; Price et al., 2010; Pustina et al., 2017; Seghier et al., 2016; Yourganov et al., 2016). One important question for this work, asks what factors are expected to limit the generalization of the learned trends. In what follows, we address three of the most important, potential restrictions on the generalizability of models predicting language outcomes after stroke – and show that all three can be overcome.

The first potential restriction follows directly from our focus on language. For pragmatic reasons, research data in this field usually refer to native English-speaking stroke patients – because many of these studies are run in English-speaking countries such as the UK, the USA, and Canada. But the wider stroke survivor population is not so restricted. Post-stroke prognostic models typically work by learning associations between the site(s) and extent of lesion damage in the brain, and the language impairments (and recovery trajectories) consequent to that damage. These associations presumably depend on the functional architecture of language networks in the undamaged brain (e.g.,

* Corresponding author.
E-mail address: t.hope@ucl.ac.uk (T.M.H. Hope).

https://doi.org/10.1016/j.nicl.2019.102005
Received 6 April 2019; Received in revised form 10 September 2019; Accepted 14 September 2019
Available online 14 October 2019
2213-1582/ © 2019 Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

Fridriksson et al., 2018). If language networks vary substantially across different native languages, as some studies suggest (e.g., Ge et al., 2015; Wang et al., 2015; Wu et al., 2015), then prognostic models trained for one native language will not generalize to other languages. However, if brain networks are substantially similar across languages (e.g., Price, 2012; Rueckl et al., 2015), we can hope to be able to successfully apply prognostic models learned for language outcomes in one language, to other languages. We address this question by training our models on data from British, native English-speaking patients, and testing those models on data from Chilean, native Spanish-speaking patients.

The second potential restriction follows from differences in the pragmatic constraints, which govern neuroimaging data collection in research versus clinical environments. Specifically, research-quality brain scans are usually acquired with structural Magnetic Resonance Imaging (MRI), while clinical scans are often acquired with Computed Tomography (CT). MRI is useful for research purposes because it offers high resolution images with multiple contrasts emphasising different tissue properties. Moreover, so long as participants have no metal in their bodies (e.g. implantable medical devices like cardiac pacemakers; metallic shrapnel; or even certain kinds of tattoo; (Ross and Matava, 2011), there is no known risk of MRI to the human body. By contrast, CT is rapid, inexpensive and more readily available in many urgent care settings (Lev, 2003). CT can also be performed safely in patients with metal in their bodies, and the small risk associated with the radiation dose involved can often be balanced, favourably, against the immediate clinical needs of a patient suffering from acute neurological trauma. If our prognostic models cannot generalize from MRI to CT, we will struggle to exploit them in clinical practice. We address this risk by training our models with structural MRI data, and testing them with brain data acquired from CT.

Pragmatic differences between research- and clinically-oriented settings also drive the third potential restriction, which may be the most important of all. Research scans are typically acquired in the chronic phase (>6 months) post-stroke, on or around the same day as patients' language skills are assessed. These are 'chronic-chronic' data: i.e. both brain scans and language assessments are performed contemporaneously, in the chronic phase. Yet the most natural, target application for models learned from these data, is 'acute-chronic': i.e. we want to predict long-term language outcomes given scans acquired acutely, within a week or so after stroke onset. There is no guarantee that prognostic models learned from chronic-chronic data will generalize to acute-chronic data (Halai et al., 2018). This is because: (a) the evidence suggests that lesion data, derived from structural MRI, are crucial to credible predictions of language outcomes after stroke (Plowman et al., 2012); and (b) the same patient's lesions may appear very different in acute versus chronic scans.

Stroke lesions have been observed to grow over months or years after stroke onset (Naeser et al., 1998; Seghier et al., 2014). This growth has attracted relatively little attention, perhaps because it has not been consistently related to recovery: e.g. lesion growth does not appear to preclude effective recovery from aphasia (Naeser et al., 1998; Seghier et al., 2014). One explanation for this, is that this growth reflects structure catching up with function: i.e. the progressive atrophy of brain regions whose functional contribution to cognition has already been disrupted by stroke. But whatever the mechanism, this growth means that the same lesion may look different (bigger) when imaged 5 years post-stroke, than it did when imaged 6 months post-stroke. This is a potentially crucial confound when generalising models learned from chronic lesion data, to acute lesion data. Our design addresses this risk directly, by attempting to generalise prognostic models trained from chronic-chronic data collected for research purposes, to the more clinically significant acute-chronic prediction problem.

2. Materials and methods

2.1. British patients – training set

British patient data were drawn from the PLORAS database ((Seghier et al., 2016), which associates more than a thousand stroke patients with: (a) demographic data; (b) structural MRI; and (c) language task scores from the Comprehensive Aphasia Test (CAT; (Swinburn et al., 2004). Inclusion criteria to the PLORAS database comprise of: (i) a demonstrable previous medical history of stroke, (ii) no record of concomitant neurological or psychiatric illness (e.g. dementia or depression), and (iii) being able to provide written informed consent. Patients were excluded if they were non-native speakers of English, had lesions less than 800 mm^3 in volume (as assessed using our Automatic Lesion Identification (ALI) toolbox; (Seghier et al., 2008), or had not completed the 'spoken picture description' task of the CAT. The rationale for excluding patients whose lesions included no single, contiguous volume of at least 800 m^3 was two-fold: (1) this is the default threshold provided by the ALI toolbox; and (2) lower thresholds are more likely to identify lesions in tissue that a trained neurologist would classify as 'preserved'.

2.2. Chilean patients – test set

The clinical patient group used to validate our model was a sample of Chilean stroke patients who were recruited from the Physical Medicine and Rehabilitation Unit or the Neurology Unit of the Clinical Regional Hospital Dr. Guillermo Grant Benavente of Concepcion, Chile, after obtaining ethical approval. Each patient underwent a head CT, then on a later date their language skills were assessed with a Spanish translation of the CAT. The CAT was translated into Spanish and administered by two trained speech language therapists fluent in both Spanish and English (D.L.L.-P. and A.G.-V.). The inclusion and exclusion criteria were the same as in the training (British) set.

Four of the Chilean patients were scanned more than once. In 3/4 cases, both scans were acquired within the first months after stroke onset. In the fourth case, the first scan was at 11 months post-stroke, and the last scan was at 30 months post-stroke. In all cases, we chose the later scans, aiming to minimise the scan-assessment delays for the Chilean patients.

2.3. Language task

For the current study, patients were assessed only on the spoken picture description task of the CAT because it provides the means of obtaining a sample of connected speech with a reliable scoring system. In this task, the participant is shown a picture depicting a complex scene and prompted to verbally describe what is happening for 1 min. The speech sample is then rated based on the total number of appropriate information carrying words (i.e. words that convey exact meaning in the appropriate context and are correctly produced) minus the total number of inappropriate information carrying words (i.e. information carrying words that are incorrectly selected and/or produced), plus syntactic variety (on a 0–6 scale), grammatical well-formedness (on a 0–6 scale) and speed of speech production (on a 0–3 scale). There is no maximum or minimum score for this task, which we chose for the following three main reasons: (i) it is comparatively difficult, so patients with a range of impairments are often also impaired on this task; (ii) it is a reasonable proxy for natural language, because it requires patients to interpret a complex scene and report their interpretation in a coherent, free-form manner; and (iii) the scene depicted in this task reflects an everyday scenario encountered in both Chile and the UK, so should not be particularly sensitive to cultural differences between the two patient groups.

To ensure consistency in the scoring procedures for the British and Chilean patient samples, co-authors D.L.L.-P. and A.G.-V. thoroughly

followed the guidelines provided in the manual of the English version of the CAT, in addition to the more specific ones developed by the PLORAS recruitment team, when rating the speech samples from the Chilean stroke patients. Any scoring issues not directly addressed by these guidelines were extensively discussed between D.L.L.-P. and A.G.-V. until consensus was reached. Subsequently, the difficult-to-rate responses were analysed in conjunction with one of the speech and language therapists who scored the British data before a final score was assigned.

2.4. British imaging data

T1-weighted structural images were collected using four different MRI scanners (Siemens Healthcare, Erlangen, Germany): 388 patients were imaged on a 3 T Trio scanner, 244 on a 1.5 T Sonata scanner, 184 on a 1.5 T Avanto scanner, and 12 on a 3 T Allegra scanner. For the 1.5 T Avanto scanner, a 3D magnetization-prepared rapid acquisition gradient-echo (Mugler and Brookeman, 1990) sequence was used to acquire 176 sagittal slices with a matrix size of 256 × 224 yielding a final spatial resolution of 1 mm isotropic: repetition time/echo time/inversion time = 2730/3.57/1000 ms. For the other three scanners, an optimised 3D modified driven equilibrium Fourier transform sequence (Deichmann et al., 2004) was used to acquire 176 contiguous sagittal slices with an image matrix of 256 × 224 yielding a final resolution of 1 mm^3: repetition time/echo time/inversion time = 12.24/3.56/530 ms and 7.92/2.48/910 ms at 1.5 T and 3 T, respectively. Pre-processing was achieved using Statistical Parametric Mapping software (SPM12), each image was spatially normalized into Montreal Neurological Institute (MNI) and segmented into lesioned and healthy tissue using a unified normalization-segmentation algorithm (Ashburner and Friston, 2000) adapted for stroke patients (Seghier et al., 2008). This resulted in a binary lesion image for each patient registered in MNI space.

2.5. Chilean imaging data

CT scans were collected at the Clinical Regional Hospital Dr. Guillermo Grant Benavente of Concepcion, Chile, using a multidetector CT scanner (Aquilion 64; Toshiba America Medical Systems, Illinois, USA). The standard imaging protocol of the Imaging Unit involves the acquisition of a CT scan performed in the horizontal plane with the patient in supine position and holding his/her head stable to avoid motion artefacts during scanning. Overall, 200 axial slices were acquired per session and the scanning parameters were as follows: acquisition mode 32 × 0.5, acquisition volume 1 mm × 0.8 mm, 10-s scan time, 0.75-s rotation time, 120 kVp, 280 mA, HP 27 pitch. The whole brain volume was subsequently reprocessed in the operator console resulting in a total of 30 axial slices with a final resolution of 0.4 × 0.4 × 5 mm. For patients who were scanned more than once, we used each patient's latest scan as this was considered to be most representative of the lesion in the chronic phase.

Each image was normalized to MNI space using the SPM12 clinical toolbox (https://www.nitrc.org /projects/clinicaltbx) and then segmented into binary images of lesioned tissue using in-house software written in MATLAB based on the procedure outlined in Gillebert et al. (2014). 72 CT scans of neurologically intact controls that were collected as part of the Birmingham University Cognitive Screen (http://www.bucs.bham.ac.uk) were used to derive normative values for healthy brain tissue (Gillebert et al., 2014). Two semi-automated procedures were performed to improve the accuracy of this method: (a) the removal of false positive artefacts and (b) the expansion of binary image for underestimated lesions. For 10/59 scans, the conversion of binary lesion images from probability maps underestimated the extent of lesions. These were detected when visually inspecting the automatically identified lesions and finding that voxels surrounding the identified lesion still had abnormally low intensity values when compared with analogous healthy voxels in the contralateral hemisphere.

A local threshold was selected such that the boarder of the identified lesion extended to voxels that appeared to have similar intensity values to those of the healthy contralateral hemisphere. This had the effect of expanding the size of lesions that were previously being underestimated. Both of these semi-automated procedures were performed by the first author (R.L.) blind to the language scores of interest.

2.6. Lesion encoding and prognostic model predictors

In order to reduce the dimensionality of the imaging data, Principal Component Analysis (PCA) was performed on the lesion images. We simultaneously performed PCA reduction of the binary images (encoding the presence = 1 or absence = 0 of a lesion) for both British and Chilean patients, as such we had to ensure the binary images were of the same resolution. We used SPM's re-slice function on the Chile binary images to match to the voxel size of the British binary images (2 mm x 2 mm x 2 mm). We then concatenated all of the images into a single matrix (patients × voxels). As an effective mask we excluded all voxels with fewer than 10 patients having an identified lesion – both lesion identification algorithms only considered lesioned voxels within a whole brain mask. This resulted in 139,794 voxels. We then performed PCA on this matrix using singular value decomposition in Matlab2017a, without performing rotation of the resulting components. From the results, we selected components that individually explained at least 1% of the variance in the spatial distribution of lesions: this resulted in a total of 12 principal components, explaining a total of 62% of the variance. Along with lesion volume, these 12 components formed 13 lesion predictors. We also used 2 non-lesion predictors: age at stroke and time between stroke and assessment, for a total of 15 predictors for each model.

2.7. Analyses

2.7.1. Study 1: simple prediction

In this study, we used the British patient data to train a prognostic model, then used the model to predict the language scores of our separate sample of Chilean patients. Prognostic models were trained with a Gaussian Process regression model with default hyper-parameters, as specified for Matlab 2017a. These include (i) a squared exponential kernel function; (ii) an initial length scale = the mean of the standard deviations of the predictors (= 1, because all predictors were standardised prior to learning); and (iii) setting the signal standard deviation to the standard deviation of the responses divided by the square root of 2. We chose this inducer because it has recently been shown to perform well in this domain (Hope et al., 2018), and at least in our experience, procedures that employ more focused hyperparameter optimisation confer at best minor predictive advantages (in this domain) when performed in a properly nested manner.

To generate predictions within the British patient training set we performed 10-fold cross validation 100 times, taking the average over the 100 repetitions to be each patient's prediction. When generating predictions for the test set (Chilean patients) we used the whole training set to train the models: i.e. modelling the way we hope and expect such models might be employed in practice. We then compared the two groups by contrasting their prediction errors: significant group differences imply that our predictions are inconsistent across the two groups.

Our experience is that prediction errors are often correlated with empirical task scores in these analyses (e.g. (Hope et al., 2013): predictive models tend to over-estimate the lowest empirical scores, and under-estimate the highest empirical scores. This is consistent with the results actually observed for the British patients' predictions (see the Results sub-section 'Study 1'). Since there was also a significant difference between the patient groups' empirical task scores, this raises the possibility that apparent group differences in prediction errors may be explained by that difference in empirical task scores. To account for

this, we employed 'empirical task score' as a covariate of no interest in the comparisons. Our Frequentist test for group differences was a *t*-test on the coefficient of group as a categorical predictor of prediction error in a general linear model, which also included empirical task score as a scalar predictor. And we also employed a Bayesian ANOVA (Wetzels et al., 2012) to test for group differences, with patient group as a single factor with two levels (British versus Chilean), and the dependent variable defined as the residuals of prediction errors after regressing out empirical task scores. The Matlab script used to run this comparison (which merely implements the procedure described by its authors), is available at: [https://github.com/robloughnan/PostStrokeLanguagePrediction_NeuroimageClinical].

2.7.2. Lesion growth analysis

On the basis of our prior work (Seghier et al., 2014), we hypothesized that the most significant confound between these two groups might follow from the delay between scan acquisition and language assessment. For most of the British patients, this delay was small (92% < 1 week), whereas it was typically large (all >1 month; 33/59 > 6 months) for the Chilean patients. Since it has been shown that lesions appear to grow gradually over years post-stroke (Seghier et al., 2014), we decided to control for the confound by modelling that growth directly: using the longitudinal imaging data from the British patients to learn a series of lesion growth functions (1 for each lesion principal component and another for total lesion volume), and using the learned functions to project each Chilean patient's lesion to the date of their language assessment.

Our training data consisted of 142 English-speaking patients for whom we had repeated longitudinal MRI scans. The growth functions were learned with a linear inducer, with time post-stroke log-transformed to ensure a power relationship between lesion growth and time post-stroke, as previously reported (Seghier et al., 2014). Each observation in the training data included all of the 'lesion-predictors' (i.e. 12 principal components plus lesion volume), together with: (a) the time post-stroke at which they were first assessed; and (b) the time post-stroke at which the last scan was acquired for that same patient. The response variable for each model was the lesion variable, acquired at that second time-point, that we wanted the model to predict (i.e. one of the 12 principal components, or lesion volume). We trained a total of 13 separate models, to predict each of the 13 lesion-variables from these training data. Lesion projection involved applying all 13 models to project the lesion-predictors forward/backward to the desired times post-stroke. The Matlab scripts used to implement this procedure are available at [https://github.com/robloughnan/PostStrokeLanguagePrediction_NeuroimageClinical]. Armed with those projected lesions, we could then employ the same group comparison analysis as described previously, hoping to see more consistent prediction errors across the two patient groups.

2.7.3. Validating the lesion growth functions

Our hypothesis, on observing the apparently consistent (positive) bias in our predictions from Study 1, was that the bias followed from group differences in the delays between scan acquisition and language assessment. Specifically, our British patients' scans and assessments were usually conducted contemporaneously (92% within a week), whereas the Chilean patients' language assessments were usually conducted much later than their scans. Models learned from contemporaneous data may be expected to learn contemporaneous lesion-symptom associations: i.e. associations between a given lesion and a specified language score 'now'. By projecting the Chilean patients' lesions to the dates of their language assessments, in Study 2, we hoped to mitigate the group difference, improving the predictions. The success of study 2 begged a further question, of whether we could improve other predictions using the same approach.

To answer this question, we ran a further study (Study 3), selecting: (a) a training sample of British patients that might maximise any implicit bias in favour of contemporaneous lesion-symptom associations (i.e. with small scan-assessment delays); and (b) a testing sample that might maximally illustrate the effect of that bias (i.e. with large scan-assessment delays). We predicted that confounds introduced by scan-assessment delays should plateau with longer delays because the rate of growth decreases with increasing time post-stroke (Seghier et al., 2014). In turn, this implies that the impact of lesion-projection, used to minimise those confounds, should be more subtle in more chronic samples (i.e. the British patients) than in more acute samples (i.e. the Chilean patients). Our subsample selection, in Study 3, aimed to account for this more subtle signal. Our subsample selection, in Study 3, aimed to maximise this more subtle signal. In both Studies 2 and 3, we employed a one-tailed Wilcoxon Signed Rank test on absolute prediction errors for pre- versus post-lesion growth predictions (smaller absolute errors imply better predictions).

3. Results

3.1. Patient samples

Our training set included 828 British stroke patients (251 women; age at stroke onset median (interquartile range or IQR) = 57.2 (18.2) years; months between stroke and behavioural assessment median (IQR) = 34.1 (57.2); days between scan and behavioural assessment median (5–95 percentile range) = 0 (5), and our test set included 59 Chilean stroke patients (26 women; age at stroke onset median (IQR) = 58.2 (15.2) years; months between stroke and behavioural assessment median (IQR) = 8 (14.8); days between scan and behavioural assessment median (5–95 percentile range) = 217 (968.4)). 346 British patients were assessed as suffering from significant, aphasic impairment in the language task: i.e. their scores would put them in the bottom 5% relative to a distribution of scores in the same task, from controls (Swinburn et al., 2004). Using the same impairment threshold, 46 Chilean patients would be assessed as impaired; however, this judgment can only properly be made by reference to a Chilean control

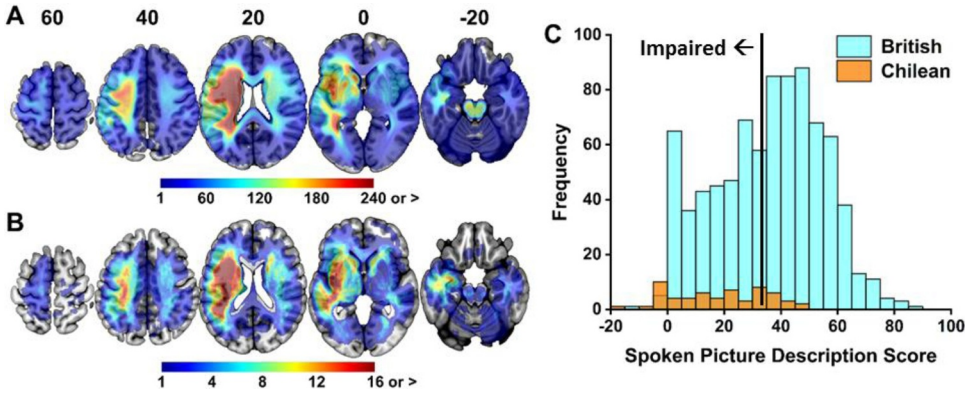

Fig. 1. Lesion overlap maps and histograms of spoken picture description scores. Lesion overlap maps for (A) British and (B) Chilean datasets are shown in axial slices (numbers above indicate z co-ordinates in MNI space; the left side of each slice is the left side of the brain). The colour scale represent the number of patients with overlapping lesions at each given voxel. (C) Histogram of Spoken Picture Description task scores for British and Chilean datasets. The Chilean patients' scores were significantly lower than the British patients' scores (independent sample *t*-test: $t(885) = 6.64$, $p < .001$) because the British sample included patients who did not have language impairments at test.

distribution of scores. Work to acquire these scores is ongoing.

Lesion frequency images for each patient group are displayed in Fig. 1A and B respectively. The median (IQR) lesion volume for the Chilean patients was 4541 (8242) voxels, and for the British patients was 5267 (12,567) voxels. Notably, we had longitudinal scan and assessment data for 142/828 British patients (months from stroke to first visit was median (IQR) = 34.9 (54.0); months between visits was median (IQR) = 19.7 (35.5)). Only these patients' first scans contribute to the lesion frequency images in Fig. 1B. Fig. 1C displays the empirical task (spoken picture description) scores for both patient groups.

Since our main interest here was in model generalization rather than group comparisons, we did not attempt to match these patient groups in a case-controlled manner. Nevertheless, two potentially important differences, between the patient groups, were observed. First, the Chilean patients had lower scores, as a group, than the British patients (see Fig. 1). Second, as expected, most of the British patients' scans were acquired contemporaneously with their language assessments (92% <1 week), whereas the delay between scan and assessment was longer (all >1 month; 33/59 >6 months) for the Chilean patients.

3.2. Study 1: simple prediction

Predicted task scores were calculated for the British patients via 100 times 10-fold cross-validation. For the Chilean patients, these predictions were made by a single model, trained on the whole of the British patient dataset. Predictions for both groups are displayed in Fig. 2. Correlations between predicted and empirical task scores were strong in both groups (British: $r = 0.71$, $p < .001$ Chilean: $r = 0.68$, $p < .001$), and not significantly different across the groups (Fisher r-to-z transform; $z = 0.42$, $p = .67$). As expected (see the Methods), prediction errors for the British patients were correlated with empirical task scores ($r = 0.71$, $p < .001$), with lower empirical scores tending to be over-estimated, and higher empirical scores tending to be under-estimated. Consistent with this, the predictions for the Chilean patients, whose empirical scores tended to be lower, are also generally positively biased (i.e. higher than the empirical task scores). Nevertheless, the group difference in prediction errors was significant even when differences in empirical task scores were taken into account: (general linear model with prediction error as the dependent variable: empirical score: $t(844) = 30.1$, $p < .001$; patient group: $t(884) = -4.21$, $p < .001$).

In other words, while our prognostic model distinguishes better outcomes from worse outcomes equally well across the two patient groups, there is nevertheless evidence of *inconsistency* in the predictions made across those groups. We can quantify this in Bayesian terms with a Bayesian ANOVA (Wetzels et al., 2012), with patient group as a single factor with two levels (British or Chilean), and the residuals of the prediction errors after regressing out empirical task score, as the dependent variable. The result indicates that there is modest evidence in favour of inconsistency across the patient groups: Bayes Factor = 4.9, so nearly five times as much evidence for inconsistency as for consistency.

3.3. Study 2: prediction accounting for scan-assessment delays

As mentioned previously, our two patient groups are significantly different in terms of the delays patients experienced between scan acquisition and language assessment: the test set is acute-chronic, whereas the training set is chronic-chronic. One potentially confounding difference between these two scenarios, follows from evidence that lesions tend to grow gradually over time post-stroke (Seghier et al., 2014). Larger lesions are typically associated with worse language outcomes (Plowman et al., 2012), and we would expect the same lesion to appear larger, in the chronic phase post-stroke, than it did in the acute phase. When generalizing lesion-deficit associations learned from chronic scans to acute scans, we therefore run the risk of making over-optimistic predictions.

To mitigate this confound, we learned 'lesion growth' functions from the longitudinal neuroimaging data available for 142/828 of the British patients. Separate functions were learned for each of the (12) principal components that were employed to encode patients' lesions in Study 1, and also for lesion volume, for a total of 13 models (see the Methods for more details). In internal cross-validation, all but one of these functions yielded correlations between predicted and empirical responses (principal components and lesion volume variables) >0.9; the exception was the model for component 5, where the correlation coefficient was 0.71. In each case, the standard deviations of the prediction error distributions was <10% of the range of the response variable. In other words, there were consistent signals to find in the British patients' longitudinal neuroimaging data. In line with prior reports (from an overlapping sample; (Seghier et al., 2014), the median rate of growth in lesion volume was ~4.5 cm^3 (~7.5%) per year, though growth also slowed with increasing time post-stroke: the correlation between log(time post-stroke at first scan) and subsequent growth rate = −0.38, $p < .001$.

Having learned these functions, we used them to project the Chilean patients' lesions forward to the times post-stroke at which their speech skills were assessed. Next, we used the same prognostic model as employed in Study 1 to predict the Chilean patients' task scores (again). The correlation coefficient between empirical and predicted scores, after lesion growth, was 0.63 ($p < .001$). Fig. 3A shows how the predictions differed before and after lesion projection. Most patients' predictions move closer to the diagonal line representing zero prediction

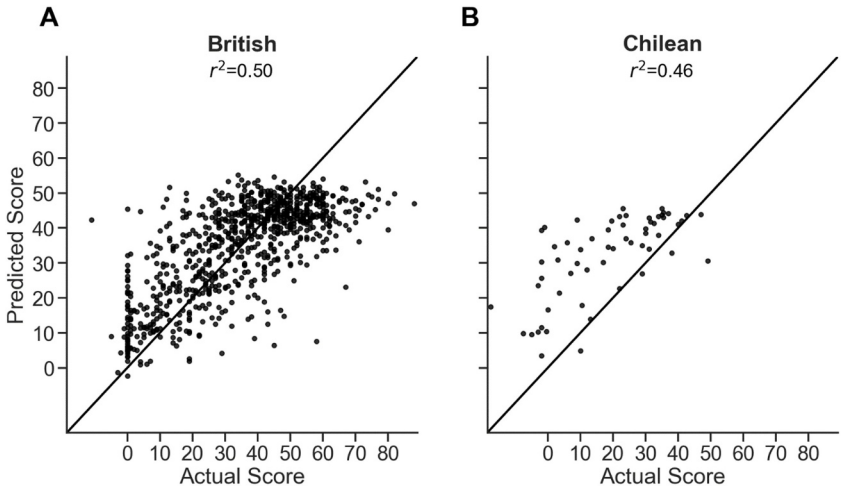

Fig. 2. Basic prediction results. Empirical vs. Predicted scores for the British (A) and Chilean (B) datasets; in each case, perfect predictions would lie on the black diagonal lines. It can be seen that most of the points for the Chilean dataset lie above the black line indicating the model is predicting patients to perform better than they actually do. Mean (standard deviation) prediction errors: British = 0.02 (13.44); Chilean = 14 (11.93).

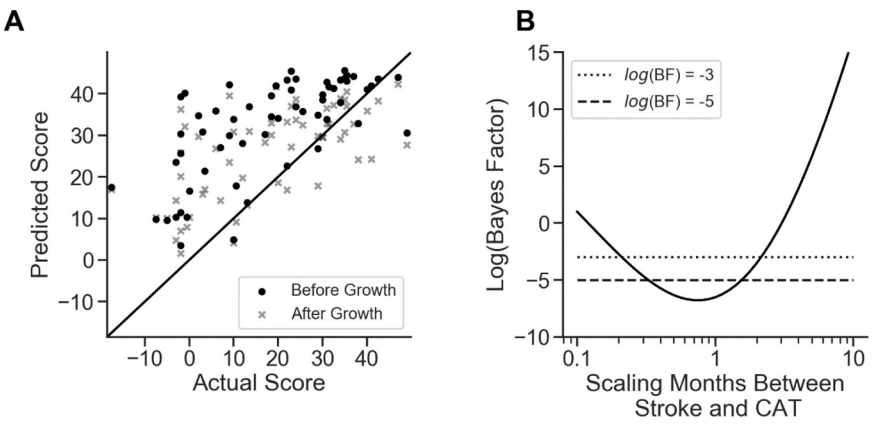

Fig. 3. The effects of lesion projection. (A) Predictions of task scores for Chilean patients before (black dots) and after (grey crosses) growing lesions to the same day as the behavioural assessment. Solid black line indicates line of perfect predictions. The mean (standard deviation) prediction error after growth was 7.87(12.53). (B) The effect of less or more lesion growth (than empirically required), on the consistency of the prediction errors across the groups. More negative log Bayes Factors indicate stronger support for consistency; this support is moderately strong below the dotted line, and very strong below the dashed line (Jeffreys, 1961).

error, and this represents a significant improvement in prediction accuracy (i.e. reduction in prediction error): one tailed Wilcoxon Signed Rank test on absolute prediction errors with vs without lesion growing ($Z = 4.62, p < .001$). Moreover, the lesion projection also extinguishes the group difference identified in Study 1: i.e. 'patient group' is no longer significantly related to speech score prediction errors, once 'empirical task score' is taken into account (empirical score: $t(884) = 30.6, p < .001$; patient group: $t(884) = 0.91, p = .36$). Using the same Bayesian ANOVA as employed in Study 1, the resulting Bayes Factor = 0.0017: i.e. the evidence for *consistency* is nearly 600 times (1/0.0017 = 596) stronger than for inconsistency here.

As a further check on those lesion-growth models, we considered what the effect might have been of growing the Chilean patients' lesions either less or more than was actually required to traverse their empirical delays between scan acquisition and language assessment. Considering lesion growth over a range from 10% to 1000% of those empirical delays, we repeated all of the analyses just described: i.e. growing the patients' lesions by the specified amount; using the prognostic model from Study 1 to make new predictions; regressing empirical task score out of the prediction errors, and using our Bayesian ANOVA (Wetzels et al., 2012) to measure any residual group effects. The results are displayed in Fig. 3B: the amount of lesion growth required to traverse the empirical delays between scan and behavioural assessment (x-axis = 1) is close to the optimum amount of growth required to maximise consistency across the patient groups (i.e. to minimise log Bayes Factor for the comparison). In other words, our lesion growth models exhibit an encouraging degree of specificity: they make the patient groups consistent only when used to 'grow' the Chilean patients' lesions to roughly the right extent.

3.4. Study 3: external validation of the lesion growth models

Study 2 demonstrates that we can improve the Chilean patients' predictions, significantly, by projecting their lesions to the dates of their language assessments. As further confirmation that the lesion projection models apply generally, rather than specifically to this particular patient sample, we asked whether we could use the same lesion-growth functions to improve our predictions for those British patients whose delays between scan acquisition and language assessment were also unusually long.

We ran this analysis by partitioning the British patients based on their own delays between scan acquisition and language assessment: distinguishing a training sample, whose delays were all less than a week (759 patients), from a test sample, whose delays were more than a year (17 patients). We then ran the same analyses as in Studies 1 and 2: i.e. training a model with the 759 patients; predicting spoken picture description scores for the test sample of 17 patients; projecting the 17 patients' lesions to the dates of their language assessments; and making new predictions for the same 17 patients. Critically, in this case we are projecting lesions identified automatically from MRI, rather than lesions identified semi-automatically from CT (as described in the Methods), as in Study 2. We are also projecting most of the British patients' lesions *backwards* in time post-stroke, rather than forwards, as in Study 2.

Nevertheless, the lesion projection process still significantly improves our predictions: Wilcoxon Signed Rank Test on the absolute prediction errors; $z = 1.75, p = .03$; see Fig. 4. This is evidence that our lesion growth models may apply generally, rather than specifically to the Chilean patient sample who first motivated their use.

4. Discussion

We began by asking whether prognostic models trained with stroke patient data could generalize to patients from another country (UK vs. Chile), with a different native language (English vs. Spanish), who were scanned at different times post-stroke (mainly chronic for British patients, and mainly acute for the Chilean patients), with different neuroimaging technology (MRI vs. CT), collected for different purposes (research vs. clinical). Perhaps surprisingly, given all of these differences between the two groups, our results suggest that models trained on one group can indeed generalize to the other.

Even before we modelled lesion growth, our prognostic models were distinguishing better from worse language outcomes approximately equally well in both datasets; correlations between predicted and

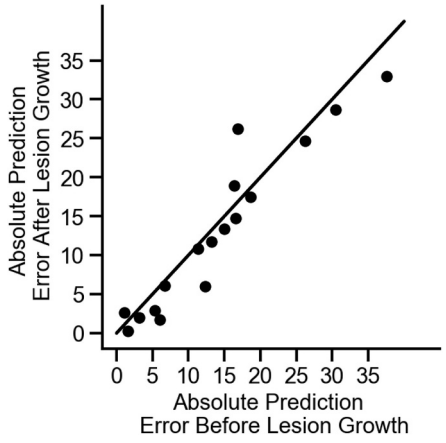

Fig. 4. British patient absolute prediction error, after growth vs. before growth: Prediction errors are compared for British imaging data, with and without projection (i.e. before and after lesion growth) to be contemporaneous with their language scores. Any point which lies on the solid black line represents a patient whose predictions did not change with projection, points below the black line indicate patients whose predictions were more accurate after lesions were projected, through time, to be contemporaneous.

empirical task scores: $r = 0.71$ and $r = 0.68$ for the British and Chilean patients, respectively. This suggests that the language networks recruited by (neurologically intact) English and Spanish speakers are substantially similar.

Nevertheless, there was moderate evidence *against* consistency across the two datasets in those predictions (Bayes Factor = 4.9). Our hypothesis was that this occurred due to differences in the delay between scan acquisition and language assessments across the two groups, and our results support that hypothesis. Using longitudinal neuroimaging data from the British patients, we learned a set of 'lesion growth functions', and used them to project the Chilean patients' lesions forward to the dates of their language assessments. Having grown the lesions, there is 'very strong' (Jeffreys, 1961) evidence for consistency (Bayes Factor ~600) in the predictions made across the two patient groups – even though only the British patient sample was used to train the prognostic model.

Our lesion growth models are still a work in progress; we are not claiming to have found the best or right way to model this change. Nevertheless, our results do suggest both that the change is worth capturing, and that our models do capture it to some extent. First, the models appeared to predict lesion growth relatively well within the longitudinal neuroimaging dataset that we used to train them; this implies that there is a systematic signal here to be found. Second – our main result – we could use those lesion growth models to demonstrate consistency between the predictions for our British and Chilean patients' language task scores. Crucially, the lesion growth models were also at least somewhat specific: the amount of growth required to traverse the Chilean patients' empirical delays between scan and assessment is close to the amount of growth which maximised consistency across the groups (see Fig. 3B). And finally, we could use the same growth models to significantly improve our prognoses for those British patients whose own delays between scan and language assessment were unusually long (Fig. 4). This suggests that our lesion growth models might be capturing a real, generally applicable feature of lesion-symptom data.

Past reports of gradual lesion growth, over time post-stroke, have emphasised that stroke survivors' behavioural skills often improve regardless of that growth (Naeser et al., 1998; Seghier et al., 2014). Growing lesions do not imply worsening behavioural skills, which suggests that recovery can recruit neural tissue far from the periphery of the original lesion cavity, for example in the preserved right hemisphere after left hemisphere stroke (Forkel et al., 2014; Hope et al., 2017; Xing et al., 2015). To the extent that this is true, it begs the question of why our lesion growth models result in improved predictions. We suggest that this is because lesion growth is implicitly captured by the prognostic models that we learned from the British patients, because most of those patients' scans and language assessments were conducted contemporaneously (92% within 1 week), and because time post-stroke is a consistent feature of our models (Hope et al., 2015a, 2018, 2013).

Quite what drives these changes, remains an open question, to which there may be more than one answer. We have learned lesion growth functions from MRI scans taken years after stroke onset and applied them, apparently successfully, to CT scans taken hours or days after stroke. This suggests that we are not just capturing some longitudinal artefact of any particular neuroimaging technology. However while our growth functions are continuous, this need not imply that there is a single, continuous neurophysiological process which drives the modelled change. Two of the best studied mechanisms of lesion change – the necrosis of initially preserved but still hypoperfused neural tissue, and the replacement of that necrotic tissue by a fluid-filled cavity ('cavitation') – may well be important in the acute phase (Rekik et al., 2012), but probably cannot explain change over years. In perhaps the first systematic study of this longer term change, Naeser and colleagues (Naeser et al., 1998) asserted that the mechanisms that drive it 'are not understood', but speculated that it may be caused by: (a) Wallerian degeneration (along white matter tracts connected to damaged tissue); (b) chronic hypoperfusion of the tissue adjacent to the lesion, leading to gradual neural atrophy; and/or (c) 'other small vessel involvement', the notion that chronic hypoperfusion may occlude other small vessels, spreading its effect beyond perilesional tissue. As far as we know, no more definite answers to this question have since been proposed.

Another open question follows from the practice, which is nearly ubiquitous in lesion-behavior mapping, of reporting analyses of groups of patients who are very variable in time post-stroke (Butler et al., 2014; Del Gaizo et al., 2017; Forkel et al., 2014; Fridriksson et al., 2018; Pani et al., 2016; Pustina et al., 2017; Tak and Jang, 2014; Xing et al., 2016; Yourganov et al., 2016; Zhang et al., 2014). Our results suggest that lesion change over time post-stroke may confound lesion-similarity analyses, which are implicit in most (mass univariate and multivariate) lesion-behavior mapping methods. Indeed, when considered together with emerging evidence that real behavioural change in aphasia also continues over years after stroke onset (Holland et al., 2017; Hope et al., 2017), the implication is that lesion-behavior mapping studies of aphasia after stroke may better be interpreted as snapshots of a dynamic system, rather than as incrementally more accurate approximations to some fixed ground truth (Del Gaizo et al., 2017; Fridriksson et al., 2018; Pustina et al., 2018; Yourganov et al., 2016). Quite how dynamic these systems really are, and as a consequence how significant are any confounds that this change may introduce, remains to be seen; these questions can best be answered with longitudinal studies.

One caveat to these results is that, despite their apparent specificity, our lesion growth models might exert only an accidental effect: i.e. rectifying a predictive bias that actually emerges for a different reason. For example, the Chilean and British patients' lesions were segmented from different imaging modalities (CT vs. MRI), with different methods; we took deliberate care to ensure that the Chilean patients' lesions were not under-estimated (as described in the Methods), but we cannot completely rule out the potential for residual bias. Similarly, the Chilean patients' language skills naturally could not be assessed in exactly the same way as the British patients' skills, because the two groups spoke different native languages. Again, we devoted significant effort to mitigating this confound, as described in the Methods: i.e. focusing on a task involving stimuli that are common in both the UK and Chile, and using two clinically trained, native Spanish speaking (and fluent English speaking) speech and language therapists to implement the scoring. But despite our best efforts, it is still conceivable that our language assessments were simply more difficult, for the Chilean patients, than they were for the British patients. Notably, the Chilean patients were referred to us by the staff who were delivering ongoing treatment to them, which in itself will likely bias the sample towards those with significant, enduring impairments (i.e. which still need treatment).

Our analyses suggest that differences in prediction errors across the two groups are not simply a function of group-differences in aphasic symptom severity at test time. This is because: (a) we control for those differences, explicitly, in our analyses; and (b) our lesion growth functions appear to explain those residual differences well (Study 2), and also to improve the predictions we make in independent data (Study 3). It remains to be seen whether they will perform as expected in further validation studies. One test of that broader performance, would be to repeat the experiments reported here for scores in a wider variety of language tasks, emphasising different language skills; this work is ongoing.

Similarly, the results presented here do not guarantee that our models will generalize across every language. For example, it seems reasonable to expect that cross-language variability in functional anatomy might be less significant for Spanish versus English, than for Chinese versus English. Notably, that latter comparison drives most reports of language-specific functional architecture in the brain (e.g., Ge et al., 2015; Wang et al., 2015; Wu et al., 2015). In the same vein, it

is possible that our models are not sensitive to these differences because they encode lesion location information in a relatively coarse way, as principal components. This kind of coarse encoding has been shown to be approximately as effective as more detailed encodings in lesion-deficit models (Ramsey et al., 2017), but there are also reports that much higher resolution encodings can be beneficial (Rondina et al., 2016). If the latter is true, then enhanced, single-language prognostic performance might depend on increased sensitivity to subtle differences in premorbid language networks, which may in turn reduce cross-language generalization.

More broadly, we cannot really know how widely these prognostic models will generalize, unless and until we test them more widely: i.e. with more and larger independent datasets, capturing more of the true variability of the stroke survivor population as a whole. First, our training and testing results were based on, and thereby limited to, patients whose lesions were larger than 800 mm^3. This has been our default criterion in many past studies (Hope et al., 2015a, 2017, 2018, 2013, 2015b), and the same applies to Chilean patients. Nevertheless, setting a minimum lesion volume threshold of 800 mm^3 resulted in the exclusion of approximately 7% of the patients (147) in the PLORAS database, from which the British patients were drawn for this study. Of note, 27 of these patients exhibited at least some language impairment at test time, which indicates the exclusion does leave some examples of aphasic performance unexplained. In addition, the requirement to provide written informed consent, in our study as in almost all others, will tend to impose a selection bias on research samples, because older and/or more severely impaired patients are less likely to grant it (or be able to grant it). Quite how to best to circumvent or minimise this barrier to clinical investigation in this domain, remains an open question.

Our questions, in this work, were: (a) whether models learned from research 'chronic-chronic' data could be applied to clinical 'acute-chronic' data; and (b) if not, why not? These questions are important because most research studies in this area are expressly motivated by the claim that their results are, or might be, clinically relevant. Having identified an apparently systematic bias in the predictions made, for clinical data, by models trained with research data, we sought to explain that bias as a function of lesion growth over time post-stroke. That we can, apparently, mitigate the bias by modelling lesion growth, is evidence that the bias is driven by lesion growth. But while the correction appears effective, and might well be improved over time, a correctable bias is still a bias. We are not suggesting that training with research scans, via lesion projection, will or should supplant training clinically applicable models with clinical acute scans. Nevertheless, we hope that our results will encourage further study of whether and how lesion growth might confound lesion-symptom analyses, and the development of better models of how, why, and what lesion growth occurs over years after stroke.

Data analysis code

Scripts used in this analysis are available at: https://github.com/robloughnan/PostStrokeLanguagePrediction_NeuroimageClinical

Acknowledgments

For the British component of the project, we thank Tom Schofield, Jenny Crinion, Sue Ramsden and Andre Selmer for setting up the patient database; Louise Lim, Rachel Bruce and Zula Haigh for setting up the language testing procedures; and the PLORAS recruitment team for their help collecting the data.

For the Chilean component of the project, we thank the Heads of the Physical Medicine and Rehabilitation Unit (Dra. Roxana Hurtado Hofer), Neurology Unit (Dr. Patricio Soto Jara), and Imaging Unit (Dr. Martin Einersen Artigues) of the Clinical Regional Hospital Dr. Guillermo Grant Benavente of Concepcion; the Chair of the Local Research Ethics Committee (Dra. Maria Antonia Bidegain Santolalla); the Dean of the Faculty of Medicine of Universidad de Concepcion (Dr. Raul Gonzalez Ramos); the Chief Radiologist of the Imaging Unit (Mr. Pablo Bustos Contreras) who provided the neuroimaging data; and the Chief Nurse of the Physical Medicine and Rehabilitation Unit (Ms. Nicolle Luna Escalona) who oversaw the recruitment process.

We also thank the Birmingham University Cognitive Screen recruitment team for their help with the CT data collection in neurologically intact patients.

Last but not least, we would like to thank the stroke survivors (and their relatives) who participated in the study.

Funding: This study was supported by the Medical Research Council (MR/M023672/1), Wellcome (091593/Z/10/Z, 205103/Z/16/Z and 101253/A/13/Z), and the Stroke Association (TSA PDF 2017/02; TSA 2014/02). Co-authors 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. The funders had no participation in the design and results of this study.

Author contributions: All authors conceived the study. C.J.P. acquired the funding to collect both datasets, and to ensure anonymized access to them. D.L.L.-P., A.G.-V. and V.E.-V. set up the Chilean component of the project. T.M.H.H. and R.L. designed the statistical analysis methods, and R.L. implemented those methods. T.M.H.H., R.L. and C.J.P. interpreted the data. T.M.H.H. and R.L. led the writing of the manuscript, though all authors contributed to this.

Data and materials availability: Analysis code are available on request from T.M.H.H.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.nicl.2019.102005.

References

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry—the methods. Neuroimage 11, 805–821.

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.

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.

Del Gaizo, J., Fridriksson, J., Yourganov, G., Hillis, A.E., Hickok, G., Misic, B., Rorden, C., Bonilha, L., 2017. Mapping language networks using the structural and dynamic brain connectomes. eNeuro 4 ENEURO.0204-0217.2017.

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, 2027–2039.

Fridriksson, J., den Ouden, D.-B., Hillis, A.E., Hickok, G., Rorden, C., Basilakos, A., Yourganov, G., Bonilha, L., 2018. Anatomy of aphasia revisited. Brain.

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.

Ge, J., Peng, G., Lyu, B., Wang, Y., Zhuo, Y., Niu, Z., Tan, L.H., Leff, A.P., Gao, J.H., 2015. Cross-language differences in the brain network subserving intelligible speech. Proc Natl. Acad. Sci. U S A 112, 2972–2977.

Gillebert, C.R., Humphreys, G.W., Mantini, D., 2014. Automated delineation of stroke lesions using brain CT images. NeuroImage 4, 540–548.

Halai, A.D., Woollams, A.M., Lambon Ralph, M.A., 2018. Predicting the pattern and severity of chronic post-stroke language deficits from functionally-partitioned structural lesions. NeuroImage Clin. 19, 1–13.

Holland, A., Fromm, D., Forbes, M., MacWhinney, B., 2017. Long-term recovery in stroke accompanied by aphasia: a reconsideration. Aphasiology 31, 152–165.

Hope, T.M., Parker, J., Grogan, A., Crinion, J., Rae, J., Ruffle, L., Leff, A.P., Seghier, M.L., Price, C.J., Green, D.W., 2015a. Comparing language outcomes in monolingual and bilingual stroke patients. Brain 138, 1070–1083.

Hope, T.M.H., Leff, A.P., Prejawa, S., Bruce, R., Haigh, Z., Lim, L., Ramsden, S., Oberhuber, M., Ludersdorfer, P., Crinion, J., Seghier, M.L., Price, C.J., 2017. Right hemisphere structural adaptation and changing language skills years after left hemisphere stroke. Brain.

Hope, T.M.H., Leff, A.P., Price, C.J., 2018. Predicting language outcomes after stroke: is structural disconnection a useful predictor? NeuroImage 19, 22–29.

Hope, T.M.H., Seghier, M.L., Leff, A.P., Price, C.J., 2013. Predicting outcome and recovery after stroke with lesions extracted from MRI images. NeuroImage 2, 424–433.
Hope, T.M.H., Seghier, M.L., Prejawa, S., Leff, A.P., Price, C.J., 2015b. Distinguishing the effect of lesion load from tract disconnection in the arcuate and uncinate fasciculi. Neuroimage.
Jeffreys, H., 1961. Theory of Probability. Oxford University Press, Oxford, UK.
Lev, M.H., 2003. CT versus MR for acute stroke imaging: is the "obvious" choice necessarily the correct one? Am. J. Neuroradiol. 24, 1930–1931.
Mah, Y.-H., Husain, M., Rees, G., Nachev, P., 2014. Human brain lesion-deficit inference remapped. Brain 137 (Pt 9), 2522–2531.
Mozaffarian, D., Benjamin, E.J., Go, A.S., Arnett, D.K., Blaha, M.J., Cushman, M., de Ferranti, S., Despres, J.P., Fullerton, H.J., Howard, V.J., Huffman, M.D., Judd, S.E., Kissela, B.M., Lackland, D.T., Lichtman, J.H., Lisabeth, L.D., Liu, S., Mackey, R.H., Matchar, D.B., McGuire, D.K., Mohler 3rd, E.R., Moy, C.S., Muntner, P., Mussolino, M.E., Nasir, K., Neumar, R.W., Nichol, G., Palaniappan, L., Pandey, D.K., Reeves, M.J., Rodriguez, C.J., Sorlie, P.D., Stein, J., Towfighi, A., Turan, T.N., Virani, S.S., Willey, J.Z., Woo, D., Yeh, R.W., Turner, M.B., 2015. Heart disease and stroke statistics–2015 update: a report from the American Heart Association. Circulation 131, e29–322.
Mugler, J.P., Brookeman, J.R., 1990. Three-dimensional magnetization-prepared rapid gradient-echo imaging (3D MP RAGE). Magn. Reson. Med. 15, 152–157.
Nachev, P., 2015. The first step in modern lesion-deficit analysis. Brain 138 e354-e354.
Naeser, M.A., Palumbo, C.L., Prete, M.N., Fitzpatrick, P.M., Mimura, M., Samaraweera, R., Albert, M.L., 1998. Visible changes in lesion borders on CT scan after five years poststroke, and long-term recovery in aphasia. Brain Lang 62, 1–28.
Pani, E., Zheng, X., Wang, J., Norton, A., Schlaug, G., 2016. Right hemisphere structures predict poststroke speech fluency. Neurology.
Plowman, E., Hentz, B., Ellis Jr., C., 2012. Post-stroke aphasia prognosis: a review of patient-related and stroke-related factors. J. Eval. Clin. Pract. 18, 689–694.
Price, C.J., 2012. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. Neuroimage 62, 816–847.
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 115, 154–166.
Pustina, D., Coslett, H.B., Ungar, L., Faseyitan, O.K., Medaglia, J.D., Avants, B., Schwartz, M.F., 2017. Enhanced estimations of post-stroke aphasia severity using stacked multimodal predictions. Hum. Brain Mapp. 38, 5603–5615.
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.
Rekik, I., Allassonnière, S., Carpenter, T.K., Wardlaw, J.M., 2012. Medical image analysis methods in MR/CT-imaged acute-subacute ischemic stroke lesion: segmentation, prediction and insights into dynamic evolution simulation models. A critical appraisal. NeuroImage: Clin. 1, 164–178.
Rondina, J.M., Filippone, M., Girolami, M., Ward, N.S., 2016. Decoding post-stroke motor function from structural brain imaging. NeuroImage Clin. 12, 372–380.
Ross, J.R., Matava, M.J., 2011. Tattoo-induced skin "burn" during magnetic resonance imaging in a professional football player: a case report. Sports Health 3, 431–434.
Rueckl, J.G., Paz-Alonso, P.M., Molfese, P.J., Kuo, W.-J., Bick, A., Frost, S.J., Hancock, R., Wu, D.H., Mencl, W.E., Duñabeitia, J.A., Lee, J.-R., Oliver, M., Zevin, J.D., Hoeft, F., Carreiras, M., Tzeng, O.J.L., Pugh, K.R., Frost, R., 2015. Universal brain signature of proficient reading: evidence from four contrasting languages. Proc. Natl. Acad. Sci. U S A 112, 15510–15515.
Seghier, M.L., Patel, E., Prejawa, S., Ramsden, S., Selmer, A., Lim, L., Browne, R., Rae, J., Haigh, Z., Ezekiel, D., Hope, T.M.H., Leff, A.P., Price, C.J., 2016. The PLORAS database: a data repository for predicting language outcome and recovery after stroke. Neuroimage 124 (Part B), 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.
Seghier, M.L., Ramsden, S., Lim, L., Leff, A.P., Price, C.J., 2014. Gradual lesion expansion and brain shrinkage years after stroke. Stroke 45, 877–879.
Swinburn, K., Porter, G., Howard, D., 2004. Comprehensive Aphasia Test. Psychology Press.
Tak, H.J., Jang, S.H., 2014. Relation between aphasia and arcuate fasciculus in chronic stroke patients. BMC Neurol. 14, 46.
Wang, X., Yang, J., Yang, J., Mencl, W.E., Shu, H., Zevin, J.D., 2015. Language differences in the brain network for reading in naturalistic story reading and lexical decision. PLoS ONE 10, e0124388.
Watila, M.M., Balarabe, S.A., 2015. Factors predicting post-stroke aphasia recovery. J. Neurol. Sci. 352, 12–18.
Wetzels, R., Grasman, R.P.P.P., Wagenmakers, E.-J., 2012. A default bayesian hypothesis test for ANOVA designs. Am. Stat. 66, 104–111.
Wu, J., Lu, J., Zhang, H., Zhang, J., Yao, C., Zhuang, D., Qiu, T., Guo, Q., Hu, X., Mao, Y., Zhou, L., 2015. Direct evidence from intraoperative electrocortical stimulation indicates shared and distinct speech production center between Chinese and English languages. Hum. Brain Mapp. 36, 4972–4985.
Xing, S., Lacey, E.H., Skipper-Kallal, L.M., Jiang, X., Harris-Love, M.L., Zeng, J., Turkeltaub, P.E., 2015. Right hemisphere grey matter structure and language outcomes in chronic left hemisphere stroke. Brain.
Xing, S., Lacey, E.H., Skipper-Kallal, L.M., Jiang, X., Harris-Love, M.L., Zeng, J., Turkeltaub, P.E., 2016. Right hemisphere grey matter structure and language outcomes in chronic left hemisphere stroke. Brain 139, 227–241.
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.