




A data-driven approach to post-stroke aphasia classification and lesion-based prediction

Jon-Frederick Landrigan,¹ Fengqing Zhang¹ and  Daniel Mirman²

Aphasia is an acquired impairment in the production or comprehension of language, typically caused by left hemisphere stroke. The subtyping framework used in clinical aphasiology today is based on the Wernicke-Lichtheim model of aphasia formulated in the late 19th century, which emphasizes the distinction between language production and comprehension. The current study used a data-driven approach that combined modern statistical, machine learning, and neuroimaging tools to examine behavioural deficit profiles and their lesion correlates and predictors in a large cohort of individuals with post-stroke aphasia. First, individuals with aphasia were clustered based on their behavioural deficit profiles using community detection analysis (CDA) and these clusters were compared with the traditional aphasia subtypes. Random forest classifiers were built to evaluate how well individual lesion profiles predict cluster membership. The results of the CDA analyses did not align with the traditional model of aphasia in either behavioural or neuroanatomical patterns. Instead, the results suggested that the primary distinction in aphasia (after severity) is between phonological and semantic processing rather than between production and comprehension. Further, lesion-based classification reached 75% accuracy for the CDA-based categories and only 60% for categories based on the traditional fluent/non-fluent aphasia distinction. The results of this study provide a data-driven basis for a new approach to classification of post-stroke aphasia subtypes in both research and clinical settings.

1 Department of Psychology, Drexel University, Philadelphia, PA 19104 USA

2 Department of Psychology, University of Edinburgh, Edinburgh EH8 9JZ, UK

Correspondence to: Jon-Frederick Landrigan Ph.D

Drexel University, 3201 Chestnut Street

Philadelphia, PA 19104, USA

E-mail: Jon.Landrigan@gmail.com

Keywords: aphasia; lesion-based diagnosis; language processing; machine learning

Abbreviations: CDA = community detection analysis; MNF = mild, non-fluent, fluent; SMOTE = synthetic minority over sampling technique; WAB = Western Aphasia Battery

Introduction

Aphasia is an impairment of language production and/or comprehension that is a common and severe consequence of stroke.¹⁻³ Aphasia diagnosis continues to follow a 19th century model of the

neural basis of language, the Wernicke-Lichtheim model, which primarily focuses on three functional aspects of language: fluent speech production, auditory comprehension, and speech repetition. A patient's aphasia subtype diagnosis can affect both treatment (by affecting treatment strategy selection) and research

Received October 18, 2018. Revised October 21, 2020. Accepted November 01, 2020

© The Author(s) (2021). Published by Oxford University Press on behalf of the Guarantors of Brain. All rights reserved.

For permissions, please email: journals.permissions@oup.com

because many studies recruit and group participants based on their diagnosis.⁴⁻⁶

Over the course of more than 100 years of research and clinical practice based on this framework, there have been numerous critiques of it and many issues have been identified. One of the primary issues is that the deficit profiles of individuals with the same subtype are highly variable, suggesting that the co-occurrence of symptoms is not adequately captured by the diagnostic framework.⁶⁻¹² Further, there is poor agreement between diagnostic instruments^{13,14}: only 27% agreement between the two major English-language aphasia diagnostic tests, the Western Aphasia Battery (WAB) and Boston Diagnostic Aphasia Examination (BDAE).¹⁵ Similarly, Crary et al.¹³ performed a cluster analysis on the subtests from the BDAE and the WAB and found less than 40% agreement between their clusters and the subtype diagnoses.

A more recent analysis of 65 left hemisphere stroke cases found that 26.5% were 'unclassifiable' and that there was very poor correspondence between lesion location and aphasia subtype.¹² This is echoed in other recent studies that found that the lesion-deficit correspondence proposed in the classic models of aphasia is not well-supported by the data.^{12,16-18} For example, there have been cases of global aphasia in the presence of an intact Wernicke's area, and fluent aphasia in the presence of anterior lesions.¹⁹

Several recent studies have used data-driven approaches to describe the primary dimensions of post-stroke aphasic deficits and their lesion correlates.²⁰⁻²⁴ These studies have identified three primary dimensions (for a review see Mirman and Thye²⁵): phonological processing (recognition and production of speech sounds), semantic cognition (conceptual knowledge), and fluency (sentence-level speech planning and production). The current study similarly used data-driven methods, but leveraged tools and techniques from machine learning to focus on the problem of classifying behavioural deficit profiles and evaluating whether these classes are predictable from lesion patterns. The current study differs from prior studies because it attempts to cluster patients based on their behavioural profiles, whereas those prior studies performed principal component analysis (PCA) to cluster behavioural measures together to identify the primary dimensions. Classifying deficit profiles is inherently imperfect because no two patients have exactly the same profile, but an effective classification scheme has substantial utility for both clinical research (e.g. many treatment studies group patients based on their diagnosis) and clinical practice (i.e. clinicians use diagnoses as a shorthand to convey deficit information). In the current study, community detection analysis (CDA) was used to cluster individuals based on their pattern of language deficits following stroke. Then, a random forests classifier was built to categorize individuals based solely on their lesion profiles. Taking this approach allowed us to directly attack the issue of finding a robust classification scheme that can be utilized in clinical research and practice to classify patients and identify treatment options to improve their quality of life.

Materials and methods

Behavioural and lesion location data

The initial behavioural data for this study were downloaded from the Moss Aphasia Psycholinguistics Project Database (MAPPD) in March 2017²⁶ and contained data from 296 participants and 43 features (i.e. six demographic measures and 37 behavioural/cognitive measures). The dataset contains test scores from participants who partook in studies at the Moss Rehabilitation Research Institute (MRRI) beginning in 1991 through the date of download, who had language impairments following left hemispheric stroke. Participants were between

the ages of 18 and 80 and primarily monolingual English speakers (<5% reported speaking a second language). Most participants had chronic aphasia (i.e. >6 months post stroke). These are retrospective data gathered over 20+ years of research, and the contents of the testing battery evolved and changed (i.e. tests were added and removed). As a result, some participants were missing data due to the changing composition of the test battery. The missing data are a result of evolving research interests and methods at MRRI and do not reflect properties participants (other than the year they were tested); therefore, the data are considered to be missing at random.

Redundant measures (e.g. Sentence Comprehension Lexical A and Sentence Comprehension Lexical B) were collapsed to a single score: the mean of the two test scores was used if a participant had both; if only a single test score was available, then that score was used. Subset measures (e.g. Synonymy Triplet Nouns and Synonymy Triplets Verbs are subsets of Synonymy Triplets Total) were dropped because they are dependent on each other; only the total score was used in such cases. The final dataset contained 20 measures (for a more detailed description of each test and measure, see Mirman et al.²⁶): WAB aphasia quotient, Philadelphia Naming Test (PNT) name verification (word-to-picture matching), PNT semantic errors, PNT formal errors, PNT non-word errors, auditory discrimination (collapsed), synonymy triplets total, rhyme discrimination, immediate serial recall, semantic short term memory span, Peabody Picture Vocabulary Test, phonological short term memory span, auditory lexical decision (d'), semantic category discrimination, Camel and Cactus Test, Pyramids and Palms Test, non-word repetition (collapsed), sentence comprehension lexical (collapsed), sentence comprehension reversible (collapsed), Philadelphia Repetition Test accuracy. To control the amount of missing data further, participants who were missing more than 60% of the scores were excluded from further analyses (see [Supplementary Fig. 1](#) for the distribution of per cent missing data). This threshold resulted in 226 participants being included in the final dataset. See [Table 1](#) for the distribution of WAB diagnoses, aphasia severity, and chronicity.

After the initial data cleaning, 15.9% of the data were missing. As complete case analysis can lead to biased estimates and reduce statistical power²⁷ we conducted multivariate imputation by chained equations (MICE) implemented in the *mice* package for R (version 2.46.0²⁸). MICE is one type of multiple imputation technique that has emerged as a principled method of handling

Table 1 Overview of participant characteristics in the clustering analysis

| WAB diagnosis | n | WAB AQ | Months post onset |
|----------------------------|----|------------------|-------------------|
| Anomic | 90 | 87.6 (60.2–97.9) | 23.49 (1–290) |
| Broca's | 58 | 56.3 (25.2–84.7) | 49.57 (2–195) |
| Conduction | 41 | 72.1 (44.0–84.0) | 23.46 (2–170) |
| Global | 1 | 32.8 | 10 |
| PCA aetiology ^a | 3 | 96.8 (95.2–99.3) | 24.67 (7–42) |
| Transcortical motor | 2 | 73.9 (72.2–75.5) | 15.50 (4–27) |
| Transcortical sensory | 6 | 62.1 (53.0–69.9) | 52.50 (4–234) |
| Wernicke's | 25 | 57.4 (39.3–82.2) | 34.52 (1–381) |

Values are presented as mean (range).

^aPosterior cerebral artery (PCA) aetiology refers to participants who had suffered a stroke in the posterior cerebral artery. A formal WAB diagnosis was unavailable for these patients, but they had very mild deficits, approximately consistent with the anomic subtype.

missing data in the literature.^{28,29} In short, the MICE procedure models each variable with missing value conditional upon other available data via a series of regression models. Importantly, MICE incorporates random variation in its computations of the missing data. The introduction of random variation across multiple imputations produces multiple plausible values for each missing data-point, thus restoring the error variance that would be lost from regression-based single imputations and reducing the bias from estimating the value based on the other data-points in the data set.^{29,30} Therefore, the imputations must be done multiple times, creating multiple complete datasets. Analyses must then be performed on each complete dataset and collapsed, providing a better estimate of the outcome of interest. (For a full review of multiple imputations see Sinharay *et al.*³¹). Further studies have shown that multiple imputations can reduce bias even when the proportion of missing data is large.³² Five complete datasets were constructed using multiple imputations; these datasets were identical in terms of the number of participants and measures, but differed in terms of the values that were imputed. A separate clustering analysis (see next section) was applied to each of the datasets and the results were collapsed post-clustering using majority vote.

Lesion location data were available for 134 of 226 participants included in the CDA clustering. These data were part of a larger ongoing project and subsets of these data have been used in a number of other studies.^{24,33} The structural scans were composed of 117 research scans (75 MRI and 42 CT) and 17 clinical scans (five MRI and 12 CT). A technician manually segmented lesions that were imaged with MRI. These segmentations were then reviewed by an experienced neurologist for accuracy. Images were first registered to a custom template constructed from images acquired on the same scanner, then registered to the Montreal Neurological Institute space 'Colin27' volume. For the CT scans, a neurologist drew lesions onto the Colin27 volume, after rotating (pitch only) the template to approximate the slice plane of the patient's scan.

Analysis methods

Clustering using community detection analysis

CDA was used to identify groups of participants with similar behavioural deficits. In short, CDA comes from graph theory, which is a discipline concerned with the study of graphs and networks. Networks consist of sets of nodes that are connected by edges. CDA attempts to uncover sets of nodes (i.e. communities or clusters) that are densely connected to one another but only sparsely connected to other communities of nodes (for a review see Fortunato³⁴). CDA has a distinct advantage over other clustering algorithms, such as k-means and hierarchical clustering, in that it provides the optimal number of clusters based on the data, as opposed to the investigator predefining the number of clusters, selecting the number of clusters through subjective inference or through *post hoc* analyses. In a preliminary analysis a hierarchical clustering analysis was performed; however, after applying the elbow method and the average silhouette method, there was no consensus on the optimal number of clusters. Further, cluster membership was not stable across the different imputed data sets: the average entropy h for the hierarchical clustering analysis was 0.81. By comparison, for the CDA clusters mean $h = 0.16$ (lower values indicate more stable clustering results). CDA was developed to examine the community structure of networks (e.g. social

networks), but has also been recently used to uncover subtypes of attention deficit hyperactivity disorder.^{35,36}

To apply CDA to the aphasic deficit data, we considered each participant as a node in the network and participants with similar deficit profiles were defined as having a connection between them. First, all scores were normalized and pairwise correlations were calculated for each pair of participants (i.e. correlations were computed using the full set of behavioural measures for each pair of participants). Second, a correlation threshold was applied to the pairwise correlation matrix to define whether or not two patients were connected. The threshold was set at the highest value such that there were no isolated nodes in the network; in other words, so that every participant node was connected to at least one other participant node (for specific threshold values see [Supplementary Table 1](#)). Note that because the value of the threshold is dependent on the data and the data were variable due to multiple imputations, the threshold was separately calculated for each of the imputed data sets and ranged from 0.391 to 0.474. The [Supplementary material](#) also describes results from alternative network construction strategies. Put simply, if two participants had strongly correlated scores, then they were linked in the network; on the other hand, if participants had weakly correlated scores, then they were not linked ([Fig. 1](#)). Constructing the network in this manner allowed the CDA to uncover distinct groups of participants who had highly correlated test scores and therefore similar behavioural profiles. The particular CDA algorithm was the edge-betweenness community detection algorithm,^{37,38} implemented through functions provided by the *iGraph* package available for R (Version 1.1.2³⁹). This algorithm operates by first determining the edge betweenness (i.e. number of shortest paths that run through an edge) of all edges in the network and removes the edge with the greatest edge betweenness value. The algorithm runs until all edges have been removed from the network. Cluster membership is determined by separating nodes based on the maximum modularity value obtained while removing edges. The modularity value is considered to be a goodness of cluster metric that compares the number of edges within a given cluster to the number of edges within a cluster if connectivity of the network were completely random while preserving the node degree distribution.^{40,41}

CDA was applied to each of the five imputed data sets separately and the results of the individual CDAs were then combined by aligning the clusters across CDAs and using majority vote to determine overall cluster assignment for each participant. That is, if a participant was placed into the same cluster in three or more of the analyses, then that participant was assigned to that cluster. If a participant was not placed into the same cluster in at least three of five CDAs, then they were considered non-clustered. To facilitate qualitative characterization of the differences in pattern of deficits between clusters, permutation-based ANOVAs were run for each of the measures because data were non-normally distributed and there were unequal variances across measures (permuco package for R version 1.0.2⁴²). For measures with statistically significant differences between clusters, *post hoc* pair-wise comparisons were made using a non-parametric permutation-based t-test due to the violation of t-test assumptions (DAAG package for R version 1.22.1⁴³).

Lesion location differences

For each cluster, a comparison of lesion location was carried out using a χ^2 test in each voxel (i.e. voxel-wise 2×2

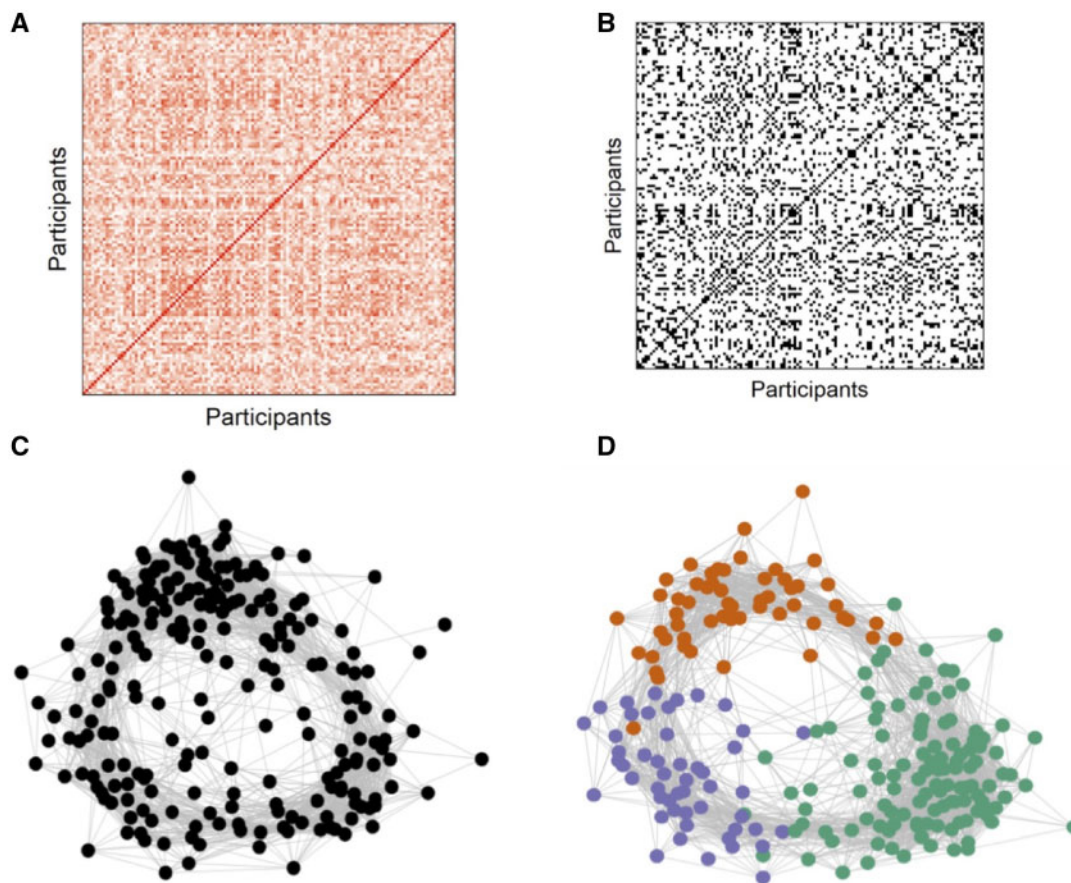


Figure 1 CDA steps. (A) Correlation matrix containing all pairwise correlations between participants (white = low correlation, red = high correlation). (B) Adjacency matrix based on thresholded correlation matrix (black dots represent links). (C) Network produced from the adjacency matrix. (D) Results of CDA with colours denoting cluster membership.

contingency table of whether the voxel was lesioned or not and whether the participant belonged to that cluster or not). In this approach, each cluster was compared to all the remaining clusters in order to identify neural regions where damage was specifically associated with membership in that individual cluster. To reduce mislocalization of effects, only voxels where at least 10% of the patients had lesions were included. Although this inclusion threshold makes it impossible to detect effects in regions where only a few participants have lesions, it prevents overfitting brain regions where very little lesion data is available.⁴⁴ The continuous family wise error rate (cFWE) correction method was used to control false positives due to multiple comparisons.⁴⁵ The cFWE is a generalization of standard permutation-based FWER correction that allows specifying the upper limit of expected number of false positive voxels at some value >1 [because a single false positive voxel rarely, if ever, affects interpretation of a voxel-based lesion-symptom mapping (VLSM) result]. For the present VLSM analyses, the upper limit was set at $v = 100$ (i.e. no more than 100 false positive voxels would be present in the results). Analyses were carried out using ANTSR (Version 0.7.2⁴⁶).

Lesion-based diagnosis

As a final test of this data-driven classification of individuals with aphasia, a prediction model was developed to evaluate whether individual lesion patterns could predict behavioural cluster membership. First, the lesion map for each participant

($n = 134$) was parcellated according to the cortical regions identified in the HCP atlas⁴⁷ and the white matter tracts identified in the ICBM-DTI white matter tractography atlas from FSL,⁴⁸⁻⁵⁰ resulting in 150 total parcels within the lesion coverage of this participant sample. The per cent damage in each parcel was calculated for each participant and combined into a per cent damage vector. Per cent damage vectors were then binarized: if 50% or more voxels were damaged in a given region for a participant, then it was considered 'damaged', otherwise it was not. To reduce the dimensionality of the feature space (i.e. number of included regions) two sequential steps were taken. First, regions were only included if at least one participant had damage in a given region post binarization ($n = 81$). Second, feature dimensions were further reduced by fitting initial random forest classifiers on the features from step 1 and selecting the features with non-trivial importance (i.e. gini feature importance > 0.005), resulting in a subset of ~ 50 features. While other approaches were considered for feature selection and modelling (i.e. non-binarized features and varying the gini importance thresholds); the reported approach appeared to produce the best and most robust performance. Feature reduction steps such as these are standard in machine learning when dealing with a large number of predictors.⁵¹ Reducing the number of predictors was especially important in the current context because the number of predictors (lesion features) was greater than the number observations (i.e. there were 180 initial features and 134 scans to train and test the model).

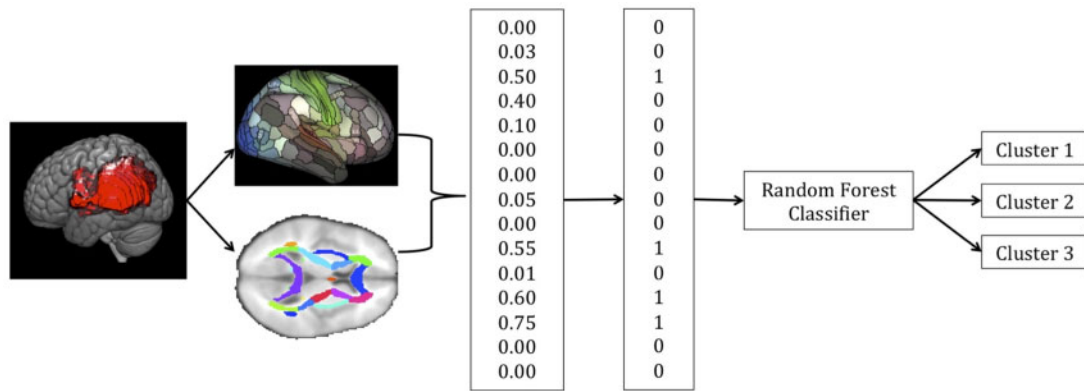


Figure 2 Schematic diagram of lesion-based diagnosis. Beginning with the participant's individual lesion map (left), per cent damage is calculated in the regions identified in the cortical and white matter tract atlases and converted into a per cent damage vector. The per cent damage vector is then binarized and used as the input feature vector for the random forest and the output is one of the clusters identified by CDA.

Table 2 Number of participants in each class for the lesion-based diagnosis analyses

| Classifier type | Class | Number of observations |
|-----------------|-----------------------|------------------------|
| CDA cluster | 1 | 74 |
| | 2 | 34 |
| | 3 | 26 |
| WAB diagnosis | Anomic | 54 |
| | Broca's | 37 |
| | Conduction | 24 |
| | Global | 1 |
| | PCA aetiology | 3 |
| | Transcortical motor | 2 |
| | Transcortical sensory | 2 |
| MNF | Wernicke's | 11 |
| | Mild | 57 |
| | Fluent | 37 |
| | Non-fluent | 40 |

PCA = posterior cerebral artery.

Leaving all features in the model could have led to overfitting.^{52,53} These vectors were then used as the feature vector input for a random forest classifier, which was trained to learn the appropriate CDA cluster for each participant. A schematic of this analysis pipeline is shown in Fig. 2.

To compare with the traditional aphasia subtyping system, classifiers were trained in the same manner as above, but using WAB diagnosis as the outcome instead of CDA cluster. For the number of observations in each class (i.e. CDA clusters and WAB diagnoses) see Table 2. A final classifier was developed for a WAB-based three-group classification of mild, fluent, and non-fluent (MNF) groups. The mild group consisted of participants with anomic aphasia ($n = 54$) and those with posterior cerebral artery aetiology ($n = 3$), the fluent group consisted of 37 participants with fluent aphasia subtypes (conduction $n = 24$, Wernicke's $n = 11$, and transcortical sensory $n = 2$), and the non-fluent group consisted of 40 participants with non-fluent aphasia subtypes (Broca's $n = 37$, global $n = 1$, and transcortical motor $n = 2$).

For a first set of analyses, the classifiers were trained on the base set of observations and the reduced feature set. In a second set of analyses, the datasets were up-sampled using the

synthetic minority oversampling technique (SMOTE⁵⁴) in order to balance the number of training observations in each class. This was used for the CDA-based and WAB-based MNF classes to balance the class sizes and prevent the models from being biased towards the majority class during training. More specifically, during training, machine learning models are designed to maximize the classification accuracy—the proportion of cases classified correctly by the model. Unbalanced outcomes can lead to a biased model because the model may achieve a high classification accuracy during training by simply classifying everything as the majority class.^{55–57} One important advantage of SMOTE is that the artificial samples are unique and are not repetitions of original observations, so classification accuracy is not inflated by repetition or leakage from training to test sets. In addition, the up-sampled data provide a more balanced and robust testing set that is not biased toward the majority class (e.g. Cluster 3 only had 26 samples to be trained and tested, so without up-sampling, a typical test set only contained about two samples per fold). Using SMOTE on the full set of WAB diagnoses was not viable because some classes were far too small (there was only a single participant diagnosed with global aphasia, and only two each with transcortical motor and transcortical sensory aphasia). The small number of observations in these classes are insufficient for the SMOTE algorithm to create synthetic observations because the synthetic observations are based on averages from the true observations. All classifiers were assessed by running 10-fold cross validation 100 times and taking the mean accuracy across runs. Mean chance performance was assessed by permuting the true labels 20 times and running 10-fold cross validation on each permutation. All modelling was carried out in python (Version 3.6.6). SMOTE was performed using functions provided by the *imblearn* package (Version 0.3.3). Random forest classifiers were developed using functions provided by the *scikit-learn* package (Version 0.19.1).

Data availability

Anonymized behavioural data are available at www.mappd.org (requires free account registration and agreement to abide by terms of data use). Lesion data are available from Daniel Mirman (dan@danmirman.org) upon reasonable request and subject to approval by the appropriate regulatory committees and officials.

Table 3 Mean test performance by CDA cluster

| Majority vote cluster | 1 | 2 | 3 | 1 versus 2 | 1 versus 3 | 2 versus 3 |
|---|-------------|-------------|-------------|------------|------------|------------|
| n | 116 | 55 | 50 | – | – | – |
| WAB Aphasia Quotient, standard score | 85.6 (9.1) | 58.3 (17.7) | 61.6 (11.8) | <0.001 | <0.001 | 0.297 |
| PNT Name Verification, % correct | 93.7 (5.3) | 76.7 (18.7) | 69.9 (23) | <0.001 | <0.001 | 0.127 |
| Semantic errors, % | 5.7 (3.6) | 5.7 (4.5) | 12.7 (5.3) | 0.938 | <0.001 | <0.001 |
| Formal errors, % | 2.6 (3.1) | 19.2 (10.1) | 7.7 (5.4) | <0.001 | <0.001 | <0.001 |
| Non-word errors, % | 8.5 (7.4) | 41.7 (20.9) | 13.1 (10.9) | <0.001 | 0.004 | <0.001 |
| Auditory discrimination, % correct | 89.8 (6.5) | 78.9 (11.4) | 82.2 (9.4) | <0.001 | <0.001 | 0.109 |
| Synonymy Triplets Total, % correct | 85.2 (10.4) | 74.1 (18.4) | 59.6 (13.6) | <0.001 | <0.001 | <0.001 |
| Rhyme discrimination, % correct | 94.8 (6.6) | 80.6 (13.7) | 85.9 (14.5) | <0.001 | <0.001 | 0.144 |
| Short-term memory ISR, span | 3.5 (0.8) | 1.6 (0.7) | 2.3 (1) | <0.001 | <0.001 | 0.003 |
| Peabody Picture Vocabulary Test, standard score | 86.6 (12.1) | 72.6 (19.7) | 59.3 (17.4) | <0.001 | <0.001 | 0.005 |
| Semantic short-term memory, span | 2.9 (1.1) | 1.6 (1) | 0.7 (0.4) | <0.001 | <0.001 | <0.001 |
| Phonological short-term memory, span | 3.9 (1.7) | 1.7 (1.4) | 1.9 (1.1) | <0.001 | <0.001 | 0.702 |
| Auditory Lexical Decision, <i>d'</i> | 2.6 (0.9) | 1.9 (0.8) | 2.4 (0.7) | <0.001 | 0.138 | 0.007 |
| Semantic Category Discrimination, % correct | 89.1 (6.7) | 77.2 (13.4) | 67.8 (11.8) | <0.001 | <0.001 | 0.004 |
| Camel and Cactus Test, % correct | 80.2 (8.6) | 72.4 (12.5) | 52.9 (15.2) | <0.001 | <0.001 | <0.001 |
| Pyramids and Palms Test, % correct | 91.1 (6.1) | 88.3 (8.8) | 77.1 (12.5) | 0.019 | <0.001 | <0.001 |
| Non-word repetition, % correct | 60 (20.3) | 21.8 (18.4) | 50.8 (21.3) | <0.001 | 0.023 | <0.001 |
| Sentence Comprehension: Lexical, % correct | 96.3 (5.6) | 88.2 (11) | 82.8 (12.1) | <0.001 | <0.001 | 0.025 |
| Sentence Comprehension: Reversible, % correct | 80.3 (14) | 64 (14.9) | 56.8 (11.5) | <0.001 | <0.001 | 0.014 |
| Philadelphia Repetition Test, % correct | 93.2 (5.8) | 62.8 (20.5) | 89.5 (9.5) | <0.001 | 0.003 | <0.001 |
| WAB aphasia subtype, n | | | | | | |
| Anomic | 78 | 2 | 9 | | | |
| Broca's | 18 | 16 | 22 | | | |
| Conduction | 15 | 22 | 2 | | | |
| Global | 0 | 0 | 1 | | | |
| PCA aetiology | 3 | 0 | 0 | | | |
| Transcortical motor | 1 | 0 | 1 | | | |
| Transcortical sensory | 0 | 1 | 5 | | | |
| Wernicke's | 1 | 14 | 10 | | | |

The right-most columns show *P*-values for pair-wise comparisons between clusters. Values are presented as mean (SD). ISR = Immediate Serial Recall; PCA = posterior cerebral artery.

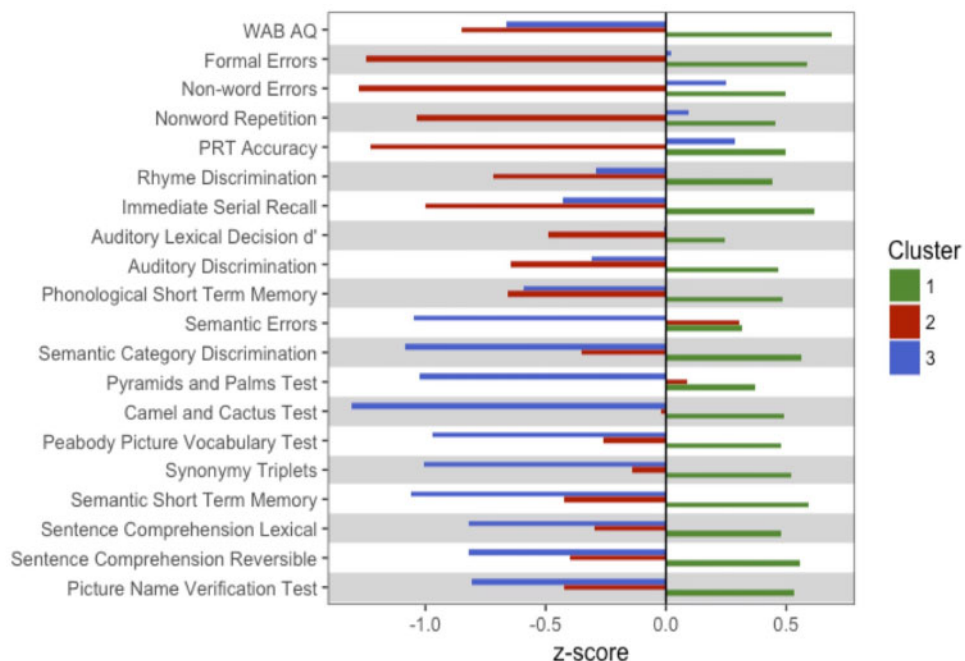


Figure 3 Standardized test performance by CDA clusters.

Results

Clustering using community detection analysis

The primary CDA resulted in three main clusters containing 97.8% of the participants. There were 116 participants in cluster 1, 55 participants in cluster 2, and 50 participants in cluster 3. The remaining five participants were deemed non-clustered because they were not placed in the same cluster in at least three of the analyses. Table 3 shows the mean performance for each cluster on each test or measure, and Fig. 3 shows the mean standardized test performance on each measure. The non-parametric permutation-based ANOVAs found statistically significant differences on every measure and pairwise comparisons revealed a coherent pattern of differences between pairs of clusters. Participants in cluster 1 typically had milder aphasia (mean WAB AQ 85.57, more than 20 points higher than the other two clusters, both $P < 0.001$) and generally scored better on all of the measures than the participants in clusters 2 and 3. The differences between cluster 1 and 2 were statistically significant for each measure and the differences between cluster 1 and 3 were statistically significant for all but two measures (Table 3). Most participants in cluster 1 had been diagnosed with anomic aphasia ($n = 78$), but this cluster also contained a substantial portion of the participants with Broca's ($n = 18$) and conduction aphasia ($n = 15$). Participants in clusters 2 and 3 had similar overall aphasia severity (WAB AQ means: 58.31 and 61.63, respectively; $P = 0.297$); however, their profiles differed in regard to the other measures. Participants in cluster 2 typically performed worse on phonological measures. For example, they had higher rates of formal errors (cluster 2 mean = 19.2, cluster 3 mean = 7.7, $P < 0.001$) and non-word errors (cluster 2 mean = 41.7, cluster 3 mean = 13.1, $P < 0.001$) in picture naming, poor word repetition (cluster 2 mean = 62.8, cluster 3 mean = 89.5, $P < 0.001$), and very poor non-word repetition (cluster 2 mean = 21.8, cluster 3 mean = 50.8, $P < 0.001$), but their performance on semantic tasks was relatively preserved. Participants in cluster 3 performed relatively well on phonological measures but performed poorly on semantic measures. For example, they had high rates of semantic errors in picture naming (cluster 2 mean = 5.7, cluster 3 mean = 12.7, $P < 0.001$) and low scores on the Camels and Cactus Test (cluster 2 mean = 72.4, cluster 3 mean = 52.9, $P < 0.001$) and synonym judgements (cluster 2 mean = 74.1, cluster 3 mean = 59.6, $P < 0.001$). They also performed more poorly on sentence comprehension (Lexical: cluster 2 mean = 88.2, cluster 3 mean = 82.8, $P < 0.05$; Reversible: cluster 2 mean = 64.0, cluster 3 mean = 56.8, $P < 0.05$). Although sentence comprehension relies on both phonological and semantic processing, as well as syntactic processes, the overarching deficit profile of patients in cluster 3 suggests that the semantic demands may be more prominent in this task. Both clusters 2 and 3 contained substantial numbers of participants with Broca's aphasia ($n = 16$ and $n = 22$, respectively) and Wernicke's aphasia ($n = 14$ and $n = 10$, respectively). Consistent with the phonological-semantic divide between clusters 2 and 3, the remaining participants with Conduction aphasia (those who were not in cluster 1) tended to be in cluster 2 ($n = 22$) rather than cluster 3 ($n = 2$), while the participants with transcortical sensory aphasia tended to be in cluster 3 ($n = 5$) rather than cluster 2 ($n = 1$). See Table 3 for the full breakdown of WAB diagnoses per CDA cluster.

A subsequent analysis was run to determine if cluster 1 could be broken down further. The analysis followed the same CDA procedure as described above. The CDA revealed a subcluster of 31 patients within cluster 1. This cluster of participants had lower WAB AQ scores (mean = 79.6) and more severe phonological impairments (e.g. higher rates of formal and non-word errors on the PNT and poorer performance on the Philadelphia Repetition

Test) relative to the rest of the patients in cluster 1. However, given that only one subcluster was identified and the rest of the participants were deemed unclassified or as singlet clusters, this subcluster was not included in further analyses.

A series of follow-up analyses were conducted to check that the cluster structure reported above was not an artefact of particular analysis choices. A different CDA algorithm [order statistics localization optimization method (OSLOM)⁵⁸] produced qualitatively the same three-cluster structure. These results were also replicated using Spearman correlations instead of Pearson correlations, running a Louvain CDA using on weighted networks as opposed to binarized networks, and under moderate increases or decreases (± 0.05 , ± 0.10 , ± 0.20) in the correlation threshold that was used to determine whether two participant nodes were linked or not. The clustering analysis was also repeated without participants who were missing data (i.e. no imputations were performed). The overall cluster structure and behavioural deficit profiles of the clusters the same, and 93% of participants were placed in the same cluster as they were in the main analysis, suggesting the imputations did not bias the results. Finally, a support vector machine classifier (implemented with the *e1071* package for R, Version 1.6–8⁵⁹) was used to evaluate how well the behavioural assessment scores predict cluster membership (i.e. support vector machines were trained to predict cluster membership or WAB diagnoses using the behavioural test scores). The 10-fold cross-validation classification accuracy was 94.0% (chance classification accuracy estimated by permutation was 43.0%), indicating that the cluster structure and membership robustly corresponded to individual participants' assessment scores (for more details about these robustness checks see the [Supplementary material](#)).

Lesion location differences

Participants in cluster 1 typically had the smallest lesions [mean = 69.39 cm³, standard deviation (SD) = 61.12 cm³], participants in cluster 3 had the largest lesions (mean = 167.64 cm³, SD = 90.01 cm³) and those in cluster 2 were intermediate (mean = 117.08 cm³, SD = 88.11 cm³). Lesion coverage for the full sample is shown in Fig. 4, top row. Comparisons of lesion locations revealed that cluster 1 was associated with damage to the core perisylvian portion of the middle cerebral artery distribution, primarily consisting of the inferior parietal lobe (including the angular gyrus, supramarginal gyrus, and postcentral gyrus) and the superior temporal gyrus, extending to the superior temporal pole and into the middle temporal gyrus (Fig. 4, second row). That is, cluster 1 membership was associated with relatively smaller lesions in the left perisylvian regions typically associated with post-stroke aphasia. Cluster 2 was primarily associated with damage to parietal areas and more specifically with damage to the supramarginal gyrus extending anteriorly into the postcentral gyrus (Fig. 4, third row). Finally, cluster 3 was associated with damage to frontal areas, including the precentral gyrus, inferior frontal gyrus pars opercularis and pars triangularis, the insula, and extending sub-cortically into the putamen and globus pallidus (Fig. 4, bottom row). For the number of above threshold voxels in each region by cluster see [Supplementary Table 2](#). To rule out artefacts due to inclusion of 22 participants with subchronic aphasia (<6 months post-stroke), additional VLSM comparisons excluding these participants were conducted. Because of reduced statistical power, the results included fewer above threshold voxels, but the overall pattern of results remained largely the same. See [Supplementary Table 3](#) for details. Although the neuroimaging and behavioural testing were conducted at approximately the same time, providing a valid snapshot of their lesion-symptom relationship chronicity can have a large impact on both cognitive performance and lesion

pattern, so a subsequent analysis was run excluding patients less than 6 months post stroke. Importantly while the number of above threshold voxels was slightly reduced, the overall patterns remained unchanged ([Supplementary Table 3](#)).

Lesion-based diagnosis

Lesion-based classification accuracy was highest for the CDA clusters and lowest for the WAB aphasia subtype diagnoses. Not surprisingly, the classifier performed better on the three-group WAB diagnosis (MNF) than on the full set of WAB diagnoses, but even this three-group WAB classification accuracy was substantially lower than its performance on the CDA clusters ([Table 4](#)). For both the CDA clusters and the three-group WAB diagnoses, using SMOTE to equate class sizes substantially improved performance. The overall highest mean lesion-based classification accuracy was for the SMOTE CDA clusters dataset, at 76.9% (chance performance estimated by permutation = 33.9%). The most direct WAB comparison was the SMOTE three-group WAB diagnoses, which

produced slightly lower classification accuracy, at 64.1% (chance = 34.6%).

Without SMOTE, lesion-based classification accuracy was approximately equal for CDA clusters (59.6%) and three-group WAB diagnosis (55.9%), and somewhat lower for the full range of WAB diagnoses (49.3%). The base MNF and WAB classifiers had larger performance difference from chance (19.1% and 16.9%, respectively) than the CDA classifier did (12.2%) because, without SMOTE, chance performance was higher for CDA clusters. This is because unbalanced classes make simple probability matching a more effective strategy (CDA clusters were more unbalanced: the largest CDA cluster was 55% of the participants, but the largest MNF cluster was 43% of participants). The chance performance difference disappeared when cluster size was equated using SMOTE and the SMOTE CDA classifier substantially outperformed the MNF classifier relative to chance performance (i.e. 43.0% above chance versus 29.5% above chance). Statistical comparisons of the classifier performances were carried out using a logistic regression (i.e. outcome was correct/incorrect with model type as the predictor). All pairwise comparisons were significant at $P < 0.0001$ ([Table 4](#)).

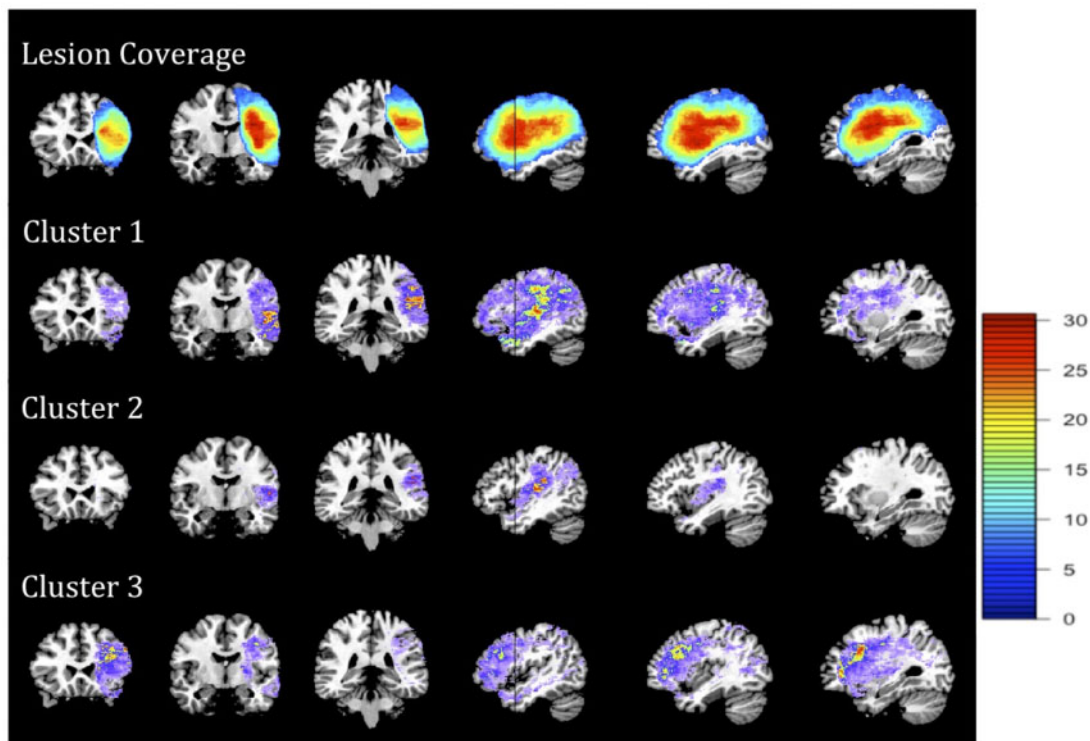


Figure 4 Lesion location patterns for CDA clusters. Top row: Lesion coverage, colour corresponds to proportion of sample with damage in each voxel, ranging from 0.1 (blue, minimum for inclusion in these analyses) to 0.5 (red). Rows 2–4: Lesion location comparison results for each cluster. Regions in the green-to-red spectrum survived correction for multiple comparisons; regions in blue-to-purple are below that threshold. For all rows, from left to right: coronal slices at $y = 150, 120$ and 90 and sagittal slices at $x = 45, 52$ and 60 , respectively.

Table 4 Pairwise comparisons of lesion-based diagnosis models

| Outcome | CDA | WAB | MNF | CDA versus WAB | CDA versus MNF | MNF versus WAB |
|--------------|------------|------------|------------|----------------|----------------|----------------|
| Base | 59.6 (2.6) | 49.3 (2.4) | 55.9 (2.1) | <0.0001 | <0.0001 | <0.0001 |
| Base chance | 47.4 (3.5) | 32.4 (5.2) | 36.8 (5.9) | – | – | – |
| SMOTE | 76.9 (1.5) | – | 64.1 (1.7) | – | <0.0001 | – |
| SMOTE chance | 33.9 (4.5) | – | 34.6 (4.0) | – | – | – |

For model fits see [Supplementary Table 4](#).

Discussion

The present study took a data-driven approach to identifying clusters of individuals with post-stroke aphasia who have similar deficit profiles. To identify these clusters, we applied CDA to a large dataset of 20 psycholinguistic measures from 226 participants with aphasia. The CDA revealed three distinct clusters. Cluster 1 consisted of individuals who generally had milder deficits as indicated by higher WAB AQ scores and better performance on all assessment measures. Individuals in clusters 2 and 3 had similar WAB AQ severity scores, but different deficit profiles. Individuals in cluster 2 typically performed worse on measures of phonological abilities, indicative of a phonological processing deficit, and individuals in cluster 3 typically performed worse on measures of semantic abilities, indicative of a semantic cognition deficit. This clustering was substantially different from the traditional aphasia subtypes as defined by the WAB (Table 3): individuals with Broca's aphasia were nearly evenly distributed across the three clusters, individuals with conduction aphasia were approximately evenly divided between clusters 1 and 2, individuals with Wernicke's aphasia were approximately evenly divided between clusters 2 and 3. That is, CDA clustered individuals according to a distinction between phonological and semantic deficits that was nearly orthogonal to the traditional aphasia subtypes. The main point of agreement was that both CDA and the WAB diagnostic framework identified a relatively large group of individuals with mild aphasia (cluster 1 in CDA, anomic aphasia in WAB).

Lesion location comparisons revealed brain regions where damage was uniquely associated with each of the CDA clusters. Cluster 1 was generally associated with smaller lesions and damage to left perisylvian portions of the middle cerebral artery territory (inferior parietal and superior temporal areas). Smaller lesions in the core of the MCA territory is consistent with the comparatively mild deficit profile of this cluster. Cluster 2 was primarily associated with damage to the supramarginal gyrus extending anteriorly into the postcentral gyrus. These regions comprise the dorsal speech production system,^{21,24,60} consistent with this cluster's prominent phonological deficits, which were particularly pronounced in speech production tasks. Cluster 3 was associated with damage to frontal regions, converging with other findings suggesting that, in post-stroke aphasia, broad semantic deficits across verbal and non-verbal tasks arise from impaired semantic control systems that are needed to select the context- and task-relevant semantic knowledge.^{61,62} It was not the purpose of these comparisons to determine the neural correlates of phonological and semantic deficits, which have been investigated extensively and more effectively in prior studies. Rather, these neural analyses provide converging evidence that the CDA-based patient clusters reflect distinct mild, phonological, and semantic deficits, both behaviourally and neuroanatomically. A secondary goal was to provide a preliminary test of whether lesion-based diagnosis would be viable, which was then tested using random forests classifiers. In other words the overall purpose of the present study was a data-driven approach to aphasia classification, so it was concerned with the lesion location → cluster analysis, which was implemented by random forests classifiers, rather than cluster → lesion location analysis, as would be implemented by lesion symptom mapping (LSM) approaches. Such LSM analyses are a potential avenue for future research and would benefit from multivariate and/or connectome lesion symptom mapping methods, which are better able to detect distributed networks of regions associated with particular deficits.^{23,63-66}

The random forest classifier was relatively successful in categorizing individuals into their CDA clusters based solely on structural information about their lesion locations, indicating strong correspondence between lesion location and cluster

membership. Lesion-based diagnosis is an important and difficult challenge.⁶⁷ It is important for clinical research and practice because accurate lesion-based predictions of language deficits could provide valuable information for guiding selection of personalized treatment strategies. It is challenging because lesion patterns are structured by the cerebral vasculature, not by functional systems, and because group-level lesion-symptom associations are highly variable at the individual level. Data-driven discovery of lesion-symptom associations and testing predictive inference (e.g. our random forests classifier) provide a way to improve lesion-based diagnosis.⁶⁸ In the current study, lesion location comparisons suggested that the deficit clusters had distinct lesion patterns and the random forests classifiers showed that the lesion patterns were distinct enough to be predictive of deficit cluster. Further, the current study shows that lesion-based prediction accuracy depends, in part, on what deficit is being predicted: data-driven deficit categories (e.g. those produced by CDA) may be more predictable as suggested by the overall higher accuracy of the CDA classifiers. It should be noted that classifier performance is strongly influenced by the distribution of class labels. Up-sampling small classes to balance class sizes gives the classifier more opportunities to learn predictors for the small classes and reduces chance performance (because just using the class label distribution is less effective). The base CDA classifier performed better than the base WAB and MNF classifiers on raw accuracy, but not relative to chance because the CDA clusters were more unbalanced. When SMOTE was used to balance the class sizes, the CDA classifier substantially outperformed the MNF classifier, both in terms of raw accuracy and relative to chance performance (which was approximately equal for CDA and MNF once the class sizes were balanced).

The finding that deficit clusters coalesced around phonological and semantic systems converges with other recent data-driven studies that combined principal components analysis with lesion-symptom mapping (for a review see Mirman and Thye²⁵). These studies also identified phonological and semantic systems as core deficit dimensions in post-stroke aphasia, and fluency deficits (impaired sentence- or utterance-level speech production) were the next most consistent deficit dimension. Our CDA was not able to capture fluency deficits because the dataset did not include any measures of sentence-level or utterance-level fluency. It is possible that there is another deficit cluster, one characterized by sentence-level production fluency deficits, that was missed by our CDA. It seems unlikely that this cluster would substantially reorganize the observed phonological and semantic clusters because our random forest lesion-based classification accuracy was substantially higher for the CDA clusters than for a WAB-based fluent/non-fluent distinction. That is, non-fluency seems more likely to be an additional deficit cluster rather than the core deficit dimension. Critically, a key novel contribution of the current study is the categorical clustering of patients rather than graded degrees of impairment produced by the posterior cerebral artery. This is important for clinical translation because clinical contexts often call for simple, categorical syndrome labels to aid in the development and selection of treatment strategies for patients with aphasia.

The current study builds on and converges with other recent studies that not only provide complimentary information to WAB data, but also potentially point to a 'primary systems' approach in the study and classification of aphasia.^{12,24,69} Whereas the classic models of aphasia define the primary distinction as between production and comprehension deficits (canonically between Broca's and Wernicke's aphasia), the present results and other recent data-driven studies suggest that, after severity, the primary distinction is between semantic and phonological processing. This distinction has several advantages over the classic model. First, it grounds aphasia in core cognitive systems as opposed to functions

that arise from interactions of multiple cognitive systems (e.g. spoken language production requires phonological-articulatory processes, lexical-semantic processes, as well as sentence planning and narrative monitoring processes). This fits with a more general trend in cognitive neuropsychology to emphasize 'primary systems' as the basis of neurocognitive deficits.⁷⁰ Better alignment with primary systems may explain why the CDA clusters were more robust and predictable from lesion patterns. This study provides a classification framework, but additional work is required to develop behavioural classification instruments—tests that could be administered in clinical settings to provide reliable classification data.

Second, many current therapy strategies directly target the semantic system, such as semantic feature analysis,^{71,72} or the phonological system, such as phonological or orthographic cueing⁷³⁻⁷⁵ or phonomotor therapy.⁷⁶ Clinical research on such therapies would benefit from being able to classify participants as 'semantic variant' or 'phonological variant' in order to determine which individuals benefit the most from the therapy, to uncover the therapeutic mechanism of action, and to guide clinical decisions about personalized treatment selection.

Finally, the success of the random forests classifier suggests that a classifier based solely on lesion structure could be used in clinical settings, where MRI and CT scans are standard tools used to assess damage post-stroke. The prediction accuracy is likely to be improved with the addition of more sophisticated neuroimaging data such as functional connectivity⁷⁷; however, many stroke survivors are unable or unwilling to undergo such scanning protocols because of contraindications (e.g. metal in their bodies), claustrophobia, and financial or other practical constraints. Therefore, a classifier based only on structural information could be extremely useful to clinicians if it can provide reliable diagnostic information. A critical next step is to refine the classifier's performance further and evaluate it on a larger sample of patients with a broader range of cognitive-linguistic assessments. The current classifiers were trained and tested on a relatively small sample in comparison to most machine learning development and in part relied on artificial samples to help it correctly identify patients from the minority classes, therefore a larger sample would help to test its true robustness in classifying real patients.

The results of this study and others^{20,24,25,78} suggest that data-driven diagnostic tools and assessments centred around semantic and phonological systems could provide valuable information that is different from the current standard tools (e.g. WAB diagnostic tools) for both the study and classification of post-stroke aphasia. In future research it would be useful to include measures of fluency and executive functions. Moreover, these results demonstrate the usefulness of applying machine learning and data-driven techniques in the study of aphasia. Increased data sharing and cooperation between investigators and clinicians provides new opportunities to understand neurological impairments better and to improve the outlook for individuals with post-stroke aphasia.

Acknowledgements

We would like to thank the Myrna Schwartz, Erica Middleton, and the Moss Rehabilitation Research Institute aphasia research group for providing the data for this project and for helpful discussions.

Funding

This project was supported by Drexel University and University of Alabama at Birmingham, and by the National

Institutes of Health grant R01DC017137 to D.M. and Jerzy Szaflarski. The data analysed here came from a project that was funded by a grant from the National Institutes of Health (R01DC000191) to Myrna F. Schwartz.

Competing interests

The authors report no competing interests.

Supplementary material

Supplementary material is available at *Brain* online.

References

- Boehme AK, Martin-Schild S, Marshall RS, Lazar RM. Effect of aphasia on acute stroke outcomes. *Neurology*. 2016;87(22):2348–2354.
- Flowers HL, Skoretz SA, Silver FL, et al. Poststroke aphasia frequency, recovery, and outcomes: A systematic review and meta-analysis. *Arch Phys Med Rehabil*. 2016;97(12):2188–2201.e8.
- Hilari K. The impact of stroke: Are people with aphasia different to those without? *Disabil Rehabil*. 2011;33(3):211–218.
- Basso A. *Aphasia and its therapy*, 1st edn. Oxford University Press; 2003.
- Caramazza A. The logic of neuropsychological research and the problem of patient classification in Aphasia. *Brain Lang*. 1984;21(1):9–20.
- Tremblay P, Dick AS. Broca and Wernicke are dead, or moving past the classic model of language neurobiology. *Brain Lang*. 2016;162:60–71.
- Schwartz MF. What the classical aphasia categories can't do for us, and why. *Brain Lang*. 1984;21(1):3–8.
- Dronkers NF, Plaisant O, Iba-Zizen MT, Cabanis EA. Paul Broca's historic cases: High resolution MR imaging of the brains of Leborgne and Lelong. *Brain*. 2007;130(Pt 5):1432–1441.
- Croquelois A, Bogousslavsky J. Stroke aphasia: 1,500 consecutive cases. *Cerebrovasc Dis*. 2011;31(4):392–399.
- Caplan D. Aphasic syndromes. In: Heilman MKM, Valenstein E, eds. *Clinical Neuropsychology*. 5th edn. Oxford University Press; 2012.
- Robson H, Sage K, Lambon Ralph M. A. Wernicke's aphasia reflects a combination of acoustic-phonological and semantic control deficits: A case-series comparison of Wernicke's aphasia, semantic dementia and semantic aphasia. *Neuropsychol*. 2012;50(2):266–275.
- Kasselimis DS, Simos PG, Peppas C, Evdokimidis I, Potagas C. The unbridged gap between clinical diagnosis and contemporary research on aphasia: a short discussion on the validity and clinical utility of taxonomic categories. *Brain Lang*. 2017;164:63–67.
- Crary MA, Wertz RT, Deal JL. Classifying aphasias: Cluster analysis of Western Aphasia Battery and Boston Diagnostic Aphasia Examination results. *Aphasiology*. 1992;6(1):29–36.
- Swindell CS, Holland AL, Fromm D. *Classification of Aphasia: WAB type versus clinical impression*. Univ. Pittsburgh; 1984. Accessed 2018. <http://aphasiology.pitt.edu/id/eprint/795>
- Wertz RT, Deal JL, Robinson AJ. Classifying the Aphasias: A Comparison of the Boston Diagnostic Aphasia Examination and the Western Aphasia Battery. [Clinical Aphasiology Paper] In: *Clinical Aphasiology: Proceedings of the Conference 1984*, Seabrook Island, SC, 1984; 40–7. Accessed 2018. <http://aphasiology.pitt.edu/794/>
- Henseler I, Regenbrecht F, Obrig H. Lesion correlates of patho-linguistic profiles in chronic aphasia: Comparisons of

- syndrome-, modality-and symptom-level assessment. *Brain*. 2014;137(Pt 3):918–930.
17. Mesulam M, Rogalski EJ, Wieneke C, et al. Primary progressive aphasia and the evolving neurology of the language network. *Nat Rev Neurol*. 2014;10(10):554–569.
 18. Yourganov G, Smith KG, Fridriksson J, Rorden C. Predicting aphasia type from brain damage measured with structural MRI. *Cortex*. 2015;73:203–215.
 19. Charidimou A, Kasselimis D, Varkanits M, Selai C, Potagas C, Evdokimidis I. Why is it difficult to predict language impairment and outcome in patients with aphasia after stroke? *J Clin Neurol*. 2014;10(2):75–83.
 20. Fridriksson J, den Ouden D-B, Hillis AE, et al. Anatomy of aphasia revisited. *Brain*. 2018;141(3):848–815.
 21. Fridriksson J, Yourganov G, Bonilha L, Basilakos A, Den Ouden D-B, Rorden C. Revealing the dual streams of speech processing. *Proc Natl Acad Sci U S A*. 2016;113(52):15108–15113.
 22. Halai AD, Woollams AM, Ralph MAL. Using principal component analysis to capture individual differences within a unified neuropsychological model of chronic post-stroke aphasia: Revealing the unique neural correlates of speech fluency, phonology and semantics. *Cortex*. 2017;86:275–289.
 23. Lacey EH, Skipper-Kallal LM, Xing S, Fama ME, Turkeltaub PE. Mapping common aphasia assessments to underlying cognitive processes and their neural substrates. *Neurorehabil Neural Repair*. 2017;31(5):442–450.
 24. Mirman D, Chen Q, Zhang Y, et al. Neural organization of spoken language revealed by lesion–symptom mapping. *Nat Commun*. 2015;6:6762.
 25. Mirman D, Thye M. Uncovering the neuroanatomy of core language systems using lesion-symptom mapping. *Curr Dir Psychol Sci*. 2018;27(6):455–461.
 26. Mirman D, Strauss TJ, Brecher A, et al. A large, searchable, web-based database of aphasic performance on picture naming and other tests of cognitive function. *Procedia Soc Behav Sci*. 2010;6:132–133.
 27. Graham JW. Missing data analysis: Making it work in the real world. *Annu Rev Psychol*. 2009;60:549–576.
 28. van Buuren S, Groothuis-Oudshoorn K. Multivariate imputation by chained equations in R. *J Stat Softw*. 2011;45:1–67.
 29. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: What is it and how does it work? *Int J Methods Psychiatric Res*. 2011;20(1):40–49.
 30. Allison PD. Multiple imputation for missing data a cautionary tale. *Sociol Methods Res*. 2000;28(3):301–309.
 31. Sinharay S, Stern HS, Russell D. The use of multiple imputation for the analysis of missing data. *Psychol Methods*. 2001;6(4):317–329.
 32. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. *J Clin Epidemiol*. 2019;110:63–73.
 33. Schwartz MF, Kimberg DY, Walker GM, et al. Neuroanatomical dissociation for taxonomic and thematic knowledge in the human brain. *Proc Natl Acad Sci U S A*. 2011;108(20):8520–8524.
 34. Fortunato S. Community detection in graphs. *Phys Rep*. 2010;486(3–5):75–174.
 35. Fair DA, Bathula D, Nikolas MA, Nigg JT. Distinct neuropsychological subgroups in typically developing youth inform heterogeneity in children with ADHD. *Proc Natl Acad Sci*. 2012;109(17):6769–6774.
 36. Karalunas SL, Fair D, Musser ED, Aykes K, Iyer SP, Nigg JT. Subtyping attention-deficit/hyperactivity disorder using temperament dimensions. *JAMA Psychiatry*. 2014;71(9):1015–1024.
 37. Girvan M, Newman MEJ. Community structure in social and biological networks. *Proc Natl Acad Sci U S A*. 2002;99(12):7821–7826.
 38. Newman MEJ. Detecting community structure in networks. *Eur Phys J B*. 2004;38(2):321–330.
 39. Csardi G, Nepusz T. The igraph software package for complex network research. *InterJ Complex Syst*. 2006;1695. Accessed 2018. http://interjournal.org/manuscript_abstract.php?361100992
 40. Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. *Physical Rev E Stat Nonlin Soft Matter Phys*. 2004;70:066111.
 41. Newman MEJ. Modularity and community structure in networks. *Proc Natl Acad Sci USA*. 2006;103(23):8577–8582.
 42. Frossard J, Renaud O. permuco: Permutation Tests for Regression, (Repeated Measures) ANOVA/ANCOVA and Comparison of Signals. 2019. Accessed 2018. <https://rdrr.io/cran/permuco/>
 43. Maindonald JH, Braun WJ. DAAG: Data Analysis and Graphics Data and Functions. 2019. R package version 1.22.1. 2019. Accessed 2019. <https://CRAN.R-project.org/package=DAAG>
 44. Sperber C, Karnath H. Impact of correction factors in human brain lesion-behavior inference. *Hum Brain Mapp*. 2017;38(3):1692–1701.
 45. Mirman D, Landrigan J-F, Kokolis S, Verillo S, Ferrara C, Pustina D. Corrections for multiple comparisons in voxel-based lesion-symptom mapping. *Neuropsychol*. 2018;115:112–123.
 46. Tustison NJ, Cook PA, Holbrook AJ, et al. The ANTsX ecosystem for quantitative biological and medical imaging. *Sci Rep*. 2021;11(1):9068.
 47. Glasser MF, Coalson TS, Robinson EC, et al. A multi-modal parcellation of human cerebral cortex. *Nature*. 2016;536(7615):171–178.
 48. Hua K, Zhang J, Wakana S, et al. Tract probability maps in stereotaxic spaces: analyses of white matter anatomy and tract-specific quantification. *Neuroimage*. 2008;39(1):336–347.
 49. Mori S, Wakana S, Zijl van PC, Nagae-Poetscher LM. *MRI Atlas of human white matter*. Elsevier; 2005.
 50. Wakana S, Caprihan A, Panzenboeck MM, et al. Reproducibility of quantitative tractography methods applied to cerebral white matter. *Neuroimage*. 2007;36(3):630–644.
 51. Saey Y, Abeel T, Van de Peer Y. Robust feature selection using ensemble feature selection techniques. In: W Daelemans, B Goethals, K Morik, eds. *Machine Learning and Knowledge Discovery in Databases*. ECML PKDD 2008, vol. 5212. Lecture Notes in Computer Science. Springer; 2008. 313–325.
 52. Granitto PM, Furlanello C, Biasioli F, Gasperi F. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics Intell Lab Syst*. 2006;83(2):83–90.
 53. Menze BH, Kelm BM, Masuch R, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*. 2009;10:213.
 54. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321–357.
 55. Ali A, Shamsuddin SM, Ralescu AL. Classification with class imbalance problem. *Int J Adv Softw Comput*. 2015;7:176–204.
 56. Sun Y, Wong AKC, Kamel MS. Classification of imbalanced data: A review. *Int J Pattern Recognit Artif Intell*. 2009;23(04):687–719.
 57. Visa S, Ralescu A. Issues in Mining Imbalanced Data Sets - A Review Paper. In: Proceedings of the 16th Midwest Artificial Intelligence and Cognitive Science Conference, Dayton, OH, 2005. 67–73.
 58. Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S. Finding statistically significant communities in networks. *PLoS One*. 2011;6(4):e18961.
 59. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics,

- Probability Theory Group (Formerly: E1071). 2017. <https://CRAN.R-project.org/package=e1071>. Accessed 2018.
60. Hickok G, Poeppel D. The cortical organization of speech processing. *Nat Rev Neurosci*. 2007;8(5):393–402.
 61. Mirman D, Britt AE. What we talk about when we talk about access deficits. *Philos Trans R Soc Lond B Biol Sci*. 2014;369(1634):20120388.
 62. Ralph MA, Jefferies E, Patterson K, Rogers TT. The neural and computational bases of semantic cognition. *Nat Rev Neurosci*. 2017;18(1):42–55.
 63. Boes AD, Prasad S, Liu H, et al. Network localization of neurological symptoms from focal brain lesions. *Brain*. 2015;138(10):3061–3075.
 64. Gleichgerrcht E, Fridriksson J, Rorden C, Bonilha L. Connectome-based lesion-symptom mapping (CLSM): A novel approach to map neurological function. *Neuroimage Clin*. 2017;16:461–467.
 65. Pustina D, Avants B, Faseyitan OK, Medaglia JD, Coslett HB. Improved accuracy of lesion to symptom mapping with multivariate sparse canonical correlations. *Neuropsychol*. 2018;115:154–166.
 66. Zhang Y, Kimberg DY, Coslett HB, Schwartz MF, Wang Z. Multivariate lesion-symptom mapping using support vector regression. *Hum Brain Mapp*. 2014;35(12):5861–5876.
 67. Price CJ, Hope TM, Seghier ML. Ten problems and solutions when predicting individual outcome from lesion site after stroke. *Neuroimage*. 2017;145(Pt B):200–208.
 68. Yarkoni T, Westfall J. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspect Psychol Sci*. 2017;12(6):1100–1122.
 69. Butler RA, Lambon Ralph MA, Woollams AM. Capturing multidimensionality in stroke aphasia: Mapping principal behavioural components to neural structures. *Brain*. 2014;137(Pt 12):3248–3266.
 70. Lambon Ralph MA, Graham KS, Patterson K, Hodges JR. Is a picture worth a thousand words? Evidence from concept definitions by patients with semantic dementia. *Brain Lang*. 1999;70(3):309–335.
 71. Boyle M. Semantic feature analysis treatment for aphasic word retrieval impairments: What's in a name? *Top Stroke Rehabil*. 2010;17(6):411–422.
 72. Doesborgh SJC, Van De Sandt-Koenderman MWE, Dippel DWJ, Van Harskamp F, Koudstaal PJ, Visch-Brink EG. Effects of semantic treatment on verbal communication and linguistic processing in aphasia after stroke: A randomized controlled trial. *Stroke*. 2004;35(1):141–146.
 73. Dede G, Parris D, Waters G. Teaching self-cues: A treatment approach for verbal naming. *Aphasiology*. 2003;17(5):465–480.
 74. Hickin J, Best W, Herbert R, Howard D, Osborne F. Phonological therapy for word-finding difficulties: A re-evaluation. *Aphasiology*. 2002;16(10-11):981–999.
 75. Nickels L. Therapy for naming disorders: revisiting, revising, and reviewing. *Aphasiology*. 2002;16(10-11):935–979.
 76. Pompon RH, Bislick L, Elliott K, et al. Influence of linguistic and nonlinguistic variables on generalization and maintenance following phonomotor treatment for aphasia. *Am J Speech Language Pathol*. 2017;26(4):1092–1104.
 77. Pustina D, Coslett HB, Ungar L, et al. Enhanced estimations of post-stroke aphasia severity using stacked multimodal predictions. *Hum Brain Mapp*. 2017;38(11):5603–5615.
 78. Ingram RU, Halai AD, Pobric G, Sajjadi S, Patterson K, Lambon Ralph MA. Graded, multidimensional intra- and intergroup variations in primary progressive aphasia and post-stroke aphasia. *Brain*. 2020;143(10):3121–3135. 10.1093/brain/awaa245.