1　　**Lesion site and therapy time predict responses to a therapy for anomia after stroke:**

2　　**a prognostic model development study**

3

4　　Thomas M.H. Hope[1,2]*, PhD, Davide Nardo[1,3], PhD, Rachel Holland[4], PhD, Sasha

5　　Ondobaka[1], PhD, Haya Akkad[1] MSc, Cathy J. Price, PhD[2], Alexander P. Leff[1,5], PhD, Jenny

6　　Crinion[1], PhD.

7　　　1. Institute of Cognitive Neuroscience, University College London, UK.
8　　　2. Wellcome Centre for Human Neuroimaging, University College London, UK
9　　　3. MRC Cognition and Brain Sciences Unit, Cambridge University, UK
10　　　4. Division of Language and Communication Sciences, City University London, UK.
11　　　5. UCL Queen Square Institute of Neurology, London, UK

12

13　　　　　　　*Corresponding author; E-mail: t.hope@ucl.ac.uk

14

15　　**Short title: Predicting anomia treatment responses**

16　　**Corresponding author:** Thomas Hope, t.hope@ucl.ac.uk, Institute of Cognitive

17　　Neuroscience, 17-19 Queen Square, London WC1N 3AR.

18　　**References:** 36

19　　**Words:** (abstract): 297; (body): 4,795

20

21

22   **Abstract**

23   BACKGROUND: Stroke is a leading cause of disability, and language impairments (aphasia)

24   after stroke are both common and particularly feared. Most stroke survivors with aphasia

25   exhibit anomia (difficulties with naming common objects), but while many therapeutic

26   interventions for anomia have been proposed, treatment effects are typically much larger in

27   some patients than others. Here, we asked whether that variation might be more systematic,

28   and even predictable, than previously thought.

29   METHODS: 18 patients, each at least 6 months after left hemisphere stroke, engaged in a

30   computerised treatment for their anomia over a 6 week period. Using only: (a) the patients'

31   initial accuracy when naming (to-be) trained items; (b) the hours of therapy that they devoted

32   to the therapy; and (c) whole-brain lesion location data, derived from structural MRI; we

33   developed Partial Least Squares regression models to predict the patients' improvements on

34   treated items, and tested them in cross-validation.

35   RESULTS: Somewhat surprisingly, the best model included only lesion location data and the

36   hours of therapy undertaken. In cross-validation, this model significantly out-performed the

37   null model, in which the prediction for each patient was simply the mean treatment effect of

38   the group. This model also made promisingly accurate predictions in absolute terms: the

39   correlation between empirical and predicted treatment response was 0.62 (95%CI: 0.27, 0.95).

40   DISCUSSION: Our results indicate that individuals' variation in response to anomia treatment

41   are, at least somewhat, systematic and predictable, from the interaction between where and

42   how much lesion damage they have suffered, and the time they devoted to the therapy.

43

1. Introduction

44

45   Stroke is a leading cause of disability [1], and language impairments (aphasia) after stroke are

46   both common [2] and particularly feared [3]. Most stroke survivors with aphasia exhibit

47   anomia, a difficulty finding words when naming common objects [4], but while many

48   therapeutic interventions for anomia have been proposed, treatment effects typically vary,

49   substantially, from patient to patient [5]. Inter-individual variation in treatment responses is

50   ubiquitous in medicine (e.g. in psychiatry and pharmacology, respectively [6, 7]), but emerging

51   evidence suggests that variation in responses to therapy for aphasia after stroke might be more

52   systematic than previously thought [5, 8, 9]. To the extent that this is true, the implication is

53   that we can use pre-treatment (e.g. behavioural and / or brain imaging) data to explain and even

54   predict patients' likely treatment responses.

55   For example, we recently showed that a model derived from: (a) pre-treatment scores

56   on standardised cognitive and language tasks; and (b) lesion location data, derived from pre-

57   treatment structural MRI, could be used to predict 23 aphasic stroke patients' responses to a

58   treatment for acquired reading impairments (central alexia) [8]. Like the treatment considered

59   here, for anomia, this earlier treatment for alexia was a computerised application designed to

60   engage participants in massed practice of trained items at home, over a period of weeks. Using

61   stepwise forward feature selection, we selected specific predictors from the pre-treatment data

62   for entry into a multiple linear regression model, which explained over 90% of the variance in

63   patients' empirical treatment responses. This result is biased by over-fitting, because all of the

64   patients' data were used to select features, but even when the feature selection was nested

65   within each fold of a leave-one-out cross-validation process (i.e. removing the bias), the

66   resulting predictions were significantly correlated with patients' empirical treatment responses

67    $(r = 0.48, p < 0.05)$ [8].

3

68   In what follows, we add to the evidence that responses to aphasia treatment might be

69   predictable from patients' pre-treatment data. Here, we focus on a computerised treatment for

70   naming difficulties (anomia) that is the most common language impairment after stroke [4].

71   This treatment's effectiveness at the group level has already been verified [10]; here, we

72   attempt to explain and predict the same treatment effects at the individual level. Our main

73   hypothesis was that the individual patients' responses to the treatment are systematic and

74   predictable given where and how much lesion damage they had suffered. We tested this by

75   comparing predictions made by models derived from those data (alone or in combination

76   with their pre-treatment anomia severity, demographic data and the hours that they devoted to

77   the therapy), to the predictions made by a 'null' model, which simply predicts the mean

78   treatment response of its training sample. If (any of) the former are significantly more

79   accurate than the latter, the implication is that individual variation in responses to this

80   treatment is systematic and predictable, at least to some extent.

81

2. Methods

83   The current analysis employs pre-treatment: (a) demographic data (age at stroke onset, time

84   post-stroke at assessment, and sex); (b) patients' initial impairment severity (i.e. their

85   accuracies when naming to-be-treated items, pre-treatment; (c) the hours of therapy actually

86   undertaken; and (d) structural MRI, which we use to extract lesion location data. We use

87   these data to predict and explain patients' responses to therapy, measured as the absolute

88   change in naming accuracy, from pre- to post-treatment, for 'trained' items (i.e. items

89   practiced during the therapy).

90   The therapy was designed to engage participants in massed practice of object naming,

91   over a 6 week period at home. A variety of phonemic cues (e.g. an audio recording of the

4

92   object's name, or of the name's first phoneme) were presented concurrently with the picture

93   to be named during treatment to encourage error-reducing learning. The approach was both

94   effective and specific to spoken word production, significantly improving patients' overall

95   object naming accuracy and reaction time immediately post-treatment (unstandardized effect

96   size: 29% and 17%, respectively; Cohen's d: 3.45 and 1.83). Longer term gains in naming

97   were maintained three months later, though in this study we focus only on the immediate

98   gains made for items trained during the therapy.

99   **2.1 Participants**

100   The study participants were 18 right-handed native English-speakers, with normal

101   hearing, no history of psychiatric disease and no prior history of neurological disorder before

102   suffering a left-hemisphere stroke, causing language impairment (aphasia). Participants were

103   recruited either from an aphasia clinic, run by JC, or via the Predicting Language Outcomes

104   After Stroke (PLORAS) study, run by CJP, between 2009 and 2012. The study size was

105   arrived at via a power calculation based on the expected effect size of the treatment

106   considered.

107   Participants were only included if they had: (i) naming difficulties (anomia), as assessed

108   via the Boston Naming Test (cut-off <56); (ii) relatively preserved single-word

109   comprehension as assessed via the Comprehensive Aphasia Test (CAT) [11]; (iii) good

110   mono-syllabic word repetition as assessed via the Psycholinguistic Assessments of Language

111   Processing in Aphasia [12]; (iv) no speech apraxia as determined by the Apraxia Battery for

112   Adults [13]; and, (v) at least partially spared left inferior frontal cortex (thought to support

113   speech re-learning [10]). All gave written informed consent to take part in the study, which

114   was approved by the Central London Research Ethics Committee and conducted in

5

115 accordance with the ethical principles stated by the Declaration of Helsinki. A table of the

116 participants' key characteristics, reproduced from [10], is included in supplementary material.

2.2 Stimuli and procedure

118 The procedure for the treatment study [10] involved behavioural assessments and

119 neuroimaging data acquisition both pre- and post-treatment. Here, we use pre-treatment data

120 only, to predict treatment response, calculated as the change in the number of trained items

121 that patients could name correctly.

122 Stimuli were drawn from a pool consisting of 299 black and white line drawings of

123 objects adapted from the International Picture-Naming Project

124 (https://crl.ucsd.edu/experiments/ipnp/). All object names were monosyllabic, with a

125 consonant-vowel-consonant structure and high name agreement (e.g. 'car'). The treatment

126 employed 150 of the 299 stimuli: i.e. for each patient, there were 150 treated items and 149

127 untreated items. 54/150 to-be-trained items and 53/149 un-trained items were kept common

128 across all patients (for use in an FMRI experiment [10], which we do not consider here). The

129 remaining items (96/150 to be trained; 95 /150 to be untrained) were determined for each

130 patient on the basis of their individual pre-treatment naming performance (accuracy) on the

131 299 items, to match each patient's pre-treatment performance on treated and untreated lists,

132 respectively.

133 After baseline assessment and pre-treatment structural MRI, patients were given a laptop

134 and asked to complete a minimum of two hours of naming practice 5 days a week, over a six-

135 week period. The pictures and auditory cues were presented using the 'StepByStep' aphasia

136 treatment software (http://www.aphasia-software.com). The naming practice was designed to

137 be completed in an error-reducing manner [14]. For example, in naming a picture of a car the

6

138   patient was asked to name it three times: (i) after a whole word auditory cue /ka:r/; (ii) after

139   an initial phonemic cue /ka/; (iii) after a whole word cue again. Only then would the patient

140   proceed to the next item to be named. Patients completed on average a total of 73 ($\pm$ 25)

141   hours of naming practice: i.e. within one standard deviation of the mean therapy dose found,

142   in a meta-analysis of aphasia treatment studies [15], to improve patients' communicative

143   ability. After the six-week period, patients were assessed again exactly as at baseline. Naming

144   accuracy was scored according to the standardized Comprehensive Aphasia Test guidelines

145   [11]. Our analyses here are separately focused on absolute change in naming accuracy (from

146   pre- to post-treatment) on the 150 treated items.

147   **2.3 Imaging acquisition and analysis**

148   The same scanner and hardware were used for the acquisition of all images. Whole-brain

149   imaging was performed on a 3 T Siemens TIM-Trio system (Siemens) at the Wellcome

150   Centre for Human Neuroimaging. Lesion images were derived from structural MRI using the

151   Automatic Lesion Identification toolbox [16], and then double-checked for accuracy by a

152   researcher experienced in manual lesion-tracking (DN), , working on individual axial slices.

153   Lesion data were then encoded as lesion load in a series of 398 anatomically defined

154   regions of interest, derived from four publicly available atlases (two focused on grey matter

155   and two focused on white matter) [17-20]. Where regions were represented in probabilistic

156   format, they were re-encoded as binary images at a 50% threshold. For each region, lesion

157   load was calculated as the number of ($2mm^3$) voxels shared by the lesion and the region,

158   divided by the total number of voxels in that region. Notably, there was significant overlap

159   between these regions, across atlases. Rather than deciding *a priori* what the best or most

160   useful atlas might be, our goal was simply to reduce the dimensionality of the lesion data in a

161   manner that retained an explicit link with familiar brain regions and / or tracts.

7

162

2.4 Modelling Methods

164 Our key aim here was to assess whether individual patients' treatment responses could be

165 predicted from pre-treatment data alone. Here, we define 'treatment responses' as the

166 absolute change in patients' naming accuracies from pre- to post-treatment.

167 Treatment studies in this domain are resource intensive and typically involve massed

168 practice, so take time to complete. Like most others in the field, our sample is therefore

169 smaller (n=18) than is usually desirable when building predictive models, increasing the risk

170 of over-fitting. That risk is further increased because we have so much pre-treatment data to

171 consider, including behavioural data, and lesion data derived from structural MRI.

172 One way to manage this risk is via feature selection, as we employed in similar, previous

173 work [8]. But though successful, that work still revealed significant over-fitting, because our

174 in-sample results (using the whole dataset to select features) were so much stronger than our

175 out-of-sample results (i.e. nesting feature selection in cross-validation): R^2 (predicted

176 response, empirical response) = 0.94 (in-sample); 0.23 (out-of-sample). Accordingly, we took

177 a simpler approach in this work by using dimensionality reduction, rather than feature

178 selection, to manage the high dimensionality of the pre-treatment (behavioural and brain

179 imaging) predictors.

2.4.1 Predictive models

181 We used Partial Least Squares (PLS) regression, as implemented in Matlab 2019a, to develop

182 our models, using either: (a) demographic variables, including age at stroke onset, sex and

183 time post-stroke; (b) pre-treatment naming accuracy (i.e. measuring the initial severity of

184 each patient's anomia); and / or (c) lesion data, derived from pre-treatment structural MRI.

185   We additionally considered one further variable, both singly and in combination with the

186   other data: the hours of therapy actually completed by each patient. There were no missing

187   data for any patient for any of these variables. All predictor variables were standardised (z-

188   scored) prior to entry into models.

189   PLS regression is appropriate, here, because it employs dimensionality reduction

190   analogous to, but more efficient than, that implemented by Principal Components Analysis

191   (PCA): i.e. where PCA identifies latent variables which explain maximal variance in the

192   predictors, PLS regression identifies variables that explain maximal variance in the response

193   variable(s). PLS regression thus allows us to build potentially effective models that are (at

194   least somewhat) robust to irrelevant predictors, rather than excluding those predictors

195   explicitly.

196   The behavioural model employed 28 predictors: i.e. scores on our battery of pre-

197   treatment language and cognitive assessments (as described in detail in [10]). The lesion data

198   were encoded as described previously, into 398 lesion load variables: however, all patients

199   had left-hemisphere lesions, and in fact all patients had zero lesion load in 220/398 regions.

200   These were removed from the analysis (leaving 178 variables), but their removal had no

201   substantive effect on the results. We trained models employing predictors derived from each

202   data type separately, and all higher order combinations of data types.

203

204   **2.4.2   Model assessment and model comparison**

205   Predictive performance was assessed with cross-validation. We report results using 1,000

206   times 10-fold cross-validation here, but analyses employing different types of cross-

207   validation were substantially similar. Absolute measures of predictive performance are

208   suspect in small samples, so we assessed our models in relative terms, by comparing them to

9

209   an empty, or baseline (i.e. null) model, which simply predicts the average treatment response

210   for all patients in the group. This model reflects our null hypothesis, that treatment responses

211   are *not predictable at the individual level*, leaving group-level averages as the only recourse.

212   When our empirical models outperform the empty model, we reject the null hypothesis,

213   concluding that individual treatment responses are predictable, at least to some extent. We

214   compare models by recording the Root Mean Squared Error (MSE) for predictions made in

215   each of 1,000 repetitions of a 10-fold cross-validation. These folds are kept identical across

216   models, so the MSE values can be compared pair-wise.

217   Traditional paired tests are not appropriate on their own here because different partitions

218   of the data will create overlapping training datasets, which are therefore not independent of

219   each other. Accordingly, while we use the traditional, paired, non-parametric Wilcoxon

220   signed rank test to compare MSEs across models, we further threshold those statistics with

221   paired permutation test. The test construes the two vectors to be compared as having labels,

222   reflecting the models used to generate them. The null hypothesis is that those labels are

223   arbitrary, because the models' performance do not differ except by chance. We therefore

224   create a null distribution of paired test statistics by randomly permuting those labels *within*

225   *each pair*, and repeating the original paired (signed rank) test. If the original statistic is

226   extreme relative to the null distribution, we conclude that the performance difference between

227   the models is significant ($p < 0.05$) after a correction for FamilyWise Error (FWE).

228

2.4.3   Model interpretation

230   PLS regression models can be interpreted by examining the weights of each of their

231   components on each of the original variables. However, this approach can be challenging

232   when there are multiple components to consider, and because the sign of each component is

233   arbitrary: i.e. positive weights on a given component do not necessarily imply a positive

234   relationship between the highly weighted independent variables and the dependent

235   variable(s). We circumvent these issues with 'data perturbation'.

236   The data perturbation procedure involves permuting random subsamples of the empirical

237   independent variables and recording the effect of the perturbation on the model's predictions.

238   The PLS model beta weights themselves are fixed based on the original empirical data: our

239   goal is not to fit further models, but rather to better understand the relationships that have

240   already been encoded. We do this by: (a) permuting random subsets of the independent

241   variables; (b) observing the effect on the models' predictions; and (c) relating perturbed

242   variable values to the resulting predictions with mass univariate correlation analyses. The

243   resultant correlation coefficients approximate the influence that each variable has on the

244   model's predictions. We ran 1,000 iterations of the process per model, yielding a total sample

245   size of 18 (patients) * 1,000 (iterations), including both perturbed independent variables and

246   the resultant, predicted dependent variable (treatment response). Repeated analyses with this

247   number of iterations yielded very consistent coefficients for all of the models we report

248   across ten repetitions of 1,000 iterations of this process, all pairwise correlations between

249   derived weights on behavioural and lesion variables were >0.99.

250

251   **3. Results**

252   **3.1 Predictive performance**

253   Table 1 reports predictive performances (median and inter-quartile ranges of Mean Square

254   Errors, or MSEs) of models driven by all combinations of the data we considered. All but one

255   of the models that out-performed the null model, with lower MSEs, included lesion data. The

256   exception was a model including hours of therapy only, with a median MSE of 300 (IQR =

11

257   16): i.e. a very small difference relative to the null model, albeit a significant one (FWE

258   adjusted p < 0.05). The best combination was hours of therapy plus lesion data (MSE median

259   / inter-quartile range = 182 / 21), and indeed this was the only combination which improved

260   upon lesion data alone: see Table 1. The mean predictions of that best model, across the

261   1,000 repetitions, were strongly and significantly correlated with empirical treatment

262   responses (r = 0.62, p = 0.006, 95% CI = 0.27, 0.95): see Figure 1.

263

Figure 1: Predicted responses versus empirical responses, for the best model identified in Table 1 (lesion load variables + hours of therapy undertaken).
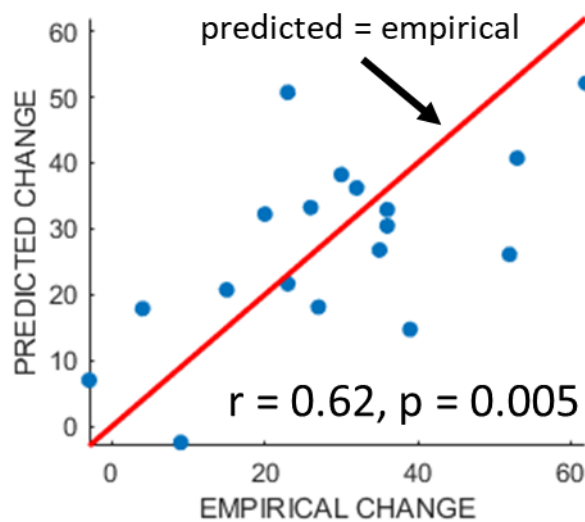
264

265

Data types	Median / IQR MSE
Null	303/16
Hrs (Therapy)	300/16
Initial (severity)	396/31
Demographics	321/24
Lesions	205/30
Initial + Hrs	364/30
Demographics + Hrs	316/26
Lesions + Hrs	**182/21**
Initial + Demographics	364/30
Initial + Lesions	253/29
Lesions + Demographics	267/21
Hrs + Initial + Demographics	355/30
Hrs + Initial + Lesions	220/32
Hrs + Demographics + Lesions	253/21

12

Initial + Demographics + Lesions	274/23
Hrs + Initial + Demographics + Lesions	261/23

266

Table 1: Data configurations and predictive performance, as assessed across the same 1,000 10-fold cross-validation runs. MSE = Mean Squared Errors of the model predictions; IQR = Inter-Quartile Range of the model predictions. These quantities are employed in preference to mean and standard deviation because MSEs typically have a Poisson distribution rather than a normal distribution. Lower MSEs imply more accurate predictions. The best model configuration is underlined (Hrs + Lesions): the most accurate predictions are derived from these data.

274

275

3.2 Interpreting the best model

Variable weights for the best models predicting change on treated items, using a combination of the hours of therapy undertaken and lesion location data, were calculated via data perturbation, as described in the Methods.

First, as expected, the best model predicted better improvement when patients devoted more hours to practice (r = 0.33). Regional weights for the lesion data in this model (i.e. taking therapy hours into account) are displayed in Figure 2, with the most negative weights (predicting lesser treatment benefit with more damage) in and around the left inferior frontal gyrus, and positive weights (predicting greater treatment benefit with more damage) in the middle, superior and anterior temporal lobe regions. Where voxels appear in two overlapping regions with different weights (e.g. we had one region covering the whole of the hippocampus and others covering only its cornu ammonis and dentate gyrus subfields), the most extreme of those two weights is displayed.

13

289
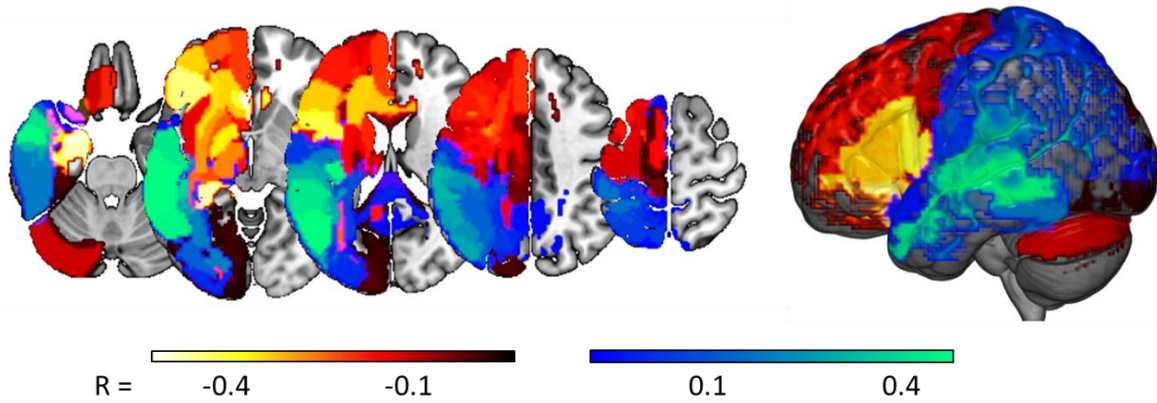
Figure 2: Relating lesion locations to predicted treatment responses. Correlation coefficients, derived via data perturbation, relating the lesion load in each of 177 regions, to treatment responses predicted by our best model (appending lesion data to the hours of therapy actually undertaken).

Notably, weights in many brain areas, including the auditory cortex and the superior, middle and anterior temporal lobes, are positive. The potentially curious implication here, is that more damage predicts larger treatment responses. Instead, we suggest that these positive regions are driven by the contingent distribution of the patients' lesions: more damage in those positively weighted regions implies less damage in the negatively weighted regions (where the latter make the more intuitive association between more damage and smaller treatment responses). As an illustration of this relationship, we considered area TE11 of the primary auditory cortex, where the strongest, positive weight was observed (0.47). Pairwise correlations, between lesion loads in this region and lesion loads in each of the other (177) regions under consideration, were very strongly correlated with the weights displayed in Figure 2, which were assigned to those regions by our best PLS regression model (r = 0.90): see Figure 3.

308

309
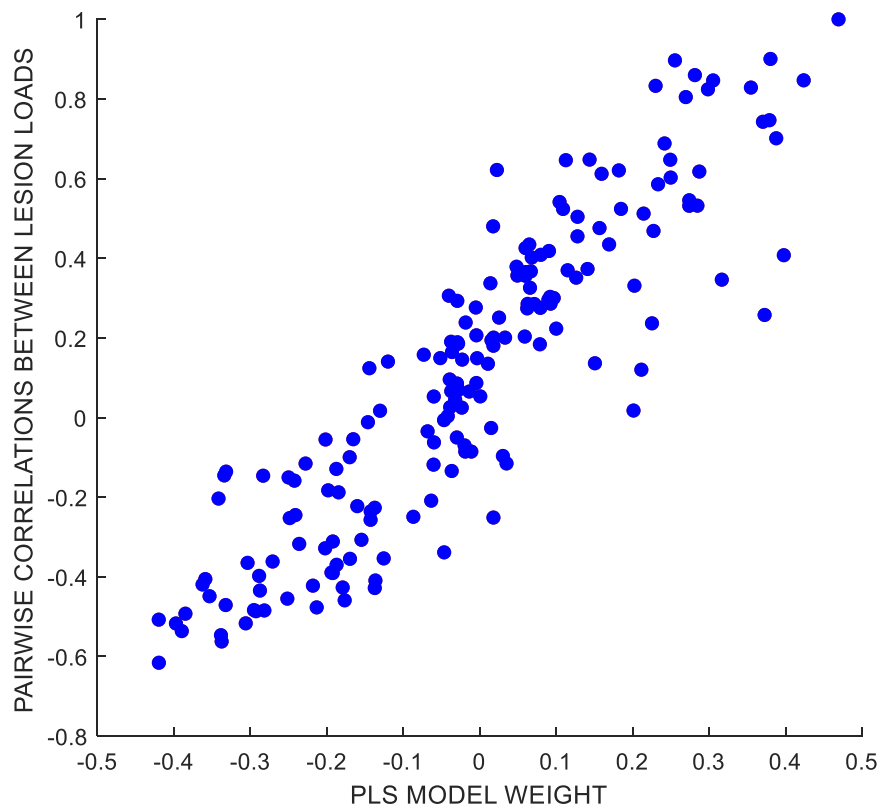
Figure 3: Scatter plot relating: (i) coefficients of the pairwise correlations between lesion load values in primary auditory cortex area TE11, and lesion loads in all of the 177 brain regions that we considered (y-axis); to (ii) the weights assigned to each of those same brain regions by our best PLS regression model, as derived via data perturbation (described in the Methods). The strong correlation between these two quantities implies that lesser lesion load in primary auditory cortex area TE11 serves as a proxy for greater lesion load in areas where that extra damage most strongly predicts poorer treatment responses.

317

4. Discussion

Recent results suggest that individual stroke patients' responses to aphasia treatment are to some extent systematic and predictable [8]. Our results add to this evidence, showing that responses to a behavioural treatment for anomia are at least somewhat predictable at the

15

322   individual level. We assessed models derived from demographic variables, and from pre-

323   treatment behavioural and lesion data. Models were derived via PLS regression, drawing on

324   efficient predictor dimensionality reduction and thus obviating the need for either algorithmic

325   or *a priori* feature selection. These models provide a sound way to establish at an individual

326   level whether pre-treatment data include signals that might be used to predict treatment

327   responses.

328   Many of the models we tested made significantly better predictions than those of a

329   baseline model, in which each patient's prediction was simply the mean response of the

330   group (see Table 1). However, the best model employed lesion location data, derived from

331   MRI, plus the hours of therapy undertaken by each patient. As hours of therapy alone has

332   very little predictive power, the results suggest that the benefit of increased therapy depends

333   on lesion location. This may explain why detecting therapy dose effects has been so

334   challenging [21]. Notably, we could not predict training effort, as indexed by hours of

335   therapy undertaken at the individual level, from any of the other data considered here.

336   Our best prognostic model, including lesion data and hours of therapy, is broadly

337   sensible. The negative weights assigned to the left inferior frontal gyrus, the hippocampus

338   and the cerebellum (more damage = less improvement) are consistent with prior work

339   emphasising the importance of the preservation of these regions in the response to aphasia

340   therapy (e.g. [22-24]). And the positive weights may best be explained as emphasising those

341   regions where more extensive damage predicts better preservation of the regions that appear

342   to support better responses to treatment (see Figure 3).

343   Notably, we did not employ any feature selection in this work: i.e. we did not attempt

344   to select the subset of lesion location variables that might best explain the patients' treatment

345   responses. This is a limitation of the current work, made necessary because feature selection

346   encourages over-fitting in small samples [8]: the only general way to circumvent this issue is

347   via external validation: testing the best model from this study in a second, completely

348   independent sample. But this is no simple endeavour, because the time and effort required to

349   run these studies is substantial, and we do not yet know how similar such a study would need

350   to be to that reported here. Does the treatment have to stay exactly the same? How much can

351   the inclusion criteria vary? Work to address these questions, by measuring how prognostic

352   models generalise across independent samples (e.g. as in [25]) and different therapy studies,

353   is ongoing.

354   Perhaps surprisingly, our models did not benefit from the addition, either of the initial

355   severity or the demographic data that we considered – suggesting that this treatment's

356   efficacy did not depend on the patients' ages, sex, time post-stroke or pre-treatment

357   impairment severity (once lesion location had been taken into account). Whether these null

358   results generalise in larger samples, is a question for future work. But our results do suggest

359   that pre-treatment structural neuroimaging (lesion data), in combination with treatment dose,

360   can be used to predict individual patients' therapeutic anomia intervention response. This is

361   consistent with prior results, suggesting that the individual responses to treatment for aphasic

362   stroke might interact with where and how much lesion damage individual patients have

363   suffered [8]. We hope that these results will encourage further attempts to explain and predict

364   inter-individual differences in treatment responses, with pre-treatment data, opening the way

365   for a more positive and personalised treatment approach for aphasia.

366

367   **ACKNOWLEDGMENTS**

371

372 **DATA**

373 The data described in this study is available to accredited researchers from JC, on request.

374

375

REFERENCES

1.   Corraini, P., et al., *Comorbidity and the increased mortality after hospitalization for stroke: a population-based cohort study.* J Thromb Haemost, 2018. **16**(2): p. 242-252.

2.   Engelter, S.T., et al., *Epidemiology of aphasia attributable to first ischemic stroke: incidence, severity, fluency, etiology, and thrombolysis.* Stroke, 2006. **37**(6): p. 1379-84.

3.   Lam, J.M. and W.P. Wodchis, *The relationship of 60 disease diagnoses and 15 conditions to preference-based health-related quality of life in Ontario hospital-based long-term care residents.* Med Care, 2010. **48**(4): p. 380-7.

4.   Laine, M. and N. Martin, *Anomia: Theoretical and clinical aspects*. 2013: Psychology Press.

5.   Crinion, J.T. and A.P. Leff, *Using functional imaging to understand therapeutic effects in poststroke aphasia.* Current Opinion in Neurology, 2015. **28**(4): p. 330-337.

6.   Helldin, L., et al., *Neurocognitive variability in schizophrenia spectrum disorders: relationship to real-world functioning.* Schizophrenia Research: Cognition, 2020. **20**: p. 100172.

7.   Kadiev, E., et al., *Role of pharmacogenetics in variable response to drugs: focus on opioids.* Expert opinion on drug metabolism & toxicology, 2008. **4**(1): p. 77-91.

8.   Aguilar, O.M., et al., *Lesion-site-dependent responses to therapy after aphasic stroke.* Journal of Neurology, Neurosurgery & Psychiatry, 2018.

9.   Lambon Ralph, M.A., et al., *Predicting the outcome of anomia therapy for people with aphasia post CVA: Both language and cognitive status are key predictors.* Neuropsychological Rehabilitation, 2010. **20**(2): p. 289-305.

10.   Nardo, D., et al., *Less is more: neural mechanisms underlying anomia treatment in chronic aphasic patients.* Brain, 2017. **140**(11): p. 3039-3054.

11.   Swinburn, K., Porter, G., and Howard, D., *Comprehensive Aphasia Test*. 2004: Psychology Press.

12.   Kay, J., R. Lesser, and M. Coltheart, *Psycholinguistic assessments of language processing in aphasia (PALPA).* Hove, UK: Laurence Erlbaum Associates, 1992.

13.   Dabul, *Apraxia battery for adults (second ed.).* 2000, Austin, Texas: Pro-Ed.

14.   Fillingham, J.K., K. Sage, and M.A. Lambon Ralph, *The treatment of anomia using errorless learning.* Neuropsychological Rehabilitation, 2006. **16**(2): p. 129-154.

15.   Bhogal, S.K., R. Teasell, and M. Speechley, *Intensity of aphasia therapy, impact on recovery.* Stroke, 2003. **34**(4): p. 987-93.

16.   Seghier, M.L., et al., *Lesion identification using unified segmentation-normalisation models and fuzzy clustering.* Neuroimage, 2008. **41**(4): p. 1253-66.

17.   Tzourio-Mazoyer, N., et al., *Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain.* NeuroImage, 2002. **15**(1): p. 273-289.

18.   Hua, K., et al., *Tract probability maps in stereotaxic spaces: analyses of white matter anatomy and tract-specific quantification.* Neuroimage, 2008. **39**(1): p. 336-47.

19.   Oishi, K.F., A. F.; van Zijl P. C. M.; Mori, S., *MRI Atlas of Human White Matter*. Vol. 2. 2011.

20.   Eickhoff, S.B., et al., *A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data.* Neuroimage, 2005. **25**(4): p. 1325-35.

21.   Harvey, S., et al., *Dose effects in behavioural treatment of post-stroke aphasia: a systematic review and meta-analysis.* Disability and Rehabilitation, 2020: p. 1-12.

22.   Sebastian, R., et al., *Cerebellar tDCS: A Novel Approach to Augment Language Treatment Post-stroke.* Frontiers in human neuroscience, 2017. **10**: p. 695-695.

23.   Mattioli, F., et al., *Early Aphasia Rehabilitation Is Associated With Functional Reactivation of the Left Inferior Frontal Gyrus.* Stroke, 2014. **45**(2): p. 545-552.

423   24.   Meinzer, M., et al., *Integrity of the hippocampus and surrounding white matter is correlated*
424   *with language training success in aphasia.* Neuroimage, 2010. **53**(1): p. 283-90.
425   25.   Loughnan, R., et al., *Generalizing post-stroke prognoses from research data to clinical data.*
426   NeuroImage: Clinical, 2019. **24**: p. 102005.

427

428

429   Supplementary Table S1: Demographic and clinical data of the patients

Patient ID	Sex	Age	Lesion volume (cm³)	Months post-stroke	BNT	CAT	PALPA 9	PALPA 8	Hours of training
P1	M	64	171	78	47	15	20	6	40
P2	F	49	44	17	12	15	21	6	31
P3	M	54	294	78	14	11	10	0	77
P4	M	41	234	65	28	14	24	8	116
P5	M	49	144	57	34	15	17	2	50
P6	M	66	109	61	52	15	24	6	63
P7	F	44	82	72	34	14	24	10	59
P8	M	54	95	34	35	15	24	8	70
P9	M	67	341	47	42	14	24	9	85
P10	M	41	75	8	23	13	23	8	89
P11	M	63	139	264	51	15	24	9	81
P12	M	47	314	52	16	15	22	6	77
P13	M	56	150	40	1	14	18	2	61
P14	F	60	104	121	27	13	22	7	120
P15	M	41	114	18	42	14	21	3	43
P16	F	21	155	33	18	15	20	3	108
P17	F	47	161	53	9	9[a]	12	0	76
P18	F	43	165	5	21	15	23	1	67
Mean (SD)		50 (12)	161 (84)	61 (58)	28 (15)	14 (2)	21 (4)	5 (3)	73 (25)
			Max score possible		60	15	24	10	

430

431

432   TRIPOD checklist for prediction model development.

Section/Topic			Checklist Item	Page
Title and abstract				
Title	1		Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	X
Abstract	2		Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	X
Introduction				
Background and objectives	3a		Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	x
	3b		Specify the objectives, including whether the study describes the development or validation of the model or both.	x
Methods				
Source of data	4a		Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	x
	4b		Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	X
Participants	5a		Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	x
	5b		Describe eligibility criteria for participants.	X
	5c		Give details of treatments received, if relevant.	X
Outcome	6a		Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	X
	6b		Report any actions to blind assessment of the outcome to be predicted.	X
Predictors	7a		Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	X
	7b		Report any actions to blind assessment of predictors for the outcome and other predictors.	X
Sample size	8		Explain how the study size was arrived at.	x
Missing data	9		Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	x
Statistical analysis methods	10a		Describe how predictors were handled in the analyses.	X
	10b		Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	X
	10d		Specify all measures used to assess model performance and, if relevant, to compare multiple models.	X
Risk groups	11		Provide details on how risk groups were created, if done.	X
Results				
Participants	13a		Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	X
	13b		Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	X
Model development	14a		Specify the number of participants and outcome events in each analysis.	X
	14b		If done, report the unadjusted association between each candidate predictor and outcome.	X
Model specification	15a		Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	X
	15b		Explain how to the use the prediction model.	X
Model performance	16		Report performance measures (with CIs) for the prediction model.	X
Discussion				
Limitations	18		Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	X
Interpretation	19b		Give an overall interpretation of the results, considering objectives, limitations, and results from similar studies, and other relevant evidence.	X
Implications	20		Discuss the potential clinical use of the model and implications for future research.	x
Other information				
Supplementary information	21		Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	x
Funding	22		Give the source of funding and the role of the funders for the present study.	x

433