MULTIVARIATE LESION SYMPTOM MAPPING FOR PREDICTING TRAJECTORIES OF RECOVERY FROM APHASIA

By

Deborah Faith Levy

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Hearing & Speech Sciences

August 13, 2021

Nashville, Tennessee

Approved:

Dr. Stephen Wilson, Ph.D.

Dr. Michael de Riesthal, Ph.D.

Dr. Melissa Duff, Ph.D.

Dr. Ipek Oguz, Ph.D.

To my grandfather, Marvin Levy, whose inexhaustible wit and curiosity at the age of 97 serve as a constant reminder that there will never be a good excuse to stop learning.

ACKNOWLEDGMENTS

There's an old adage, attributed to many potential sources but with which I am most familiar through my grandfather: "If you're the smartest person in the room, you're in the wrong room". In alignment with this quote, I'd like to acknowledge some of the numerous brilliant individuals with whom I've been lucky enough to share rooms over the last five years, and without whom this work would not have been possible.

First and foremost, to my advisor, Dr. Stephen Wilson: thank you for your clear and effective guidance, your systematicity and (to borrow a term) perspicacity, your approachable demeanor, your patience and understanding, and your trust in me to learn to do things I didn't know I was capable of. Working with you made graduate school an altogether pleasant (in addition to an extremely educational) experience, and I couldn't have asked for a better advisor.

To my lab mates in the Language Neuroscience Lab: thank you for your incisive questions, your expertise and guidance (both clinical and scientific), your kindness, your support, and your incredibly hard work. To Sarah, Jillian, Caitlin, Maysaa, and Marianne, this dissertation quite literally could not have happened without your years of effort. Thanks to all of you, as well as to Melodie, Anna, and Yev, for your camaraderie as well as your scientific counsel.

To my various committee members and mentors across multiple projects—Drs. de Riesthal, Duff, Oguz, Ramachandran, Stecker, Chang, Morgan, and Booth—thank you for your valuable insights and instructive feedback. The quality of my work was increased greatly by your oversight.

To Dominique and the members of the Aphasia Group of Middle Tennessee: thank you for welcoming me into your incredibly special community with open arms, and for your patience and openness towards me despite my lack of clinical training. It was an absolute honor getting to know you all, and I will never forget everything you've taught me about resilience, humor, the value of community, and the multitudinous ways a person can express who they are.

To my Hearing and Speech Sciences cohort: thank you for your steadfast support and inspiring successes. I am so amazed by all that you've accomplished! Thank you in particular to Sarah and Natalie for your friendship, and your constant willingness to dissect all the eccentricities of academic life with me.

To my friends—in particular Carley, Emily, Meredith, Alina, and Nic—thank you for reminding me that there is life outside of graduate school, and for always making that life so much fun.

To my parents and big sibling Rowan: thank you for the unending love and support that made me believe I was capable of anything. I am so lucky to have had you all as cheerleaders, both throughout this process and throughout my life.

To my fiancé, Isaac: thank you for flying to Nashville at 4:00 AM twice a month for three years to come visit me while we were long distance; thank you for driving cross-country to move here immediately after defending your Master's thesis, despite the fact that you hadn't slept in a week; thank you for letting me ask you endless methodological questions and tolerating it when I inevitably yelled at you for answering them too intelligently; thank you for loving me unconditionally, even at my messiest, grouchiest, least confident, and most covered in Hot Cheeto dust. Thank you for still wanting to marry me despite it all.

And finally, to all the participants in this study and their loved ones who made this work possible: thank you for your time, your tenacity, your faith in the scientific process, and your willingness to let us learn from you. This work is the result of nothing if not your strength and generosity.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Literature Review

Aphasia is an acquired disorder of communication that results from injury to regions of the brain that support language (NIDCD, 2015). While aphasia can result from a number of etiologies, including traumatic brain injury, brain tumor, and neurodegenerative disease, it is most commonly caused by stroke (Crinion et al., 2013; Shuster, 2018; Mayo Clinic Staff, 2020).

Of the approximately 800,000 individuals who have a stroke in the United States each year (Benjamin et al., 2017), anywhere from 21-38% of them present with aphasia acutely (Pedersen et al., 1995; Laska et al., 2001; Pedersen et al., 2004; Flowers et al., 2016). The vast majority of cases of aphasia result from infarct to the left hemisphere (Berthier, 2005) and most in the territory of the middle cerebral artery (MCA) (Fridriksson et al., 2018). Maximal recovery of language function tends to occur within the first three months (Kertesz and McCabe, 1977; Pedersen et al., 1995; Laska et al., 2001; Wilson, 2019), and while it was once widely believed that recovery plateaus after the first six months post-stroke (Hersh, 1998; Elman, 2016; Johnson et al., 2019), recent work has begun to seriously challenge this adage (Fridriksson and Hillis, 2021). However, there remains a significant amount of inter-individual variability in both short- and long-term recovery (Lazar and Antoniello, 2008; Dunn et al., 2016; Hope et al., 2017); some individuals will recover their language near-completely in the first few months following stroke, while others may be forced to deal with the resulting language impairment for the rest of their lives (Lendrem and Lincoln, 1985; Pedersen et al., 1995). Due to the devastating consequences of aphasia on quality of life (Spaccavento et al., 2013; Musser et al., 2015), the burden it imposes on loved ones and caregivers (Hilton et al., 2014; Musser et al., 2015), and the clinical desire to provide clear and effective guidance on navigating the recovery process (Doogan et al., 2018), individuals with aphasia, their loved ones, and their treating clinicians alike are eager to better understand whether and in what ways language can be expected to recover in individuals who have experienced aphasic

stroke (Worrall et al., 2011; Bright et al., 2013; Hope et al., 2013).

1.1 Predicting recovery from aphasia

When aiming to predict recovery from aphasia, there are three main classes of information that have been commonly considered; linguistic, demographic, and stroke-related. Linguistic predictors, generally, refer to the nature of language or aphasia immediately following stroke, including overall severity, particularities of language symptoms, and the presence and extent of therapeutic intervention. Demographic predictors describe characteristics of individuals prior to the onset of their aphasia, including information about age, gender, or years of education. Finally, stroke-related predictors cover information about the cerebrovascular event that caused the aphasia, such as the extent and location of the lesion or the subtype of stroke experienced. A summary of findings on aphasia recovery with reference to these three general classes of predictors are presented below.

1.1.1 Linguistic predictors

Linguistic predictors are commonly believed to be among the most effective predictors of aphasia recovery, with initial severity in particular being associated with poorer long-term outcomes (Lazar et al., 2010; Osa García et al., 2020). Patients who present with milder language impairments tend to recover more quickly, and often resolve to better language function in the long-term (Pedersen et al., 1995). Those with more severe presentations acutely unfortunately tend to remain significantly impaired linguistically (Kertesz et al., 1979; Lazar et al., 2010). To the extent that they do recover, more severe, non-fluent forms tend to evolve to milder, more fluent forms over time (Kertesz and McCabe, 1977; Laska et al., 2001; Pedersen et al., 2004; Bakheit et al., 2007), with cases reported of individuals transitioning from globally aphasic to mildly anomic within a year (Pedersen et al., 2004). Earlier and more intensive administration of speech-language therapy have been associated with more positive outcomes in recovery (Bhogal et al., 2003; Breitenstein et al., 2017; Fridriksson and Hillis, 2021; Ali et al., 2021), though these findings have at times been challenged (Lincoln et al., 1984; Laska et al., 2001); for discussion of some open questions and

new developments regarding the role of speech and language therapy in post-stroke aphasia, see Doogan et al. (2018) or Fridriksson and Hillis (2021).

It is commonly believed that recovery from aphasia, like other post-stroke deficits, abides by a "proportional recovery rule", which states that individuals with stroke tend to recover some fixed proportion of their lost function, in general about 70% (Lazar et al., 2010; Marchi et al., 2017). However, the legitimacy of this rule is disputed from a statistical perspective, as the strong baseline-change correlations that appear to support proportional recovery have been shown to occur even in simulated data with no true association between baseline and outcome scores (Hope et al., 2019; Hawe et al., 2019; Bonkhoff et al., 2020; Bowman et al., 2021). However, the general notion that initial severity is highly predictive of long-term outcomes is largely uncontroversial.

1.1.2 Demographic predictors

Demographic variables refer to characteristics of patients that are unrelated to their stroke, such as age, sex, race, or years of education. The extent to which this class of variables is predictive of language and recovery from aphasia is still debated. While there do not generally appear to be significant independent effects of sex on aphasia outcomes (Pedersen et al., 1995; Wallentin, 2018; Gerstenecker and Lazar, 2019) some studies have suggested that females may have slightly better recoveries than males (Basso et al., 1985; Pizzamiglio et al., 1985). Many studies have noted no effect of age on aphasia recovery (Lendrem and Lincoln, 1985; Lazar and Antoniello, 2008; Ellis and Urban, 2016), but some studies have suggested that younger individuals show better recoveries than older individuals (Pickersgill and Lincoln, 1983; Laska et al., 2001). Years of education may have a modest positive relationship with aphasia recovery (González-Fernández et al., 2011), but this effect is debated (Connor et al., 2001; Lazar and Antoniello, 2008); interestingly, however, individuals with more education have been shown to be less likely to present with aphasia at all following stroke (González-Fernández et al., 2011; Watila and Balarabe, 2015). While socioeconomic status (SES) has been shown to be a good predictor of initial aphasia severity, such that individuals with lower SES tend to present with more severe aphasia acutely, the rate of recovery

from aphasia was not shown to differ across socioeconomic groups (Connor et al., 2001). Ethnicity does not appear to independently contribute to outcomes in aphasia recovery (Holland et al., 1989), although there is very little work investigating the subject; though race-related differences in neurocognitive outcomes following stroke have been reported, they seem to be mediated by other factors (Horner et al., 2003; Johnson et al., 2017).

In line with this concept, it is important to note that even when demographic variables show associations with language outcomes, this does not inherently mean that they are *mechanisms* of recovery in and of themselves. Instead, demographic variables may covary with unmeasured variables that more directly drive recovery (e.g., years of education may covary with cognitive reserve that allows for more successful compensatory strategies; see Umarova et al., 2019), or covary with nuisance variables that can disguise recovery when it is present (e.g. age may covary with factors like hearing and vision loss which could impact assessment results without truly reflecting language abilities; see Wertz and Dronkers, 1990). In summary, prior work has revealed very few unambiguous relationships between demographic variables and language recovery following stroke.

1.1.3 Lesion-related predictors

Lesion and stroke-related factors appear to be the most reliably predictive determinants of aphasia recovery (Watila and Balarabe, 2015), although the relationship between lesions and language is far from straightforward. Whether or not stroke subtype—that is, ischemic or hemorrhagic—has effects on aphasia recovery remains somewhat unclear, with some studies showing no association between stroke type and recovery (Paolucci et al., 2003; Salvadori et al., 2020) and others suggesting that survivors of hemorrhagic stroke have better language outcomes in the long-term (Jung et al., 2011). Lesion size and lesion location, however, strongly associate with language outcomes, such that larger lesions are strongly associated with poorer recoveries, and lesions in peri-Sylvian regions in the generally language-dominant left hemisphere are associated with poorer language outcomes than lesions to other sites (Naeser and Palumbo, 1994; Gerstenecker and Lazar, 2019).

Such findings were extremely enlightening in the early days of aphasia research, when the existence of "language regions" at all was a source of significant debate (see the next section for a review of this historic period and the resulting classical model of language). However, at this point in history, the general assertion that extensive damage to known language regions will have negative effects on language leaves much to be desired in terms of its prognostic power. Contemporary research has demonstrated that a significant number of patients do not conform to theoretical expectations for their language based on their lesion location (Mohr, 1976; Basso et al., 1985; Willmes and Poeck, 1993; Berthier, 2001; Yourganov et al., 2015), and that information that is not directly attributable to the size or location of the infarct, such as hypoperfusion, diaschisis, and abnormal functional activation, also make significant contributions to language impairment (Olsen et al., 1986; Metter et al., 1989). Even so, lesion-related factors remain perhaps the most powerful predictors of aphasia recovery (Watila and Balarabe, 2015), and have formed the basis for a widely used method in research on language and aphasia, lesion symptom mapping (Wilson, 2017; Forkel and Catani, 2018). Findings from the lesion symptom mapping literature will be discussed in detail in later sections of this document.

Taken together, all of this information suggests that recovery from aphasia is a complex and multicomponent problem, one that is influenced by many interacting variables and that is extremely difficult to predict using overly simplistic models, even those that take the most robust predictors—lesion-related factors—into account. However, it would be unwise to embark upon any serious discussion of factors contributing to aphasia and aphasia recovery without first discussing the foundational models that shaped the field. The next section will explore the long and influential history of lesion-deficit models in the study of aphasia, to put into context the more contemporary methods and findings that inform the current work.

1.2 Lesion-symptom mapping (LSM)

1.2.1 An abridged history

The ability to reliably map functions of language onto neural regions has been sought after for centuries (Tesak and Code, 2008), but the search perhaps began in earnest in the early 1860's, with a series of debates among Parisian physicians leading to key breakthroughs that still influence aphasia practice today (Leblanc, 2019). In these days prior to computerized neuroimaging, such investigations into the brain-language relationship could not occur until autopsy, when the locus of neural damage could be assessed and retroactively mapped onto language symptoms displayed by a patient during their lifetime. Using this method, physicians and in-laws Jean-Baptiste Bouillaud and Ernest Auburtin had noticed a pattern in their patients: those who suffered from deficits in their language almost universally came to autopsy with lesions to their frontal lobes. Bouillaud and Auburtin became so convinced of the relationship between frontal lobe damage and language impairment that they put forth a challenge to their peers: anyone who could show them a patient with language impairment who did not have frontal lobe damage would be rewarded with 500 Francs (in the case of Bouillaud) and a public renunciation of their beliefs (in the case of Auburtin). Not long after, a new patient, Leborgne, came to the care of neurologist Paul Broca, suffering from an acute gangrenous infection in addition to a chronic and severe impairment of expressive language. Leborgne quickly became the test case for the challenge, and he did not disappoint: upon coming to autopsy less than a week later, Leborgne's brain showed extensive degeneration centered on—as judged by Broca—the third frontal convolution. Surprisingly, the fact that it was Leborgne's *left* frontal lobe that was impacted did not strike Broca as particularly important until several years later, when the pattern among later autopsy cases became too striking to ignore. It was at that time that his famous theory of language lateralization—"we speak with the left hemisphere"—was established (Berker et al., 1986). Interestingly, a relatively unknown father-son pair of physicians, Marc and Gustave Dax, actually arrived at this same conclusion several years earlier (Joynt and Benton, 1964), but attempts to publish the findings were met with little success. The final Dax paper ended up being published in the same year as Broca's—earlier, in

fact—but was greatly overshadowed; as Levelt (2013) puts it, "It is not a modern phenomenon that the process of peer review can occasionally fail" (p. 61).

In the intervening years, the new study of aphasiology spread across Europe and found a new home in Germany, where a young physician, Carl Wernicke, took up its study. Wernicke rejected the notion that language could have a single circumscribed seat in the brain, deeming it highly improbable in the context of his understanding of speech development and his findings of comprehension-based impairments to language following posterior damage (Wernicke, 1875). He instead proposed a connectionist model of language, in which multiple nodes and connections in the left hemisphere operated in tandem to produce fluent language production and comprehension. These nodes and connections could, he suggested, each be lesioned with varying effects on language, resulting in a list of seven language syndromes which could theoretically occur as a function of the language centers lesioned. This model was clarified, schematized, and demonstrated in patients by Wernicke's student, Ludwig Lichtheim (Lichtheim, 1885), and finally crystallized in 1886 into what we generally now refer to as the "classic model" (Wernicke, 1886).

The classic or "house" model of aphasia (see Fig. 1.1) consisted of three primary centers, referred to by Lichtheim as (A) the center for auditory images, (M) the center for motor images, and (B), the center for concepts; the remaining two centers were (*m*) and (*a*), referring to the articulatory system and the auditory sensory system, respectively. Lesions to the centers and commissures and their associated language syndromes as described in Wernicke (1886) are summarized in Table 1.1.

Based on autopsy, centers A and M were localized to the left inferior frontal gyrus and the left superior temporal gyrus, respectively, with center B theorized to consist of a number of distributed cortical areas dedicated to representing sensory and conceptual information. This anatomically-based connectionist model provided explanatory power both to predict lesion location from aphasic syndrome, as well as to bring together Broca's expressive aphasia and Wernicke's sensory aphasia, two distinct syndromes which had not always been considered in tandem as equally "linguistic" (as many early physicians considered receptive aphasias to be impairments of intelligence, rather

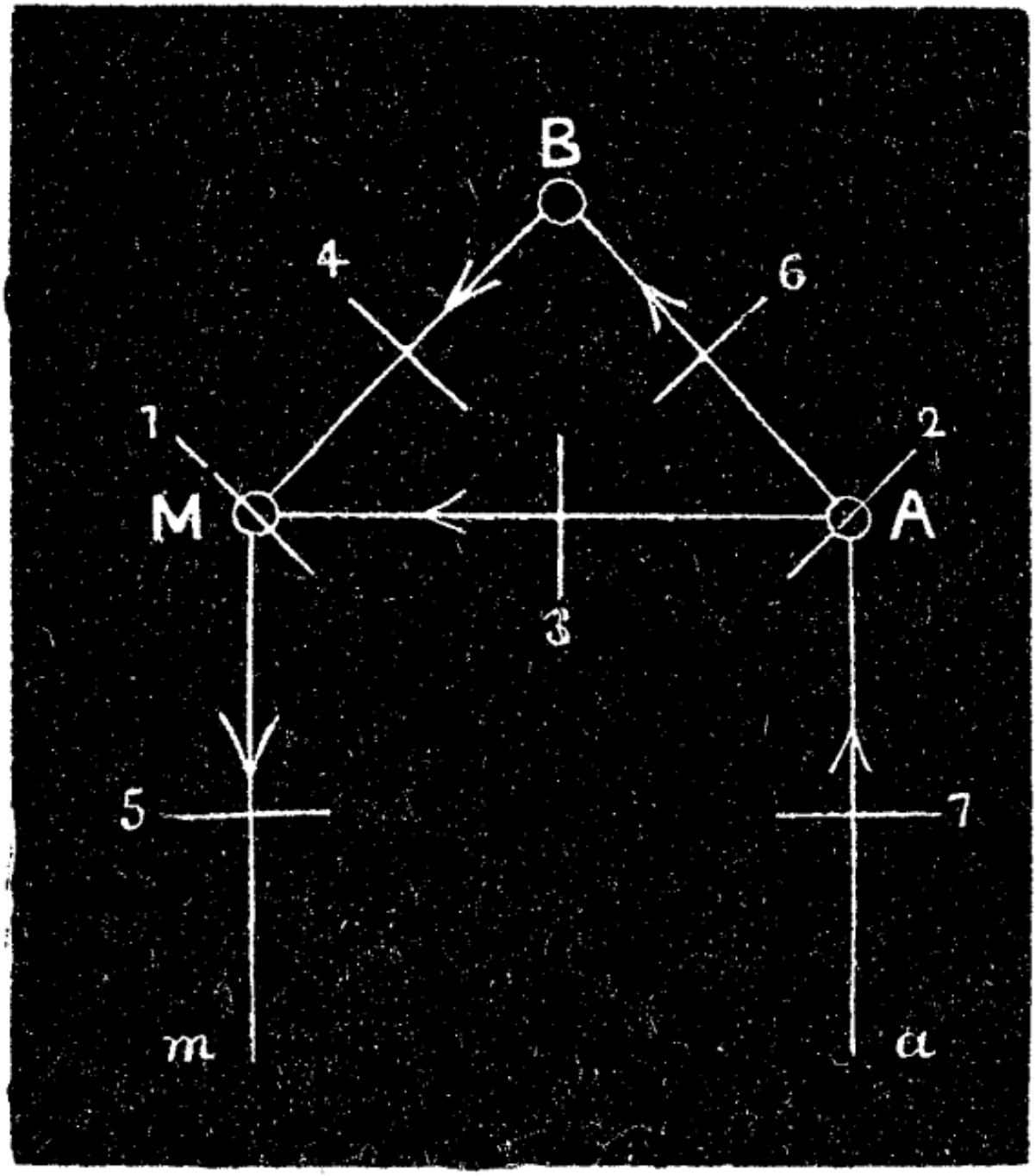

FIG. 1.

Figure 1.1: Reproduced from Lichtheim (1885). The classic model of aphasia as shown by Lichtheim (1885). See Table 1.1 for further detail.

Lesion	Wernicke Nomenclature	Symptoms
Commissure *a*–A	Subcortical sensory aphasia	- Spontaneous speech in tact - Impaired comprehension/repetition
Center A	Cortical sensory aphasia	- Spontaneous speech with semantic paraphasias - Impaired comprehension/repetition
Commissure A–B	Transcortical sensory aphasia	- Spontaneous speech with semantic paraphasias - Repetition in tact - Impaired comprehension
Commissure A–M	Conduction aphasia	- Comprehension in tact - Semantic paraphasias
Commissure B–M	Transcortical motor aphasia	- Repetition/comprehension in tact - Impaired spontaneous speech
Center M	Cortical motor aphasia	- Comprehension in tact - Impaired spontaneous speech/repetition
Commissure M–*m*	Subcortical motor aphasia	- Comprehension/"word-concept" in tact - Impaired spontaneous speech/repetition

Table 1.1: Theorized relationships between lesions and symptoms in Wernicke (1886).

than language; Tesak and Code, 2008). The model was initially celebrated for its incisiveness, but even as early as 40 years later was mocked for its oversimplification of aphasic syndromes; Head (1926) wrote, "Lichtheim's paper...reads like a parody of the tendencies of the time...it enabled...an easy dogmatism, but serious students could not fit these conceptions of aphasia to the clinical phenomena" (p. 65). Though this critique may seem harsh in the context of the model's enduring influence, it has indeed become clear over time that the ability of the classic model to predict aphasia outcomes in the real world is far from infallible.

A number of issues with the early autopsy method can account for some of the model's shortcomings. First of all, characterizations of language function in early autopsy studies were broad and unsystematic, depending heavily on what a given physician, often with specific a priori hypotheses and without specific linguistic training, deemed interesting or important to test (Heilman, 2015). Additionally, the majority of historical conclusions relied on subjective judgments of where the bounds of lesions truly laid, as there was no reliable means for standardizing lesion location across individual cases (Dronkers et al., 2007). Finally, the need to wait for autopsy meant neural findings were non-contemporaneous with symptoms; the state of the brain at death provided only

a single time point, precluding the ability to make inferences about how disease processes over the lifetime may have dynamically impacted language at different stages of illness or recovery (Mohr, 1976). Due in part to such criticisms, connectionist approaches fell out of favor for several decades, rejected in favor of more holistic models of language and intelligence in which all neural regions worked together as a system, with no special role for particular nodes or connections in particular processes. Yet the classic model remains the basis for a significant portion of modern-day practice and study around aphasia.

The classic model was revived in the late 1960's by Norman Geschwind, a Boston neurologist who felt compelled to revisit the classical literature following discoveries about the behavioral effects of severing the corpus callossum in animals (Geschwind, 1965). He brought to light several cases that conformed to the predictions of the classic model, and in concert with Harold Goodglass cemented the neo-classical Boston classification as a unified theory of brain and language. In this school of thought, the primary components of the language system were Broca's area (the left pars opercularis and triangularis), involved in the production of language; Wernicke's area (the posterior superior temporal gyrus), involved in the comprehension of language; the arcuate fasciculus, connecting Broca's and Wernicke's areas and involved in connection/feedback between productive and receptive processes; and the angular gyrus, involved in cross-modality associations (Geschwind, 1965; Tesak and Code, 2008). Damage to these areas and their connections could result in a set of syndromes very similar to those described by Wernicke and Lichtheim; Broca's aphasia, Wernicke's aphasia, conduction aphasia, transcortical sensory/motor aphasia, and anomia. This Boston classification formed the basis for both the Boston Diagnostic Aphasia Examination (BDAE) (Goodglass et al., 2001) and the Western Aphasia Battery (WAB) (Kertesz, 2007), two of the most commonly used aphasia assessments today (Spreen and Risser, 2003; Patterson, 2015).

Not long after the neo-classical revival, the advent of computerized neuroimaging brought with it new opportunities for comparatively more objective and immediate assessments of lesion locations in relation to clinical syndromes. Initially, the majority of these investigations were completed using lesion overlays, in which the lesions of patients with similar language syndromes

or symptoms were overlaid on top of each other to determine areas of maximal overlap that reliably associated with language outcomes. Results from the overlay method supported some postulates of the classic model (Kertesz et al., 1977; Naeser and Hayward, 1978), but raised a number of issues with it as well; for example, exclusive damage to "Broca's area"—anatomically, the pars opercularis—very rarely lead to a lasting Broca's aphasia (Mohr, 1976); a significant portion of patients did not conform to theoretical expectations for their language based on their lesion location (Basso et al., 1985), and language deficits often attributed to specific brain areas could arise from damage to different brain areas entirely (Berthier, 2001). Willmes and Poeck (1993) even claimed that "no unequivocal association between type of aphasia and localization of lesion" could be found among a cohort over 200 patients (p. 1527), further challenging the connectionist approach. Yet this new lesion overlay method was still plagued by some of the same issues as the autopsy method; judgments of the bounds of lesion were holistic in nature, and in most cases required the grouping of patients into aphasia subtypes into which they often only loosely fit, ignoring inter-individual variation in symptoms within those subtypes. In order to make precise, longitudinal and statistically tenable predictions about the relationship between lesions and language symptoms, both measures of language and measures of lesions would have to be appropriately quantified. A discussion of contemporary means by which first, language, and second, lesions, have been and are currently being quantified for the purposes of aphasia research is presented below.

1.2.2 Contemporary methods

1.2.2.1 Methods for quantifying language

As discussed above, the BDAE and the WAB, in their revised forms, are two of the most commonly used comprehensive aphasia assessments today. Both tests aim to assess aphasia severity and diagnose it taxonomically according to the Boston classification of aphasia. The WAB-R consists of four main subtests (Spontaneous Speech, Auditory Verbal Comprehension, Repetition, Naming and Word Finding) and four supplemental sections (Reading, Writing, Apraxia, and Constructional/Visuospatial/Calculation). The BDAE-3 consists of six sections: Conversational and Expository

Speech, Auditory Comprehension, Oral Expression, Reading, Writing, and Praxis. Both assessments collapse across scores on the subtests to arrive at one or more composite scores assessing overall severity and a subtype diagnosis. These tests are widely used, but can be time-consuming to administer in their full, most validated forms and have been criticized for their emphasis on classical syndromes and the lack of nuance in their summary scores (Crary and Gonzalez Rothi, 1989; Hula et al., 2010; Patterson, 2015). More recent tests have focused less on classification and more on assessing strengths and weaknesses across linguistic domains (for review, see Spreen and Risser, 2003; Patterson, 2015), leading to results in the form of a multidimensional language profile rather than a single diagnosis. However, the question of what should constitute a "linguistic domain" remains its own area of research.

A number of studies have aimed to reveal subcomponents of language in a data-driven manner. Dimensionality reduction techniques such as factor analysis (FA) and principal components analysis (PCA) have been used to observe how performance on multiple language subtasks in individuals with aphasia clusters together; the majority of studies have detected at least one factor that maps onto general severity, along with factors loosely mapping onto motor-speech, grammaticality, comprehension, and cognition; see Section 1 of Wilson and Hula (2019) for a review. Such factors have, in many cases, become the basis for clinical assessment, defining which dimensions are reflected in a multidimensional language profile. However, as Wilson and Hula (2019) note, it is important to remember that any factors identified in such a manner are, by necessity, a function of the tasks administered when probing for clusters, and thus task selection in any such investigation must be conducted carefully, aiming to sample widely across as many modalities as possible (Hanson et al., 1982; Wilson and Hula, 2019). Another concern relates to the way that scores or diagnoses on an aphasia battery are calculated, which can influence the methods by which their results may be validly statistically analyzed and interpreted; for example, the Porch Index of Communicative Ability (PICA) has been criticized for treating its scoring system as equal interval (i.e., a score of 1 is as different from a score of 2 as a score of 15 is from a score of 16) without evidence to suggest that it is necessarily even ordinal (Lincoln et al., 1981), with the WAB critiqued

on similar fronts (Hula et al., 2010); see Ivanova and Hallowell (2013) for a summary of related psychometric concerns in aphasia assessment at large. Particularly if ones' interest is in lesion-symptom mapping of the recovery process—that is, how the state of the brain relates to language performance over time—it is imperative that any changes between language assessments be a result of actual changes in language itself, rather than measurement error of the language assessment technique. It is therefore extremely important when choosing a quantitative method for assessing language in aphasia to consider the reliability, validity, and statistical assumptions of the language measure in question.

A newly introduced language assessment, the Quick Aphasia Battery (QAB), offers a rapid means by which to comprehensively assess language using carefully selected test items (Wilson et al., 2018b). The QAB is a non-classificatory, multidimensional aphasia battery consisting of eight subtests: (1) level of consciousness, (2) connected speech, (3) word comprehension, (4) sentence comprehension, (5) picture naming, (6) repetition, (7) reading aloud, and (8) motor speech. Scores on these subtests are used to derive eight summary measures: (1) word comprehension, (2) sentence comprehension, (3) word finding, (4) grammatical construction, (5) speech-motor programming, (6) repetition, (7) reading, and (8) QAB overall, along with measures of dysarthria and consciousness. The paper introducing the QAB provides quantitative evidence of its reliability and validity (e.g. ICC range of .91-.99 for inter-rater reliability), provides clear scoring guidelines for assessors, and supplies different testing forms for use specifically in longitudinal assessment. The QAB demonstrated its ability to capture difference in language in its norming sample—with more variability in language profiles across clinical diagnoses than within them—and correlated with subscores on the well-established WAB. Limitations of the QAB include the absence of writing in its sampled language domains, and a relatively small norming sample size. However, its psychometric characteristics, in conjunction with its extremely rapid administration time that renders it easy to administer at the bedside or even remotely, make the QAB a promising tool for the longitudinal assessment of aphasia and language recovery.

Before moving forward, it is important to note that all of the previously described measures

are impairment-based measures, meaning they assess the extent of disability across hypothesized linguistic domains. However, there is also another school of thought entirely around aphasia assessment in which language is measured from a functional perspective, that is, based on an individual's ability to communicate and participate in their life, rather than their ability to achieve high performance on strictly linguistic tasks (Crockford and Lesser, 1994; Doyle et al., 2003; Galletta and Barrett, 2014; Fama et al., 2016; Fridriksson and Hillis, 2021). For the purposes of this dissertation, the focus is on linguistic outcomes, and thus such functional measures of language will not be discussed in detail here; however, functional communication strategies and psychosocial wellbeing are both crucial aspects of effective recovery, and should not be overlooked in broader contexts. A review of functional methods for assessment is available in Chapter 8 of Spreen and Risser (2003), and some preliminary work from this dissertations' author on the benefits of aphasia groups for maximizing life participation, psychosocial wellbeing, and creative self-expression while living with aphasia is presented in Kasdan et al. (2021).

1.2.2.2 Methods for quantifying lesions

Let us pivot now back to the quantification of lesions as needed for statistical approaches to lesion-symptom mapping. One of the first efforts to explicitly quantify lesion extent in a study by Turkheimer et al. (1990), which quantified the location and extent of lesions with respect to anatomical landmarks across all axial slices of the brain, then assessed their covariance with various behavioral measures to derive "importance functions" across the cortex as related to those behaviors. This study found damage to left frontal regions to be most important for predicting verbal errors on the Aphasia Screening Exam. Similarly, Caplan et al. (1996) normalized CTs to a Talairach template, calculated the amount of each normalized slice occupied by the lesion, multiplied it by the slice thickness, and summed across slices in which the lesion appeared, to then be split across regions of interest and associated with behavioral performance on a sentence comprehension task. This study found that left-hemisphere peri-Sylvian areas were most associated with poor sentence comprehension, but detected no difference between patients with anterior

and posterior lesions. These were some of the earliest studies to use explicit quantification of lesions in terms of covariance with language performance along a spectrum of scores, rather than artificially dividing those scores or lesions into categories. A breakthrough in this method was reached, however, when advances in computational ability first allowed for statistical calculations on a voxel-wise basis.

In the late 1990's and early 2000s, two new, mass-univariate methods for lesion-symptom mapping were introduced: voxel-based lesion symptom mapping (Bates et al., 2003) and voxel-based morphometry (VBM) (Ashburner and Friston, 2000). In VLSM, lesions are drawn either manually or automatically on 3D images and then treated as binary masks, such that for each voxel participants may be divided into lesioned versus non-lesioned groups and statistically assessed for differences in the behavioral measure of interest. In VBM, each voxel's integrity is instead treated as a continuous measurement of the proportion of the voxel that is gray matter; this measure is then correlated directly on a voxel-wise basis to behavioral scores across participants. The two methodologies have been shown to produce partially overlapping but non-identical results in the localization of language functions (Geva et al., 2012), and have occasionally been blended to achieve particular theoretical or methodological aims (Leff et al., 2009; Wilson et al., 2015). Taken together, VBM and VLSM have suggested that many distinct sub-functions of language have distinct neural correlates, demonstrating, for example, differing neural bases for phonemic substitution errors versus phonemic distortion errors (Wilson et al., 2010) and semantic naming errors versus semantic conceptualization errors (Schwartz et al., 2009); see Wilson and Hula (2019) for a detailed review. However, such univariate approaches are plagued by three main issues; first, they artificially treat each voxel as a potentially independent predictor of language, even though this is conceptually untenable; second, they are plagued by the "partial injury problem", in which the ability to detect regions relevant to a function of interest is dependent upon individuals exhibiting different symptoms across lesion statuses at every voxel within a given functional region (Rorden et al., 2009; Karnath et al., 2018); and third, they are subject to issues of notable spatial mislocalization due to dependencies between lesion locations following the vascular distribution of

the brain (Mah et al., 2014; Nachev, 2015).

Mass-univariate analyses artificially treat measurement units—that is, arbitrarily delineated cubes of neural tissue in a brain image, or voxels—as functional units—that is, segments of the brain that can meaningfully underlie particular cognitive processes (Karnath et al., 2018; Pustina et al., 2018). It would seem quite unreasonable to suggest that, if a single cubic millimeter of brain was lesioned in myself but not one of my peers, our language abilities would differ in any meaningful way; however, mass-univariate analyses assume just that, treating each cubic millimeter of tissue as if it has the potential to independently predict differences in language or language recovery across individuals. Rather, lesioned voxels occur in a neural, vascular, and functional context, such that the lesion status of a given voxel is (a) highly correlated with that of its neighbor (Pustina et al., 2017), (b) highly dependent upon the vascular supply of the brain (such that certain voxels are significantly more likely to be damaged than others; see Mah et al., 2014; Pustina et al., 2018), and (c) is just one part of a much larger pattern of lesioned voxels, which in its entirety has effects on behavior. Related to this last point is the "partial injury problem", or the fact that behavioral effects may result from damage to any given part of a functional unit, without the functional unit needing to be destroyed in its entirety. However, if any set of voxels within a functional unit are all needed for a given behavior, each of those voxels may be opaque to VLSM analysis, as its statistical testing requires that behavior differs across lesion statuses within a single voxel (Rorden et al., 2009; Karnath et al., 2018). Finally, the mass-univariate approach can easily mis-localize symptoms as a result of neural regions that can be systematically damaged with, without themselves being, areas critical for a behavior of interest (see Figure 1.2 for an illustration, as well as Mah et al., 2014; Inoue et al., 2014; Herbet et al., 2015; Xu et al., 2018).

Take, for example, the insula: the insula is often damaged in large MCA stroke due to its position along the artery's M2 segment; thus, most patients who have any significant damage along the MCA will have damage to the insula, regardless of the heterogeneity in lesion distribution along more distal branches of the artery (Kodumuri et al., 2016). Statistical power for detecting lesion-symptom relationships may therefore be greatest in such vulnerable vascular territory (e.g. the

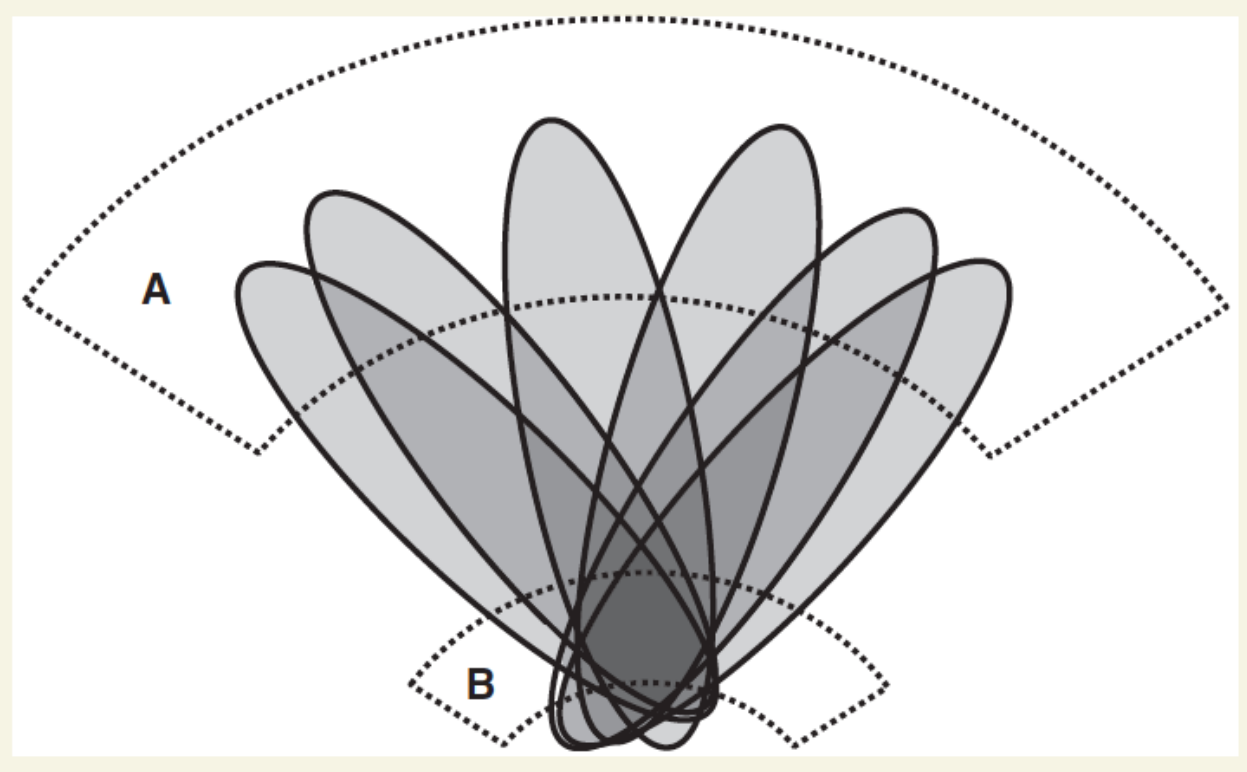

Figure 1.2: Reproduced from Mah et al. (2014): "Illustration of how stereotyped patterns of brain damage...across a set of patients can hypothetically mislocalize damage of any part of critical area A...to the non-critical area B...This will happen whenever the spatial variability of damage to a non-critical area is less for the group or factor of interest than for the critical area" (p. 2523).

insula), artificially de-emphasizing distal regions or tracts that are, in fact, behaviorally relevant for language, but may not be adequately represented in a patient sample (Mah et al., 2014; Xu et al., 2018; Wilson and Hula, 2019). Additionally, patterns of neural damage often occur in a stereotyped manner due to the innate structure of the vasculature, resulting in a "hidden deep structure in the data" (Nachev, 2015, p. 1) that spuriously influences the results of VLSM. For all of these reasons, the mass-univariate approach to lesion-symptom mapping is difficult to interpret, and some argue should not be used at all (Nachev, 2015).

1.3 Multivariate approaches

Recently, as computational power and ease of use has increased, multivariate[1] methods—that is, methods that are able to deal with extremely high numbers of variables—have become increasingly popular, and brought with them the ability to address a number of the issues presented by mass-univariate methods. While mass-univariate lesion-symptom mapping methods operate on voxels, aiming to localize functions to particular places in the brain based on differences in performance across lesioned and non-lesioned individuals at each voxel, multivariate lesion-symptom mapping methods operate on individuals, aiming to predict functional abilities based on an entire lesion map considered simultaneously (often in combination with other factors). This capacity for prediction over and above that of localization is an important difference between univariate and multivariate methods; while both univariate and multivariate methods can be used to represent topographically the neural correlates of behaviors (as we will discuss in detail shortly), only multivariate methods also possess the capacity to quantitatively predict behaviors from voxel-based lesion data on an individual basis. While this is a slightly different question than that posed by much of prior neuroscientific research historically, it is perhaps an even more clinically relevant area of research; both clinicians and patients alike would likely be much more eager to know what they can expect

[1]The term "multivariate" here is used for consistency with other work in this field (e.g. Zhang et al., 2014; DeMarco and Turkeltaub, 2018; Sperber et al., 2019; Ivanova et al., 2021); however, this may in fact be a bit of a misnomer, as models that take multiple predictors should in the strictest terms be called multi*variable* models, as opposed to multi*variate* models, a term which technically refers to models that produce multiple outcomes (Hidalgo and Goodman, 2013). For the purposes of this dissertation, this questionable terminology goes unchallenged, but future work may adjust terms accordingly.

during recovery than whether or not a given cubic millimeter of tissue is associated with the ability to speak.

There are a number of methods available for multivariate lesion-symptom mapping, influenced by the field of machine learning, many of which have recently been used in the context of research on aphasia. Broadly, these methods can be divided into supervised versus unsupervised approaches, the former making use of some known "ground truth" (that is, prior knowledge about the outcome of interest), and the latter operating in its absence, revealing instead any underlying structure in the data set that may or may not be known to the experimenters. Supervised approaches are very common in the multivariate lesion-symptom mapping literature and are the focus of this dissertation, so they will be treated with the most care here; however, it is worth noting that supervised approaches often make use of unsupervised approaches within their implementations, e.g. for the purposes of dimensionality reduction (that is, reducing the number of features/predictors to be input to a model). A discussion of some existing methods for multivariate lesion-symptom mapping, focusing on those most commonly used in research on aphasia, will be presented below.

1.3.1 General concepts

In supervised learning approaches, the overarching goal is to uncover relationships between input features and known outcomes of interest such that future, unknown outcomes can be predicted based on their input features alone. Relationships between features and outcomes are first "learned" by an inducer (or specific machine learning algorithm) during the training phase, during which some subset of the available data (the training data) is input in conjunction with its outcome labels so that a potential relationship can be modelled between the two. The resulting model is then subsequently tested on its ability to predict outcomes on the remaining held-out (testing) data using only input features. This training-testing split occurs in an iterative fashion such that all of the data is eventually used for both training and testing, in a procedure referred to as cross-validation. Ideally, an independent test set is also held out to evaluate final model performance. These hold out sets are crucial, as they simulate how well the model might do on future data with

truly unknown outcomes (what we are *actually* interested in as scientists, who are curious about the way things work in the real world); we do not want our model to overfit (that is, to only work on the specific data set we trained it on and not generalize to new exemplars). The details of how the cross-validation procedure should occur—e.g., whether a leave-one-out, in which all but one exemplar are trained at a time, or a k-fold approach, in which smaller groups are used in the training set—remains a subject of some debate (Poldrack et al., 2020), but regardless of the specifics of its implementation it is an important process for assessing the generalizability and utility of a model, with some measure of the average accuracy across training-testing splits generally mapping onto the reported model performance in a study.

In the case of lesion-symptom mapping for aphasia, generally the input features would (at minimum) reflect some representation of each patient's lesion, such that each of the P participants in question is represented as a vector in n-dimensional space where n corresponds to the number of lesion-based features of interest. For example, an experimenter might represent each patient as the binary lesion status of each voxel in their lesion image ($1 \times numberOfVoxels$, categorical) or as the percent damage they sustained to a number of a priori selected ROIs ($1 \times numberOfROIs$, scalar). The outcome would then be either some category (e.g., aphasia type) or some scalar (e.g., score on some aphasia assessment) reflecting the nature of the language phenomenon the experimenter aims to predict. Then the relationship between the lesion-based feature vector and the language-based outcome would be modelled using the inducer the experimenter deems theoretically useful for the problem at hand.

To date, the most commonly used inducers in lesion-symptom mapping for aphasia are support vector machines (SVMs; Vapnik, 1998). SVMs can be used either to classify, in the case of support vector classification (SVC), or estimate continuous values for, in the case of support vector regression (SVR), unlabeled test data. In a linear SVM analysis, a hyperplane is calculated to model a linear relationship between the input features and the behavioral outcome of interest; if (as is generally the case with multivariate data) the relationship between the input and output variables is non-linear, a "kernel trick" may be used to project the data into a higher-dimensional

space where it can be modelled linearly (DeMarco and Turkeltaub, 2018). In the case of SVC, the hyperplane is meant to maximally divide data points (in this case, individuals) that belong to different binary classes (e.g. aphasic versus non-aphasic). The position and orientation of the separating hyperplane is influenced by those hard-to-classify data points that are closest to it, the eponymous "support vectors", and is chosen so as to maximize the distance or "margin" between the boundary and those support vectors. In the case of SVR, the hyperplane is less of a separator than a plane of best fit, a series of predictions (e.g. Aphasia Quotient as measured on the WAB) with some tolerance for error, referred to as the ε-tube, built around it. In SVR the hyperplane and its surrounding ε-tube are placed so as to capture as many data points as possible while maximizing distance between the hyperplane and the support vectors, which in this case are those data points that fall just outside the bounds of the ε-tube (DeMarco and Turkeltaub, 2018). Those features that informed the SVR can also be mapped back into voxel space to create topographic lesion maps akin to those created using VLSM, though this mapping is not straightforward (Zhang et al., 2014; DeMarco and Turkeltaub, 2018). Parameters that can influence the performance of SVM-based analyses include the box constraint C, which specifies how much to penalize misclassifications or errors; the kernel shape gamma, which dictates the manner in which the data is mapped into higher-dimensional space to simulate linearity; and the threshold for what constitutes an error during training of a regression model, ε. Notably, because of the emphasis on support vectors, the majority of the data is not actually used when determining the location for the separating hyperplane, making it computationally less expensive than many other methods and less subject to overfitting to atypical data points (Vapnik, 1998).

1.3.2 Prior multivariate LSM studies of aphasia

SVM approaches have become relatively popular in lesion-symptom mapping for aphasia. One of the earliest studies to use such an approach was Wilson et al. (2009), which showed that diagnosed primary progressive aphasia (PPA) subtype (semantic, logopenic, or non-fluent variant) could be reliably predicted using structural imaging data in an SVC analysis, attaining a mean accuracy of

92.2%. This study used voxel-wise gray matter integrity in the structural images as input, and accounted for the lateralized nature of degeneration in PPA by including a lateralization image (the difference between gray matter in the right hemisphere versus the left) as a feature. However, the study was limited by small sample sizes in some of its patient groups. Additionally, the nature of PPA is quite different than that of post-stroke aphasia in both its spatial distribution and degenerative progression, making such findings difficult to extend to a post-stroke population; thus, other studies that have used SVM-based approaches in investigations of neural determinants of PPA will not be discussed here.

The same technique has recently grown popular in investigations of aphasia following stroke. Yourganov et al. (2015) used SVC to classify aphasia subtype as diagnosed using the WAB in a post-stroke population, using the proportion of damage to a series of a priori selected ROIs as the input features. A range of classification accuracies were obtained, varying depending on which atlas was used to parcellate the brain images into ROIs and which aphasia subtypes were being distinguished; classification was best (87-95% accuracy) when fluent versus non-fluent subtypes were pitted against each other, but the model was generally unreliable at distinguishing within-fluency subtypes. This study also had relatively poor representation of some clinical groups (e.g., only 7 patients with Wernicke's aphasia), and could be argued to have relied too heavily on a priori assumptions with regard to the reality of the Boston classification/classical syndromes and relevant ROIs.

The first study to move beyond classification and instead use a regression-based SVM approach was Zhang et al. (2014), which attempted both to create SVR-based lesion-symptom maps (SVR-LSM) and to predict proportions of phonological and semantic errors on the Philadelphia Naming Test (PNT) using voxel-wise lesion status as input. The lesion maps were able to localize the neural bases of functions in synthesized lesion-symptom relationships (AUC of .94 in an ROC analysis, compared to .71 for VLSM) and corresponded well with VLSM-generated maps based on real patient data (with correlations of .94 and .87 between univariate and multivariate semantic and phonological-based maps, respectively). However, the accuracy at predicting behavioral outcomes

was very low (R^2 of .10 for semantic errors, .11 for phonological errors).

Xing et al. (2016) combined VBM and SVR-LSM to investigate whether right hemisphere gray matter volumes correlated positively with language outcomes in left-hemisphere stroke, but did not aim to explicitly predict such outcomes from neural data; rather, SVR-LSM was used mainly as the method for selecting regions of interest known to be associated with language.

DeMarco and Turkeltaub (2018) created an SVR-LSM toolbox to refine the Zhang et al. (2014) approach (in particular by providing more options for accounting for lesion volume), but again focused mainly on localizing language-predictive regions rather than explicitly predicting language outcomes.

SCCAN, or Sparse Canonical Correlation Analysis (Pustina et al., 2018), is a newly developed multivariate method that maximizes overall correlation between input features and outcome variables by adjusting weights in a series of common components; the authors describe these components as "principal components...of the covariance matrix computed between two different modalities acquired in the same subject" (e.g., voxel-wise lesion statuses and behavioral outcomes; Pustina et al., 2018, p. 155). This method was compared with VLSM on its accuracy at localizing simulated lesion-symptom relationships and was shown to outperform it on a variety of measures (e.g., Dice similarity with simulated source maps, average distance from implicated to actual region as assessed based on contour and peak voxel displacement, etc.). However, the SCCAN method, as noted by the authors, is extremely new and therefore at risk for bugs and errors, and additionally was indicated only for lesion-symptom mapping, not prediction.

Ivanova et al. (2021) directly compared a variety of univariate and multivariate methods for lesion-symptom mapping and found that multivariate methods still suffered from statistical issues such as high false positive rates and spatial displacement of implicated regions, the conclusion being that the jury is still out on whether multivariate methods should universally be preferred to univariate. However, this paper did not focus on the ability of multivariate methods to directly predict behavioral outcomes, a clear benefit of multivariate over univariate approaches, focusing instead on localizationist aims.

Lately, greater emphasis has been placed on prediction as well as localization. Del Gaizo et al. (2017) used linear SVR to predict WAB Aphasia Quotient (AQ) and fluency scores from percent damage to cortical language ROIs, structural connections from probabilistic DTI, and structural brain dynamics (that is, a measure of how information might spread through a structural network). The mean R^2 (over multiple iterations of k-fold cross-validation) between actual and predicted scores for the models using all predictors was 0.58 for AQ and 0.53 for fluency. The primary aim of this study was to compare the utility of connectivity-based versus cortically-based measures for predicting outcomes, as in a similar SVR-based study out of the same lab (Yourganov et al., 2016) which predicted WAB speech fluency, auditory comprehension, speech repetition, oral naming, and AQ from gray matter and structural connectome maps. In that study, R^2 values between actual versus predicted outcome were lower, ranging from 0.21 for auditory comprehension using connectome-based predictors to 0.50 for fluency using gray-matter predictors.

Hope et al. (2018) used SVR along with 16 other inducers to assess the relative utility of lesion load versus structural connectivity-based features for predicting language outcomes on multiple domains of the Comprehensive Aphasia Test. The main finding of this paper was that structural connectivity measures do not increase the predictive power of MLSM models, regardless of the specifics of how these models are built. This study boasts a very large sample size (818 participants pulled from the PLORAS database) and demonstrates high correlation between actual and predicted scores in multiple models (average R^2=0.42, maximum R^2=0.58 among the highest performing models). Some limitations of this study include the fact that all findings were dependent upon an automated lesion segmentation procedure (not meeting the current gold standard for lesion delineation; Liew et al., 2018) and relied on coarse-grained characterization of lesions (calculated as percentages of the AAL atlas, which captures less than 60 regions in each hemisphere) and structural connections (estimated from T1 images rather than participant-specific measures of tractography). Additionally, as this study's aim was to abstract across inducers to learn about the utility of different features, its model-comparison scope is extremely broad, and its final analysis takes into account only the highest-performing models (regardless of their specifics) on each lan-

guage domain. Thus, it is difficult to interpret the meaningfulness of any one model's reliability for predicting language outcomes from this study.

Halai et al. (2020) examined the utility of multiple imaging modalities (T1 versus DTI), brain parcellation strategies (whole brain-, anatomical-, lesion clustering-, connectivity-based-, and theorized language network-based atlases), and inducer types (kernel ridge regression, relevance vector regression, GPR, and multi-kernel regression, an expansion of SVR) for predicting four dimension of language and cognition post-stroke. Similar to Hope et al. (2018), they found that DTI data did not provide a significant benefit over other imaging modalities, regardless of the model parameters or language dimensions in question. Also similarly to Hope et al. (2018), lesions were segmented automatically, and interpretability of any one particular model's utility for predicting language outcomes was limited due to the very high number of models generated and reporting of results from only the best-performing models. Even still, while the best of the best-performing models had an R^2 of 0.53 between actual and predicted values (though mean squared error was the primary measure reported), none of the runners-up exceeded R^2 values of .28.

Kristinsson et al. (2021) used SVR to predict multiple subscores of the WAB-R using task-based fMRI, DTI, cerebral blood flow (CBF), and lesion-load data in a sample of size N=116 (which skewed disproportionately male, with only 41 females in the data set). The fMRI measures utilized were based on a picture naming task in which the task condition required speaking and the control condition did not, which may confound imaging results with aphasia severity due to increased effort in the speaking condition. The maximum correlation between actual and predicted values achieved in this study was R^2=0.45, based on a model using all modalities as predictors. Though this finding suggests the multimodal model performed better than the other models generated in this particular paper, it is lower performance than many of the published models discussed above. This is the case even despite the fact that feature selection (that is, the decision of which variables should be used as input to the SVR) consisted of univariate regressions of all ROIs in all imaging modalities on the same data and language scores to be predicted by the full SVR models, an example of "leakage" of test data into training (Poldrack et al., 2020) which can inflate predic-

tion accuracy. Similarly, the selection of the radial kernel function applied in the final SVR models was accomplished via "applying various kernel functions [to the data] to select the most robust parameters" (p. 1689), and it is not described whether or not this selection process was limited to be based on training data alone. A recent paper by Hosseini et al. (2020), entitled "I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data" warns against the dangers of such an approach (which apply to regression as well as classification analyses), as it is likely to result in models that do not generalize to new exemplars.

Another supervised method that has been employed for the use of lesion-symptom mapping and outcome prediction is Gaussian process model regression (GPR). GPR is similar in concept to SVR in that it predicts real-valued outcomes based on input featural data mapped into a high-dimensional space. However, it differs from SVR in a number of ways mathematically, perhaps the most important being that it does not provide a single estimated outcome but a distribution of estimated outcomes for each set of input features, mapping approximately onto its "confidence" about its estimates. The distribution of predictions at each set of feature values is created by determining all functions that could reasonably explain the data (subject to some limiting and smoothing constraints), then providing the output to all of these functions based on the featural inputs at that point; unlike SVR, it does not optimize the placement of a single hyperplane, but generates a set of possible hyperplanes along which outcome values can be predicted and provides the output from all of them at once.

Hope et al. (2013) used GPR based on structural imaging data and clinical variables to predict speech production on the CAT at both single and multiple time points. This study had a large sample size (with 270 total patients, of whom 38 were assessed more than once), reasonable predictive ability (with R^2 ranging from .34-.59 across speech production subtasks and versions of the model), and made an explicit attempt to account for recovery, a facet of aphasia that is crucially important and often ignored in the lesion-symptom mapping literature (Flowers et al., 2016; Price et al., 2017). Hope et al. (2013) made use of language scores and brain scans available in the PLORAS database and created a composite speech production score associated with each scan,

calculated as the mean of the minimum score of two visually-based production tasks (object naming and picture description) and two aurally-based production tasks (word and sentence repetition). Lesions were delineated on structural images using an automated detection algorithm and binary thresholding procedure. A leave-one-out cross-validation procedure was used in the GPR models, of which there were several: the first, a model based only on time post-stroke, age at time of stroke, handedness prior to stroke, and gender (resulting in an R^2=0.01); the second, a model in which lesion volume was included as a predictor (resulting in an R^2=0.35); the third, a model in which the extent of lateralization of the lesion (indexed as the number of lesioned voxels in each hemisphere) was added (resulting in an R^2=0.47); the fourth, a model in which the lateralization predictor was replaced with the proportion of damage to 232 anatomically defined ROIs (resulting in an R^2=0.52); and finally, the fifth, a more constrained model consisting of time post-stroke, lesion size, and proportional damage to the 35 most relevant ROIs as indicated using an automatic relavence detection procedure (resulting in a final R^2=0.59). These results were superior to other similar studies at the time. However, some important caveats must be noted. First, although the final selected model consisted of 35 ROIs, similar results could be obtained using approximately half the number of predictors, suggesting redundancy/autocorrelation within the input variables. Similarly, the selection of these 35 ROIs appears to have been completed post-hoc (that is, using information gained on the full dataset during previous iterations of the model-building procedure); while this raises possible concerns of "leakage" as described above, the fact that results of earlier versions of the model are also reported (in which the unselected list of ROIs was used) at least offers some transparency with regard to the relative performance increases incurred. Second, the study combined images of stroke in both acute and chronic stages, which can vary notably due to either biological processes or imaging differences and lead to the potential for introducing biasing noise into the analysis (i.e., if production is generally better in individuals in later stages of recovery, artifactual characteristics of non-acute images could end up driving predictions; Loughnan et al., 2019). Third, only a small fraction of the patients (N=38) were assessed longitudinally, making any claims about the nature of "recovery" somewhat weak. Finally, the study focuses on

only a single domain of language—speech production—leaving any questions about expectations for recovery across other linguistic subdomains unanswered.

Corbetta et al. (2015) used an unsupervised PCA and FA approach to investigate what factors, across multiple imaging and behavioral domains, may predict language (and other behavioral) outcomes; they found that language outcomes across modalities are associated with integrity, function, and connectivity of multiple nodes in a widely distributed language network, extending beyond known peri-Sylvian regions. The functional connectivity finding in particular was corroborated in a study using a supervised ridge-regression approach (Siegel et al., 2016) which showed that both lesion location and changes in functional connectivity following stroke likely have effects on language, with altered interhemispheric communication exerting a large influence. Other work made use of a stacked, multi-modal random-forest approach to estimate picture naming, sentence repetition, sentence comprehension, and aphasia severity scores through a combination of lesion maps, structural connectivity, and functional connectivity, attaining strong correlations between predicted and obtained scores; however, performance on a completely untrained validation set was much lower than that reported for the preceding models, rendering the extent of the true generalizability of the model somewhat unclear (Pustina et al., 2017).

It is worth noting that, in all but one of these studies just described (Hope et al., 2013), only a single time point was considered, and in most cases, only a subsample of language functions were investigated; no SVM-based approaches have yet, to our knowledge, been used to examine complete language profiles at multiple, systematically arranged time points along the course of recovery.

To summarize, multivariate methods are an extremely promising means by which to account for the multidimensional nature of lesions and language, allowing us both to map out lesion-symptom relationships and predict language outcomes directly. However, there is yet to exist a study that has applied such methods to specifically investigate longitudinal recovery from aphasia across a comprehensive set of language domains. This is the aim of the present dissertation.

CHAPTER 2

Methods & Analysis

2.1 Characterization of data set

2.1.1 Participants

359 patients both with and without aphasia consented for this study at the bedside following stroke. Consent was acquired directly from the patient when possible, with the use of visual aids and supportive conversation techniques as necessary; in participants for whom comprehension deficits were too great for direct consent to occur, surrogate consent was acquired from a family member. In order to qualify for the study, patients must have met the following inclusion criteria:

- left hemisphere or aphasia-causing stroke confirmed by CT or MR imaging
- over 18 years old
- pre-morbid fluency in English
- no previous symptomatic stroke in a known language area or its right hemisphere counterpart
- no concurrent neurological condition (e.g. dementia or schizophrenia).

All consenting patients were deemed aphasic or not aphasic per clinical impression by a speech-language pathologist. Language evaluation using the QAB was attempted at the bedside for all patients. Follow-up language evaluation at 1 month, 3 months, and 12 months post-stroke was attempted only for those individuals who presented with aphasia acutely, barring rare exceptions. See Table 2.1 for patient characteristics and retention across time points. At the acute time point, independent samples t-tests revealed no significant difference in age between people with ($M = 62.78$, $SD = 14.04$) and without ($M = 62.31$, $SD = 15.62$) aphasia, $t(349) = 0.28$, $p = .78$, as well as no significant difference in years of education between people with ($M = 12.86$, $SD = 3.16$) and without ($M = 13.38$, $SD = 2.74$) aphasia, $t(345) = -1.50$, $p = 0.13$. Similarly, Fisher's exact

tests revealed no significant differences in the prevalence of ischemic versus hemorrhagic strokes in people with (174/223) and without (95/120) aphasia ($p = 0.89$), no significant differences in the prevalence of males versus females in people with (122/230) and without (57/121) aphasia ($p = 0.31$), and no significant differences in the prevalence of right versus non-right (i.e., left or ambidextrous) handedness in people with (205/230) and without (106/121) aphasia ($p = 0.72$).

Of 359 patients consented, 5 were considered to be outliers due to suspected right hemisphere dominance or bilaterality for language, as evidenced by aphasia given right hemisphere lesion ($N = 2$) or absence of aphasia given large left hemisphere lesion ($N = 3$). All analyses and figures herein exclude these patients. Please note that the choice to exclude these participants from analysis does not reflect a lack of regard for the importance of these unique presentations, but rather an attempt to capture generalizable patterns of the collected data set without running the risk of overfitting to atypical cases. For a detailed discussion of unexpected absence of aphasia in one of these unique participants, see Schneck et al. (2021).

2.1.2 Language evaluation

Language evaluation was attempted by a speech-language pathologist at each time point using the Quick Aphasia Battery (QAB). As described in the literature review, the QAB is a valid and reliable measure of language resulting in a multidimensional characterization of language function. Language summary scores on the QAB reflect single word comprehension, sentence comprehension, word finding, grammatical construction, speech-motor programming, repetition, reading, dysarthria, and overall severity. In some patients who were amenable to testing, assessment was prevented by either impairment or situational factors (e.g. intubation, somnolence); these patients were marked as "untestable" rather than "missing", as some knowledge about their language function was gained despite the inability to test it comprehensively. Any validity concerns that may have affected scores on a given summary measure (e.g. marked dysarthria impacting interpretability of responses in word finding) were flagged for handling in analysis. At the acute time point, the standard version of the QAB was administered; at all follow-up time points, the extended version

Clinical opinion:	**Aphasia**	**No Aphasia**
	Acute	
No. patients	230	121
Age (yrs)	62.78 ± 14.04	62.31 ± 15.62
Sex (M/F)	122/108	57/64
Handedness (R/ambi/L)	205/5/20	106/3/12
Education (yrs)	12.86 ± 3.16	13.38 ± 2.74
Stroke type (isch/hem/N.A.)	174/49/7	95/25/1
	One month	
No. patients	103	1
Age (yrs)	62.24 ± 13.83	56
Sex (M/F)	60/43	0/1
Handedness (R/ambi/L)	90/2/11	0/0/1
Education (yrs)	13.22 ± 2.70	12
Stroke type (isch/hem/N.A.)	82/21/0	0/1/0
	Three months	
No. patients	96	0
Age (yrs)	62.66 ± 13.40	N.A.
Sex (M/F)	57/39	N.A.
Handedness (R/ambi/L)	84/3/9	N.A.
Education (yrs)	13.20 ± 2.80	N.A.
Stroke type (isch/hem/N.A.)	77/19/0	N.A.
	One year	
No. patients	70	0
Age (yrs)	61.77 ± 13.37	N.A.
Sex (M/F)	38/32	N.A.
Handedness (R/ambi/L)	61/3/6	N.A.
Education (yrs)	13.57 ± 2.92	N.A.
Stroke type (isch/hem/N.A.)	55/15/0	N.A.

Table 2.1: Participant characteristics and retention across time points. No significant differences in patient characteristics were detected between individuals with and without aphasia at the acute time point. The eight cases in which stroke type is unavailable acutely come from patients who consented and for whom a clinical decision regarding aphasia was obtained, but for whom no imaging or language data was acquired.

(including additional sections on writing, written word comprehension, and extra single word and sentence comprehension) was administered. Follow-up evaluations were administered over Zoom as necessary during the COVID-19 pandemic.

Many scoring decisions on the QAB, particularly those related to ratings of connected speech, contain some subjectivity. Thus, consensus meetings were held to discuss the scores for each

patient with aphasia and to arrive at the final scores used in analysis.

2.1.3 Lesion delineation

As part of their clinical care, all patients that come through Vanderbilt University Medical Center suspected for stroke undergo a head MRI and/or CT to identify the presence, location, and extent of neural damage; consenting patients agreed to have these images collected for the purposes of the research study. Lesions were drawn manually in ITK-Snap (Yushkevich et al., 2006) on clinical imaging acquired within the first 5 days post-stroke. Lesion drawings were completed by trained students in the Language Neuroscience Lab, with guidelines determined based on consultation with the prinicipal investigator and a VUMC neuroradiologist. These guidelines are described in detail below.

Lesions due to ischemic stroke (in which blood supply to a portion of the brain is blocked by a blood clot or embolus) and hemorrhagic stroke (in which a weakened blood vessel ruptures, leading to a pooling of blood on the brain that is toxic to nearby tissue) appear differently on different modalities of MRI (Dehkharghani and Andre, 2017) and at different times post-stroke (Lin and Liebeskind, 2016). In acute ischemic stroke, the lack of blood flow due to occlusion results in swelling which restricts the motion of extracellular water (Xing et al., 2012). This restricted diffusion is visible as increased signal on diffusion-weighted magnetic resonance imaging (DWI) and decreased signal on both apparent diffusion coefficient (ADC) MRI (Baliyan et al., 2016) and CT (Lin and Liebeskind, 2016). Acute hemorrhagic stroke appears hypointense on FLAIR imaging and hyperintense on CT due to effects of pooling blood on the magnetic susceptibility and density of nearby tissue (Heit et al., 2017). Thus, when MRI was available, lesions due to ischemic stroke were drawn on DWI/ADC and hemorrhagic strokes were drawn on FLAIR, with the minority of patients who did not undergo MRI scanning having their lesions drawn on CT. VUMC-acquired imaging was preferred, with outside images used when VUMC imaging was unavailable. The order of preference for base images in lesion drawing was as follows: first, VUMC MRI; second, outside MRI; third, VUMC CT; fourth, outside CT. Participants in whom extension of the lesion

occurred within 30 days following study-eligible stroke were each represented with an additional lesion drawing depicting the extended lesion, termed an *ext* image. In participants who had had asymptomatic strokes prior to enrollment in the study, prior lesions were delineated in a different color to distinguish them from study-relevant lesions; for the purposes of this dissertation, these prior lesions were masked out and excluded from analysis.

The resulting binary lesion masks, along with their accompanying clinical images, were normalized to MNI space using both the unified segmentation procedure as implemented in SPM12 (Ashburner and Friston, 2005) and DARTEL, a top-performing deformation algorithm for normalizing structural images (Ashburner, 2007; Klein et al., 2009). All analyses presented here make use of the images warped using the unified segmentation procedure. All image warps were checked, manually adjusted as necessary, smoothed using a 4mm FWHM Gaussian blur (Cox, 1996), and approved by the author and principal investigator. Warped *ext* images were updated to reflect both the original and the extended lesion by taking the union of the two images in MNI space; this allowed for a complete representation of the tissue believed to be damaged at the time of extension, despite any changes in visibility of the initial lesion on follow-up imaging due to pseudonormalization over time (Allen et al., 2012).

2.2 Analysis

2.2.1 Mixed effects modeling

Mixed effects models are regression models that, by taking into account grouped structure within data, are able to make less biased estimates of means at different levels of a factor in the presence of randomly missing data (Cunnings and Finlayson, 2015). In order to examine basic patterns of language recovery in the data set independent of subject-level variation, mixed-effects models of QAB scores across domains were generated to assess the effect of time post-stroke on language function using the *fitlme* function in Matlab2019a (Mathworks, Inc., 2019). Time was modeled as a fixed effect (as time points at which to examine language were pre-selected as part of the study design) while participants were modeled as random effects, with random slopes (rates of

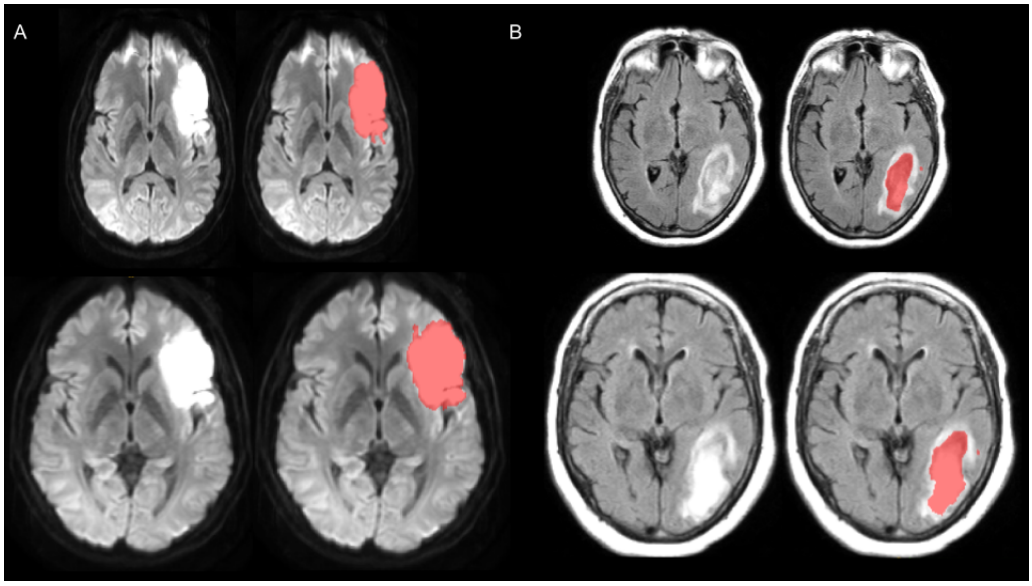

Figure 2.1: Lesion drawings and their corresponding MNI warps from two sample participants. Top row displays original modalities with and without lesion drawings, with bottom row showing the same images after warping to MNI space using the unified segmentation procedure. (A) An ischemic stroke drawn on DWI. (B) A hemorrhagic stroke drawn on FLAIR, excluding surrounding edema.

recovery), intercepts (baseline acute scores), and slope-intercept correlations (effects of baseline score on rate of recovery) modeled for each participant. Note that this analysis was completed only on individuals who presented with aphasia acutely, as data from individuals without aphasia were missing from the data set in a non-random manner due to lack of follow up. The end result was nine mixed-effects models, one for each of the QAB domains of interest (i.e. single word comprehension, sentence comprehension, word finding, grammatical construction, speech-motor programming, repetition, reading, dysarthria, and overall severity) reflecting the impact of increasing time post-onset on language function in that domain.

2.2.2 Support vector regression (SVR)

As discussed in Section 1.3, support vector regression (SVR) is a method in machine learning by which to predict real-valued numbers from high dimensional input data. Machine learning models require input in the form of vectors of real numbers corresponding to representative features of the actual data of interest. The methods by which these features were generated and subsetted to create

different models for comparison are described below.

2.2.2.1 Feature representation

Due to their frequent discussion in the existing literature (Watila and Balarabe, 2015; Gerstenecker and Lazar, 2019), age, sex, years of education, handedness, stroke type, and lesion size (calculated as the size of the binarized lesion mask in cm^3) were included as an initial set of features in the model. Each numerical predictor was normalized to a 0-1 scale using min-max scaling by theoretically informed minimums and maximums (i.e. age between 0 and 100, years of education between 0 and 25, lesion size between 0 and 640 cm^3) to speed the model-building process and increase interpretability of beta weights across predictors.

Lesion location was transformed into a vector space representation via calculation of the alignment of each patient's lesion mask with 294 spatial regions of interest (ROIs) as defined in a combined gray matter and white matter atlas (Mori et al., 2005; Fan et al., 2016). During the creation of this combined atlas, any voxels which were assigned a value in both atlases were set to the value in the gray matter atlas; then, the values of voxels in the white matter atlas were re-assigned such that each ROI (that is, both gray matter and white matter ROIs) had a distinct value in the combined atlas image (see Fig. 2.2). Individual patients' lesion vectors were then calculated as the number of voxels in the intersection between their particular lesion mask (weighted by the "certainty" of each voxel as indicated by the smoothing procedure) and each ROI in the combined atlas, divided by the size of that ROI. Thus, each patient's lesion was represented as a 1×294 vector corresponding to the extent of damage to each of the 294 ROIs in that patient. This vector will henceforth be referred to as "lesion load". Note, however, that for the purposes of analysis only left hemisphere lesion load was used, as right hemisphere damage was not well represented in this data set by design. Left hemisphere lesion load thus corresponded to a 1×144 vector, excluding all right hemisphere ROIs and 6 commissural tracts.

2.2.2.2 Model building

A series of support vector-based models were generated to predict QAB summary measures at various time points in recovery using the *fitrsvm* function in Matlab2019a (Mathworks, Inc., 2019). Models were designed in attempts to answer the following questions:

1. Given demographic information and lesion information, can we predict language scores at all time points following stroke?

2. Given demographic information, lesion information, and initial language scores, can we predict language *recovery* at all follow-up time points following stroke?

The first of these questions may be thought of as a cognitive neuroscience question, focusing on the relationship between brain and language, while the latter may be thought of as more clinically oriented, making use of all available information to predict long-term language outcomes.

The nature of each model built to answer these questions was dependent upon a number of analytical decisions, e.g. whether individuals who did not present with aphasia should be included in the data set and how to account for patients who were untestable at the time of assessment (who might reasonably be argued to be globally aphasic, and whose inclusion would allow for a higher sample size and increased power). Eight classes of models (Models 1 through 8) were therefore built corresponding to different methods of answering the questions above (see Table 2.2 for an

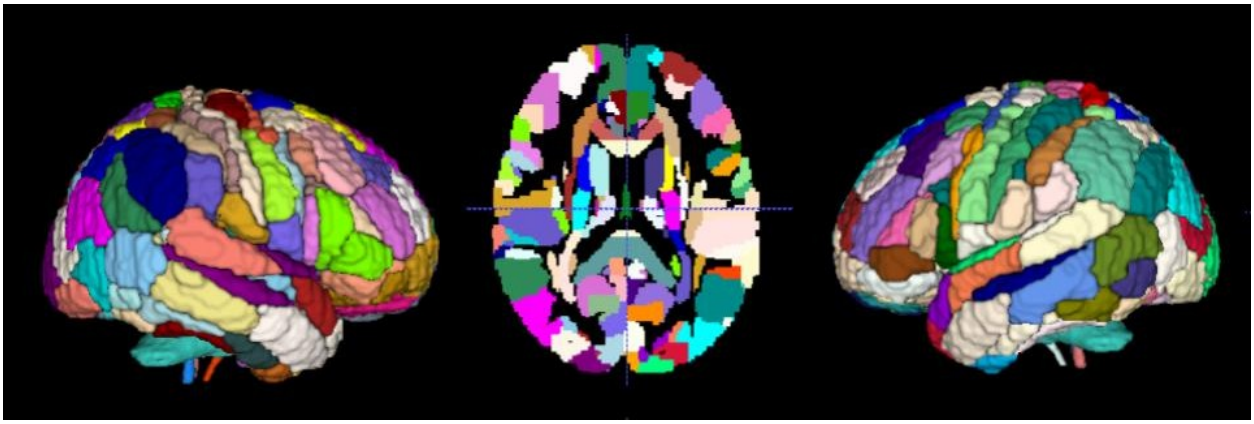

Figure 2.2: Combined gray and white matter atlas used in feature generation, based on Mori et al. (2005) and Fan et al. (2016).

at-a-glance depiction of each). All models were built up in stages corresponding to the relative ease of acquiring the relevant predictive information at the bedside, beginning with only those predictors that could be determined by conversing with the patient or looking at their medical chart (e.g. demographic information, stroke type, initial language scores), then sequentially adding in additional predictors requiring processing of the lesion image, namely lesion size (available with minimal processing) and lesion load (available with in-depth processing). The details of each set of models and their corresponding model-building procedures are described in further detail below.

		Cohort		**Predictors**		**Untestable**		**Pred. Time**			**Pred. Domain**	
		Aph	*No*	*Dem/Les*	*Lang*	*NaN*	*Zero*	*Acute*	*Follow-up*	*Diff*	*Subscores*	*Overall*
Q1	*M1*	X		X		X		X	X		X	X
	M2	X		X			X	X	X			X
	M3	X	X	X		X		X			X	X
	M4	X	X	X			X	X				X
Q2	*M5*	X		X	X	X			X		X	X
	M6	X		X	X		X		X			X
	M7	X		X	X	X				X	X	X
	M8	X		X	X		X			X		X

Table 2.2: Table depicting characteristics of different classes of models (*M1-M8*) addressing different experimental questions (*Q1*, on the neural bases of language, versus *Q2*, on the utility of all predictors including linguistic predictors for predicting outcomes and recovery). Columns 1-3 correspond to details of the input data (*Cohort* included, *Predictors* included, and handling of *Untestable* participants) while columns 4-5 correspond to details of the output data (*Predicted Times* and *Predicted Domains*).

Models 1 through 4

Models 1 through 4 address the cognitive neuroscience question regarding the neural bases of language by using only demographic and neuroimaging data to predict language outcomes at various times post-stroke. Models 1 and 2 use information only from individuals deemed to have aphasia, and are thus able to predict language outcomes acutely and at all follow-up time points. Models 3 and 4 use information from individuals both with and without aphasia, as this maximized sample size; however, they can only predict language outcomes acutely (as individuals without aphasia were not followed up). Models 1 and 3 treat untestable patients as missing data points, while Models 2 and 4 assume that, if tested, these patients would be globally aphasic (with a QAB overall score of 0); this means, however, that Models 2 and 4 can predict only the QAB overall

score, as the specifics of an untestable patient's language presentation cannot be quantified in the absence of testing. Predictors used in each stage of model building were as follows:

1. Age, sex, years of education, handedness, stroke type
2. Age, sex, years of education, handedness, stroke type, lesion size
3. Age, sex, years of education, handedness, stroke type, lesion size, lesion load

A total of 150 distinct models were therefore created to address Question 1 (Model 1: 4 time points $\times$ 9 subscores $\times$ 3 stages = 108 models; Model 2: 4 time points $\times$ 1 subscore $\times$ 3 stages = 12 models; Model 3: 1 time point $\times$ 9 subscores $\times$ 3 stages = 27 models; Model 4: 1 time point $\times$ 1 subscore $\times$ 3 stages = 3 models).

Models 5 through 8

Models 5 through 8 address the clinical question of how a patient's initial aphasia severity interacts with their stroke characteristics to determine the extent of their recovery at later time points. As acute scores were used as predictors here, no outcomes were predicted at the acute time point to avoid circularity in the models. Patients without aphasia were similarly not included in these models, as patients without aphasia had only the acute time point available by design. Models 5 and 6 predict language scores at later time points, differing from each other only in their treatment of untestable patients (see description of Models 1-4 above). Models 7 and 8, however, were not trained to predict scores at a given time point, but rather *changes* in scores *between* the acute and later time points, aligning more closely with prior work on proportional recovery in aphasia (Lazar et al., 2010; Marchi et al., 2017). Models 7 and 8 differ from each other in their treatment of untestable patients as above. Predictors used in each stage of model building were as follows:

1. Age, sex, years of education, handedness, stroke type, acute score
2. Age, sex, years of education, handedness, stroke type, acute score, lesion size
3. Age, sex, years of education, handedness, stroke type, acute score, lesion size, lesion load

A total of 180 distinct models were therefore created to address Question 2 (Model 5: 3 time points × 9 subscores × 3 stages = 81 models; Model 6: 3 time points × 1 subscore × 3 stages = 9 models; Model 7: 3 time points × 9 subscores × 3 stages = 81 models; Model 8: 3 time points × 1 subscore × 3 stages = 9 models).

All stages of all preceding models were generated using a linear SVR model, due to lower computational time, relatively high accuracy, and increased interpretability of output beta weights. However, a non-linear model was also generated for each of the final-stage models (that is, including all predictors), using an RBF kernel with hyperparameter values per the recommendations of Zhang et al. (2014) and DeMarco and Turkeltaub (2018) (RBF kernel scale of 5, box constraint C of 30, and ε value of 0.1). This allowed for the possibility of cross-featural interactions and eased comparison with much of the greater literature. Thus, 440 models were generated in total to address the experimental questions noted above.

2.2.2.3 Data handling

At each time point, patients were filtered to remove anyone in whom evaluation was missing at that time point. For patients in whom lesion extension had occurred, the latest lesion vector relative to the date of evaluation was used. In analyses where acute score was included as a predictor (Models 5-8), only patients who had the same lesion at the time of both assessments were included (i.e., no extension between assessments), as language change may have occurred in these cases due to change in lesion status rather than neuroplastic recovery.

2.2.2.4 Cross-validation and performance assessment

Model generalizability was assessed using a leave-one-out cross-validation procedure, in which each patient was held out in turn to have their score predicted from a model based on data from the remaining patients. Performance of each model was calculated using intraclass correlation coefficient (ICC) type A-1 as implemented by Salarian (2021) in Matlab2019a, corresponding to the degree of absolute agreement between actual and predicted values across all folds of the cross-validation procedure. Prior to calculation of this statistic, predictions were capped so as not

to be more extreme than theoretical bounds would allow (e.g., predictions of outcomes could be no less than 0 and no more than 10, or the possible scores to be obtained; predictions of change scores could be no more than 10, or the largest amount a participant could improve by). Note that ICC is a ratio measure of relative variance between a factor of interest (in our case, individual patients) and a nuisance factor (in our case, "raters", or measured versus predicted scores); thus, if there is very little variance in the factor of interest, we may reasonably expect low values of ICC (Liljequist et al., 2019). Standards for the quality of ICCs obtained came from Cicchetti (1994), with $ICC < 0.4$ considered "poor", $0.4 < ICC < 0.6$ considered "fair", $0.6 < ICC < 0.75$ considered "good", and $0.75 < ICC$ considered "excellent". In the cases of Model 7 and Model 8, ICC values were calculated both for actual versus predicted change scores and actual versus predicted outcome scores (in this case, the true outcome score versus the score generated by adding the predicted change score to the true initial score). This is both to put the various models on the same footing and to account for the fact that prediction accuracy on change scores is likely to be artificially inflated relative to prediction accuracy on the corresponding predicted outcomes due to mathematical coupling (see Hope et al., 2019; Hawe et al., 2019; Bonkhoff et al., 2020; Bowman et al., 2021, Results in Section 3, and Discussion in Section 4 for further detail).

2.2.2.5 Model comparison

The extent to which adding information to a model led to a significant improvement in predictive ability was assessed using permutation testing. For each of the available time points in each of the relevant language domains, a null distribution of ICC values was generated using up to 1000 iterations of the following procedure: randomly shuffle the newly added predictors across observations in the input feature matrix; generate a new SVR-based model predicting scores from that shuffled input; cross-validate and calculate ICC between true scores and predicted scores across folds; store the resulting null ICC value. Less than 1000 permutations were used only in cases where significant improvements were trivially apparent and the time required to generate null models was excessively long (Stage 2 versus Stage 1 of Models 1-4; 500 iterations). *P*-values (one-tailed) were

calculated for each model by calculating the number of null ICCs that were greater than or equal to the actual ICC divided by the total number of iterations of the procedure. Criteria for significance corresponded to an α level of 0.05 (that is, less than 5% of the ICC values in the null distribution were greater than or equal to the true ICC).

2.2.2.6 Beta weight extraction/feature importance

Though the focus of this dissertation is primarily on prediction, rather than localization of function, the beta weights output by linear SVR models may arguably be used to surmise the approximate relative importance of each predictor, and particularly each ROI in the lesion load vector, for generating predictions. In order to investigate the potential neural, clinical, and demographic bases of those long-term language outcomes that were well-predicted by our models, beta weights were extracted from the one year time point for re-fit versions of Models 1 and 5 (those that predicted all language domains without and with acute scores in the models, respectively), such that all available data was used (that is, with all available predictors and without withholding any test set). As there is currently no well-established method for assessing significance of SVR-generated beta weights in a neuroimaging context (Haufe et al., 2014; Sperber et al., 2019; Halai et al., 2020), raw beta weights (thresholded only to exclude those less extreme than 0.2) are plotted directly on their corresponding location on the brain. It is noted that, due to their experimental nature, these results should be interpreted with caution.

CHAPTER 3

Results

3.1 Descriptive statistics and figures

3.1.1 Language

Tables 3.1 and 3.2 show descriptive statistics at each time point across domains for people who presented with and without acute aphasia, respectively. Patients who were untestable are treated as missing in these tables.

Acute *(N=197*)*									
	Overall	*SWC*	*SC*	*WF*	*GC*	*SMP*	*Rep*	*Read*	*Dys*
Mean	5.81	7.47	4.57	4.52	6.17	8.27	6.23	5.02	8.08
SD	2.69	3.12	3.38	3.09	3.46	3.41	3.38	3.37	3.19
Range	0-9.75	0-10	0-10	0-10	0-10	0-10	0-10	0-10	0-10
One month *(N=97)*									
Mean	7.19	8.56	6.19	6.27	7.72	8.79	7.16	6.74	9.23
SD	2.33	2.62	3.35	2.98	2.44	2.48	2.73	3.21	1.82
Range	0-9.93	0-10	0-10	0-10	0-10	0-10	0-10	0-10	0-10
Three months *(N=96)*									
Mean	7.74	9.24	6.89	6.81	8.10	8.88	7.66	7.25	9.48
SD	2.21	1.80	3.20	2.84	2.43	2.48	2.49	3.13	1.20
Range	0.15-9.90	0-10	0-10	0-10	0-10	0-10	0-10	0-10	5.00-10
One year *(N=70)*									
Mean	8.13	9.49	7.32	7.56	8.39	8.82	8.05	7.72	9.61
SD	1.93	1.21	3.08	2.47	2.28	2.43	2.09	2.66	1.01
Range	0.90-10	2.50-10	0-10	0-10	0-10	0-10	0-10	0-10	5.00-10

Table 3.1: Mean QAB scores for people with aphasia across time. Asterisk is to indicate that $N = 196$ for reading at the acute time point, due to a single participant's inability to complete the reading task for situational reasons. Abbreviations are as follows: SWC = single word comprehension; SC = sentence comprehension; WF = word finding; GC = grammatical construction; SMP = speech-motor programming; Rep = repetition; Read = reading; Dys = dysarthria.

To visualize any potential relationship between aphasia severity and follow up retention status, an alluvial plot was generated in RawGraphs (Mauri et al., 2017) depicting the approximate

Acute *(N=121)*									
	Overall	*SWC*	*SC*	*WF*	*GC*	*SMP*	*Rep*	*Read*	*Dys*
Mean	9.19	9.82	8.40	9.00	9.26	9.96	9.25	9.07	8.37
SD	0.75	0.46	1.79	1.07	1.03	0.32	0.78	1.32	2.60
Range	4.05-10	7.08-10	0-10	2.25-10	1.00-10	7.50-10	5.42-10	3.33-10	0-10

Table 3.2: Mean QAB scores for people without aphasia acutely. Abbreviations are as follows: SWC = single word comprehension; SC = sentence comprehension; WF = word finding; GC = grammatical construction; SMP = speech-motor programming; Rep = repetition; Read = reading; Dys = dysarthria.

makeup of the full cohort (including both patients who were and were not clinically deemed to have aphasia) across evaluation times, with severity groupings generated as (reluctantly) indicated by the authors of the QAB (that is, QAB overall of 0-4.99 = severe, 5-7.49 = moderate, 7.5-8.89 = mild, 8.9-10 = very mild or no aphasia; see https://langneurosci.org/qab/). This plot demonstrates that a similar proportion of each severity is represented in the data set at all time points, suggesting little influence of severity on retention status. Note the distinction between unavailable and untestable patients; subsequent analyses treat these untestable patients as missing data points except where explicitly stated otherwise.

To visualize trajectories of recovery in individual participants, spaghetti plots of QAB overall scores in those who were clinically deemed to present with aphasia are displayed in Fig. 3.2. Grouping by severity is for purposes of visualization only, and it is emphasized that the cutoffs used are arbitrary (again, 0-4.99 = severe, 5-7.49 = moderate, 7.5-8.89 = mild, 8.9-10 = very mild or no aphasia); no claims are made about differences across these groups. Nearly all participants showed improvement across the first year of recovery with decelerating improvement across time. Similar patterns were observed across subscores of the QAB (not shown in Fig. 3.2), summarized in the mixed-effects models described below.

Mixed effects modeling generated estimates of QAB scores at each time point for individuals with aphasia on all summary measures of the QAB, revealing a decelerating trajectory of recovery across language domains (see Fig. 3.3). Results of both coding untestable patients as missing and

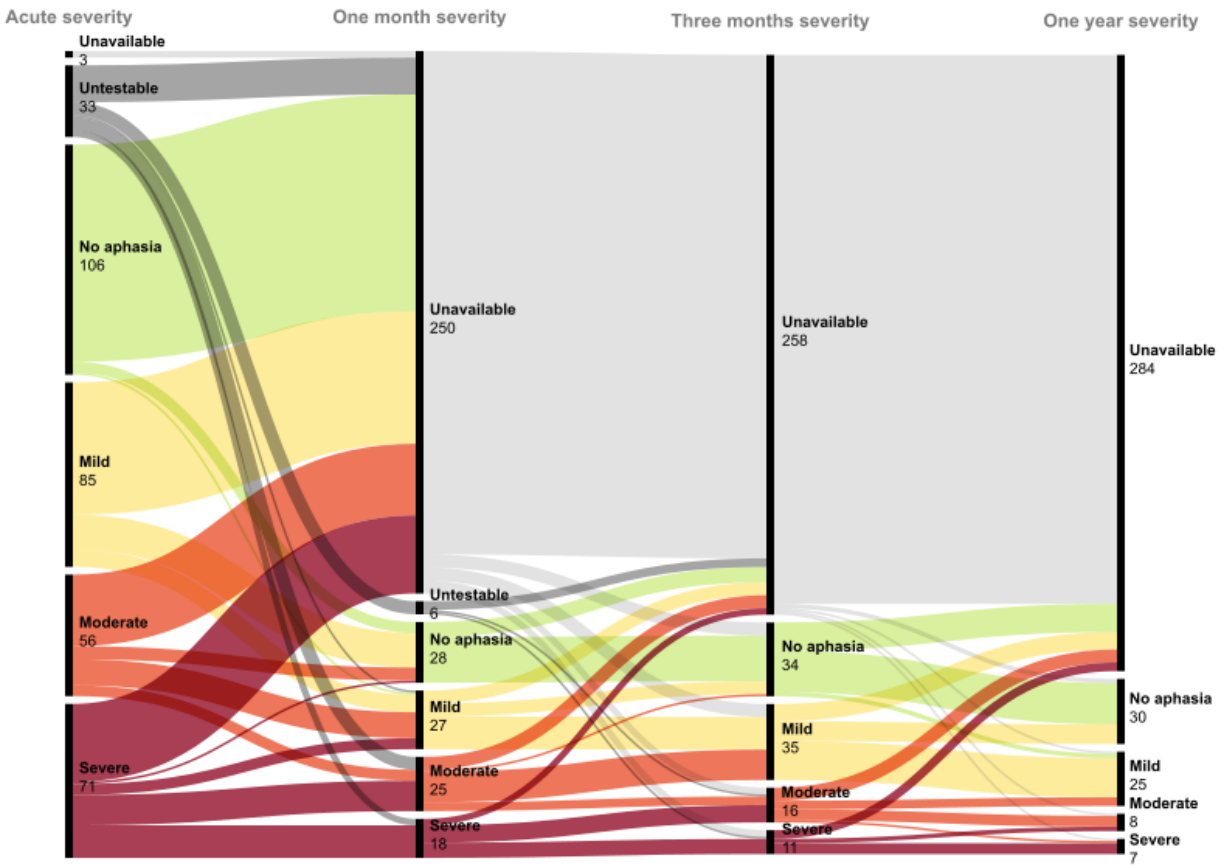

Figure 3.1: Alluvial plot showing sample makeup and retention across time points.

coding them as globally aphasic were examined. When patients who were untestable were coded as missing, the overall QAB score was estimated at 5.58 (SD 2.80) acutely, 7.33 (SD 1.54) at one month, 7.99 (SD 1.91) at three months, and 8.30 (SD 2.09) at one year. When these patients were coded as globally aphasic, estimates changed to 4.98 (SD 3.20) acutely, 6.98 (SD 1.87) at one month, 7.74 (SD 2.18) at three months, and 8.08 (SD 2.38) at twelve months. Note that only QAB overall score was examined when untestable patients were treated as globally aphasic (as detailed information about these patients' language was not available).

Two language domains did not show precisely the same decelerating trajectory as the others: speech-motor programming appeared to plateau after the one month time point, while word finding

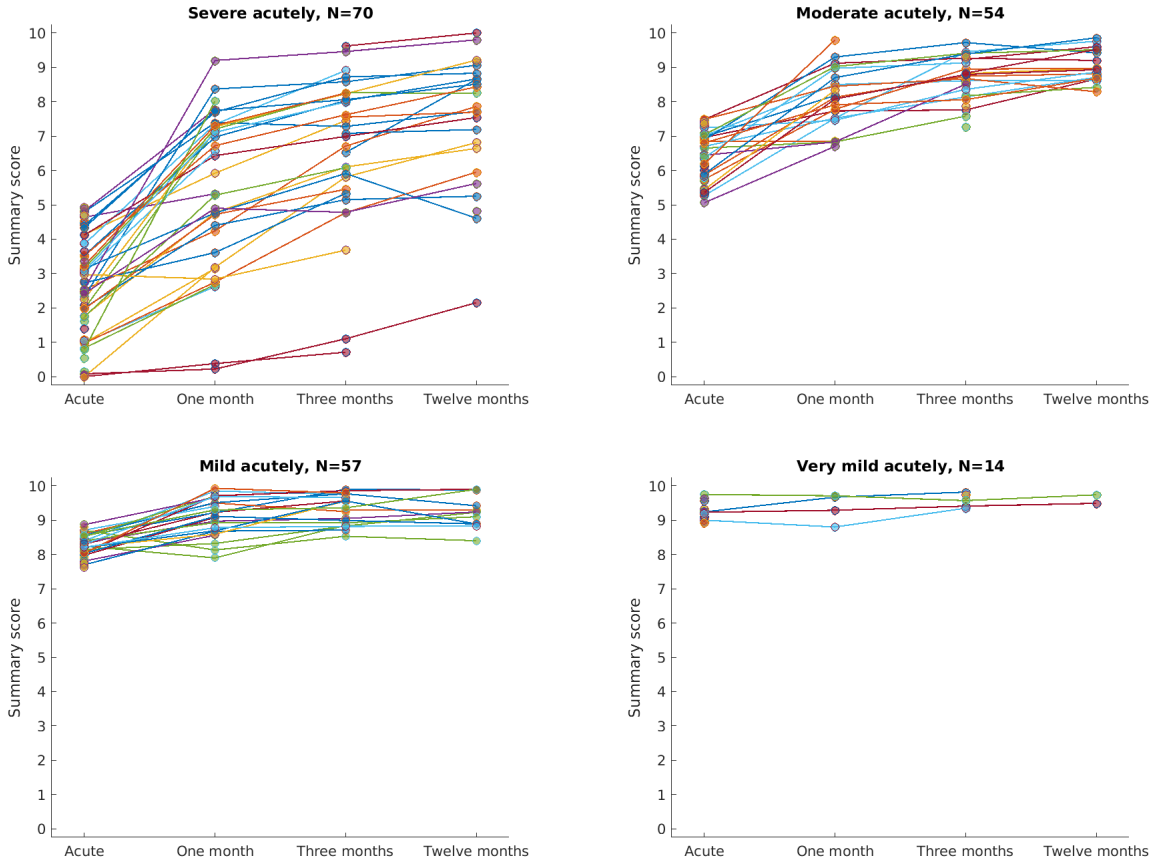

Figure 3.2: Spaghetti plots demonstrating trajectories of recovery across the first year post-stroke at the individual level as measured by QAB overall. Severity groupings are for the purposes of visualization only, based on arbitrary guidelines put forth by the authors of the QAB (0-4.99 = severe, 5-7.49 = moderate, 7.5-8.89 = mild, 8.9-10 = very mild or no aphasia).

continued to show improvements between three months and one year. Scores were estimated to be highest across time points on the dysarthria and single word comprehension domains, and lowest across time points on the sentence comprehension and word finding domains. Note that the purpose of this analysis was intended to be descriptive (that is, to reflect general patterns of language change over time in our particular data set, rather than to generalize to new exemplars); thus, results of statistical tests are not reported or interpreted here.

Correlation coefficients between the nine QAB summary measures of interest were calculated

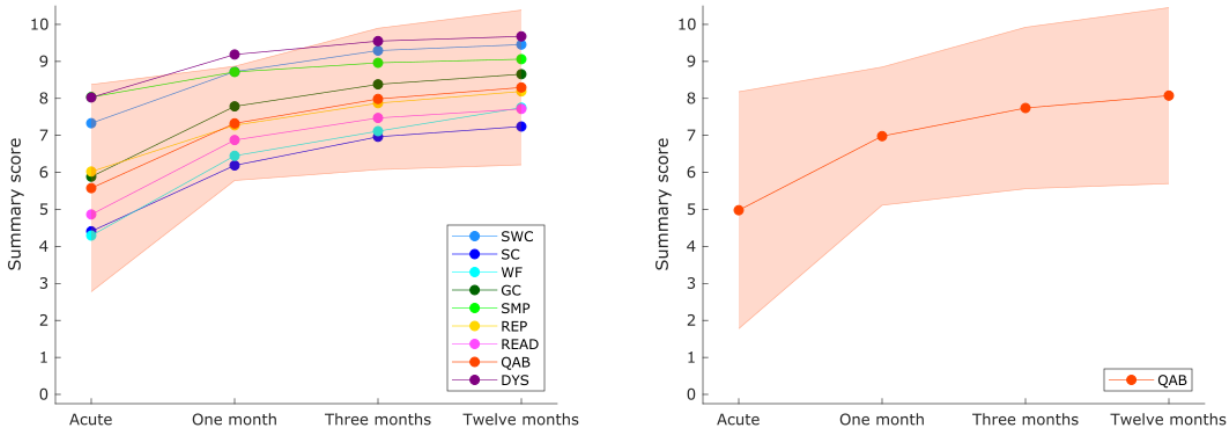

Figure 3.3: Plot of estimated QAB performance as a function of time in individuals who presented with aphasia. Left panel excludes those who were untestable at a given time point while right panel codes them as globally aphasic. Estimates reflect participants with aphasia for whom at least one valid language evaluation was available across time points (N=213 and N=230, respectively). Shaded error bars correspond to standard deviation of QAB overall scores across participants at each time point as calculated in the mixed-effects modeling procedure. Abbreviations are as follows: SWC = single word comprehension; SC = sentence comprehension; WF = word finding; GC = grammatical construction; SMP = speech-motor programming; REP = repetition; READ = reading; DYS = dysarthria; QAB = QAB overall.

across time points in order to examine how relationships between them stayed stable or fluctuated over time (see Figure 3.4). All measures correlated with all others across all time points except for dysarthria, which gradually became less correlated with other scores over time. The highest correlations across time points were observed between word finding and QAB overall, grammatical construction and QAB overall, and repetition and QAB overall. Note that these measures are not entirely independent, as the QAB is a composite score of other summary measures; however, the measures that correlate most highly with QAB overall vary in their contribution to the QAB overall score, with word finding and grammatical construction each contributing 14% and repetition contributing only 8%. Thus, the high correlations between these measures (range: $r = 0.90 - 0.94$) relative to low correlations with other measures that contribute to the QAB overall in similar ways (e.g. speech-motor programming at one month, $r = 0.55$, which like repetition also contributes 8% to QAB overall) may still be meaningfully interpreted. The lowest correlations were generally ob-

served in the presence of the motor speech measures, speech-motor programming and dysarthria, with speech-motor programming and sentence comprehension in particular showing particularly low correlations across all time points (range $r = 0.22 - 0.48$).

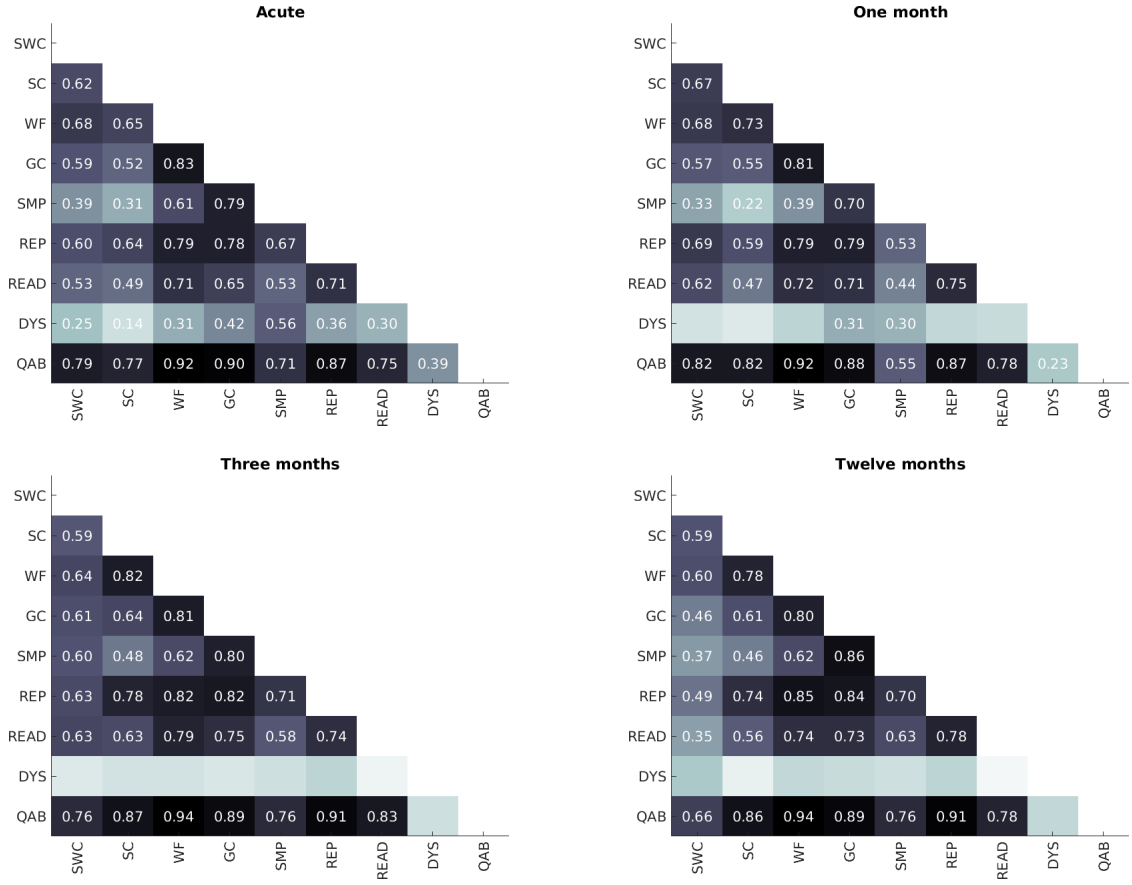

Figure 3.4: Correlations between subscores of the QAB at different time points. Coefficient values are plotted when significant at $p < 0.05$. Abbreviations are as follows: SWC = single word comprehension; SC = sentence comprehension; WF = word finding; GC = grammatical construction; SMP = speech-motor programming; REP = repetition; READ = reading; DYS = dysarthria; QAB = QAB overall.

3.1.2 Imaging

Lesion overlays demonstrated near-complete coverage of the left hemisphere in our data set, as well as distinct lesion distributions across patients with and without aphasia (see Fig. 3.5). Regions of maximum overlap associated with an aphasia diagnosis fell in the left external capsule (85/220)

and left insula (83/220), while the region of maximum overlap associated with a diagnosis of no aphasia fell in the left putamen (23/119).

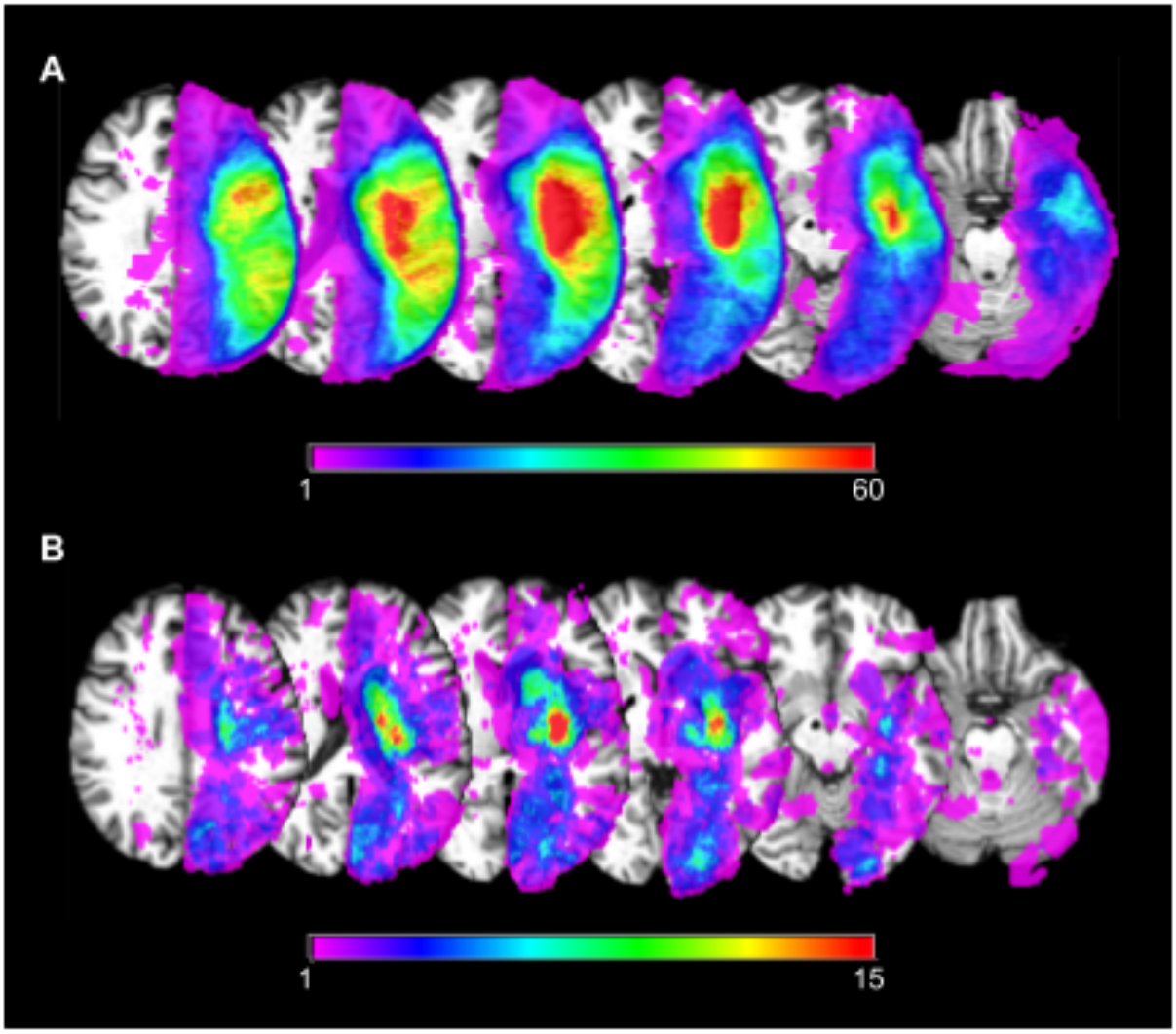

Figure 3.5: Lesion overlays for (A) individuals with aphasia (N=220) and (B) individuals without aphasia (N=119). Color for each voxel corresponds to the number of patients in whom that voxel is lesioned.

In order to assess correlations between lesion size and damage to different ROIs, a correlation matrix was generated including lesion size as well as damage to all left hemisphere regions of interest in participants with aphasia for whom imaging was available (see Fig. 3.6; for list of abbreviations, see Table 3.3). Lesion size correlated maximally with damage to the superior tem-

poral gyrus, the post-central gyrus, and white matter regions. The majority of high correlations fell within a superset region (i.e., pars opercularis and pars triangularis, both within the realm of the IFG, were highly correlated), although other high correlations were observed (e.g. insula with inferior frontal gyrus and paracentral lobule). Note that *ext* images are not reflected in this plot.

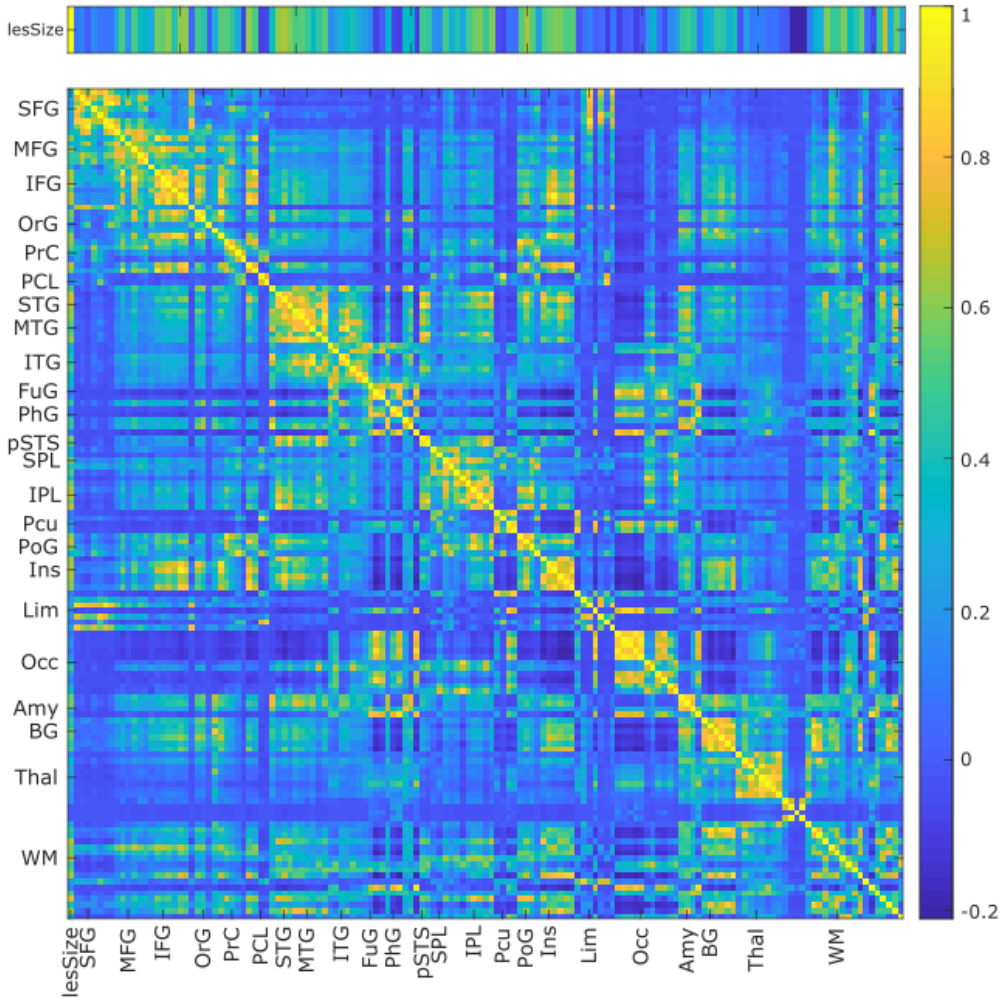

Figure 3.6: Correlations between lesion size and damage to left hemisphere ROIs in people with aphasia ($N = 220$). To aid in visualization and readability, only superset regions are labeled; see Table 3.3 for meanings of abbreviations, and Table 1 in Fan et al. (2016) for further description of comprising subregions.

Abbreviation	Expanded name	Num. LH subregions
SFG	Superior frontal gyrus	7
MFG	Middle frontal gyrus	7
IFG	Inferior frontal gyrus	6
OrG	Orbital gyrus	6
PrC	Precentral gyrus	6
PCL	Paracentral lobule	2
STG	Superior temporal gyrus	6
MTG	Middle temporal gyrus	4
ITG	Inferior temporal gyrus	7
FuG	Fusiform gyrus	3
PhG	Parahippocampal gyrus	6
pSTS	Posterior superior temporal sulcus	2
SPL	Superior parietal lobule	5
IPL	Inferior parietal lobule	6
Pcu	Precuneus	4
PoG	Postcentral gyrus	4
Ins	Insula	6
Lim	Limbic lobe (cingulate gyrus)	7
Occ	Occipital lobe	11
Amy	Amygdala/hippocampus	4
BG	Basal ganglia	6
Thal	Thalamus	8
WM	White matter	21

Table 3.3: Table of ROI abbreviations as extracted from Fan et al. (2016) and Mori et al. (2005). Far right column refers to number of subregions comprising each ROI in the left hemisphere of the combined atlas. See Table 1 of Fan et al. (2016) for further detail.

3.2 Linear SVR model performance

As discussed in the Methods section above, a variety of different models reflecting different analytical choices were created for the purposes of predicting language outcomes at various timepoints post-stroke. A recap of the characteristics of each of these models is included prior to the presentation of their results to increase interpretability; Table 2.2 may also be consulted for this purpose.

3.2.1 Models 1 through 4: Cognitive neuroscience focus

Models 1 through 4 address a cognitive neuroscience question by using only demographic and lesion-related predictors, excluding information about acute language scores. Models 1 and 3 treat untestable patients as missing while Models 2 and 4 treat them as globally aphasic (such that only QAB overall can be predicted, due to a lack of detail regarding specific characteristics of language); Models 1 and 2 include only individuals with aphasia while Models 3 and 4 include individuals both with and without aphasia (such that only the acute time point can be predicted, due to lack of follow-up of those without aphasia).

Figures 3.7 and 3.8 show model performance for models addressing the extent to which language outcomes at different time points can be predicted from demographic and lesion information alone. Demographic and stroke type information alone were insufficient to accurately predict language outcomes for any time point or language domain, regardless of methods for filtering the input data (i.e. inclusion or exclusion of individuals without aphasia or who were untestable), with the maximum ICC achieved across all Stage 1 models being 0.08 (Model 3, acute sentence comprehension). The addition of lesion size as a predictor in the Stage 2 models led to significant improvements in prediction of outcomes at all time points across all data filtering strategies and language domains except for speech-motor programming and dysarthria (though the addition of lesion size marked a significant increase in accuracy for dysarthria at the acute time point, the absolute accuracy was still extremely low), though all resulting predictions still remained below Cicchetti's standards for good or excellent reliability (Cicchetti, 1994). With the addition of lesion load information in Stage 3 came statistically significant improvements in all but one model (Model 1 reading at three months), such that 19 out of 50 models reached "good" reliability per Cicchetti (Model 1 sentence comprehension at one month and twelve months, word finding at three months and twelve months, grammatical construction at three months and twelve months, repetition acutely, overall at three months and twelve months; Model 2 overall at all time points; Model 3 word finding, grammatical construction, repetition, reading, and overall acutely; Model 4 overall acutely).

Across all models addressing the neural bases of language, the maximum predictive accuracy (ICC =0.73) was achieved when predicting QAB overall score at the one year time point using all available predictors and excluding individuals without aphasia (Models 1 and 2, Stage 3, which show identical results at the one year time point due to the absence of untestable patients). Figure 3.9 shows representative raw scatter plots of actual versus predicted overall QAB scores as generated by Model 1.

3.2.1.1 Neural correlates and other predictors of recovery

In order to get a sense of which regions are associated with language recovery in the long-term, beta weights from high-performing Model 1 models at the one year time point are displayed in Figure 3.10, with ROIs associated with negative beta weights (that is, ROIs in which larger amounts of damage were associated with lower than average language scores) plotted in hot colors and ROIs associated with positive beta weights (that is, ROIs in which larger amounts of damage were associated with higher than average language scores) plotted in cool colors.

Non-lesion-load-related predictors were interpreted as potentially associated with outcomes if their assigned beta weights were as extreme or more than two standard deviations from the mean beta weight across all predictors in the model. By this metric, years of education were positively associated with sentence comprehension (β=1.00) and grammatical construction (β=0.85) outcomes, while age was negatively associated with word finding outcomes (β=-1.18).

3.2.2 Models 5 through 8: Clinical focus

Models 5 through 8 address a clinical question by using all available predictors, including information about acute language scores, to predict language at later time points; thus, only post-acute scores in individuals with aphasia are predicted, and individuals without aphasia (who were not followed up after the acute stage) are excluded. Models 5 and 7 treat untestable patients as missing while Models 6 and 8 treat them as globally aphasic (such that only QAB overall can be predicted, due to a lack of detail regarding specific characteristics of language); Models 5 and 6 predict outcomes at particular time points while Models 7 and 8 predict score *change* between the acute and

later time points; see again Table 2.2.

Figures 3.11 and 3.14 show model performance for models addressing the extent to which language recovery can be predicted given information about acute language presentation in addition to demographic and lesion information. As the two sets of models address this question using different methods—namely, by predicting outcome versus by predicting change/recovery—their results will be discussed separately here.

3.2.2.1 Predicting outcomes at follow-up

Figure 3.11 shows model performance for models addressing the extent to which a later cross-sectional language score can be predicted given information about acute presentation. Stage 1 models including only demographic information, stroke type, and acute language score were already highly predictive of longer-term outcomes at earlier time points, with word finding, grammatical construction, speech-motor programming, repetition, reading, dysarthria, and QAB overall at the one month time point, as well as QAB overall at the three month time point, already well-predicted per the Cicchetti (1994) standards for "good" reliability. These findings held whether individuals who were untestable were coded as missing or globally aphasic. The addition of lesion size information in Stage 2 still contributed to prediction, however, with significant increases observed for single word comprehension at one month and one year, sentence comprehension at one month and three months, word finding and grammatical construction at all time points, repetition at one month and three months, reading at three months, and overall score at all time points. The further addition of lesion load information in Stage 3 significantly increased predictive accuracy for single word comprehension at one year, sentence comprehension at all follow-up time points, word finding at one year, repetition at one year, dysarthria at all follow-up time points, and overall QAB at the one year time point (as well as the three month time point when untestable patients were included in the data set). Note that the majority of the cases in which predictions were improved by the presence of lesion load information were at later time points post-stroke. Figure 3.12 shows representative raw scatter plots of actual versus predicted overall QAB scores as generated

using Model 6.

3.2.2.2 Neural correlates and other predictors of recovery

In order to get a sense of which regions are associated with language recovery in the long-term when initial presentation is accounted for, beta weights from high-performing Model 5 models at the one year time point are displayed in Figure 3.13, with negative beta weights (ROIs in which larger amounts of damage were associated with lower than average language scores) plotted in hot colors and positive beta weights (ROIs in which larger amounts of damage were associated with higher than average language scores) plotted in cool colors.

Non-lesion load related predictors were again interpreted as potentially associated with outcomes if their assigned beta weights were as extreme or more than two standard deviations from the mean beta weight across all predictors in the model. By this metric, acute score was positively associated with sentence comprehension (β=1.14), word finding (β=1.18), and QAB overall (β=1.01). Age was negatively associated with sentence comprehension (β=-0.94), word finding (β=-1.41), and QAB overall (β=-0.76).

3.2.2.3 Predicting change at follow-up

Figure 3.14 shows model performance for models addressing the extent to which recovery of language (that is, change in language score between the acute and later time point) can be predicted at later time points given information about acute presentation. ICC values indicated that prediction of change at Stage 1 was, in many cases, already excellent per Cicchetti's standards, with few significant increases in ICC as further predictors were added.

Note, however, that correlations between actual and predicted change scores will be inflated when variance is lower in outcome than baseline scores see (as is the case in our data and most stroke data sets; see Hope et al., 2019; Hawe et al., 2019; Bowman et al., 2021). Thus, it is likely to paint a rosier picture of prediction accuracy than truly is warranted. Per Kundert et al. (2019), it is a "common but inaccurate assumption" (p. 885) that being able to predict change means being able to predict outcomes by adding predicted change to baseline scores, and indeed the falsity of

this supposition is borne out in the figures and analysis described below.

Figure 3.16 shows model performance for the same models assessed in Figure 3.14, but with ICC calculated between actual outcomes and *predicted outcomes based on predicted change* (that is, actual acute score plus predicted change) rather than actual change and predicted change. Note that many of the extremely high ICC values observed in Figure 3.14, for example single word comprehension, grammatical construction, and dysarthria at the three and twelve month time points, have lowered in this version of the accuracy analysis.

Across all models in the Model 7, Stage 3 class, acute score was the highest predictor of change scores, with β values ranging from -4.53 to -6.26, all more extreme than those observed for any predictors in the previously reported models.

To further illustrate this point with regard to the apparent spuriousness of calculating ICC on change scores, let us briefly return to Stage 1 of Model 5 (demographics, stroke type, and acute score predicting cross-sectional outcomes), with a focus on single word comprehension. Predictions of single word comprehension outcomes were quite poor at this stage, with the greatest ICC being attained at the three month time point ($ICC = 0.23$). However, if we were to use this very same model but instead calculate ICC between predicted and actual *change* at the three month time point—that is, between $(actualOutcome - actualAcute)$ and $(predictedOutcome - actualAcute)$—our ICC suddenly skyrockets to $ICC = 0.90$. Clearly the model itself is not superior for practical purposes when we calculate accuracy in this manner; the correlation-based metric simply exploits the synthetic variance generated by subtracting the highly variable true baselines from the highly *invariable* predicted outcomes, which are most pronounced when ceiling effects are present (see Fig. 3.18).

3.3 Non-linear SVR model performance

Finally, a non-linear version of this analysis was completed using an RBF kernel per the recommendations of Zhang et al. (2014) and DeMarco and Turkeltaub (2018). This approach allows for the possibility of interactions between features/predictors and has the potential to improve prediction

accuracy, but at the expense of providing clearly interpretable beta weights and a clear path forward for statistical comparison with the linear models reported above. The highest predictive accuracies across these models in general were generated using Models 7 and 8 (with ICCs calculated between actual outcomes and change-based predicted outcomes, since accuracy for predicting change alone is difficult to interpret; see discussion above), with excellent accuracy reached for word finding at the one month time point, repetition at the three month time point, and QAB overall at the one and three month time points. Most other predictive accuracies fell into the "good" range, with the exception of single word comprehension at one and twelve months, grammatical construction and repetition at twelve months, and reading and dysarthria at three and twelve months, which were fair to poor per Cicchetti's standards. Results are plotted in Figure 3.19.

As stated above, there is not a straightforward way to statistically compare linear and non-linear models using the permutation approach taken by this study. To provide a general sense of any potential improvements in accuracy using a non-linear RBF versus a linear kernel in the models using all predictors, raw differences in ICC between Stage 4 RBF-based models and Stage 3 linear models are plotted in Figure 3.20. Though some models showed improved accuracy with the use of the non-linear kernel, the benefits were not strikingly apparent across language domains or time points.

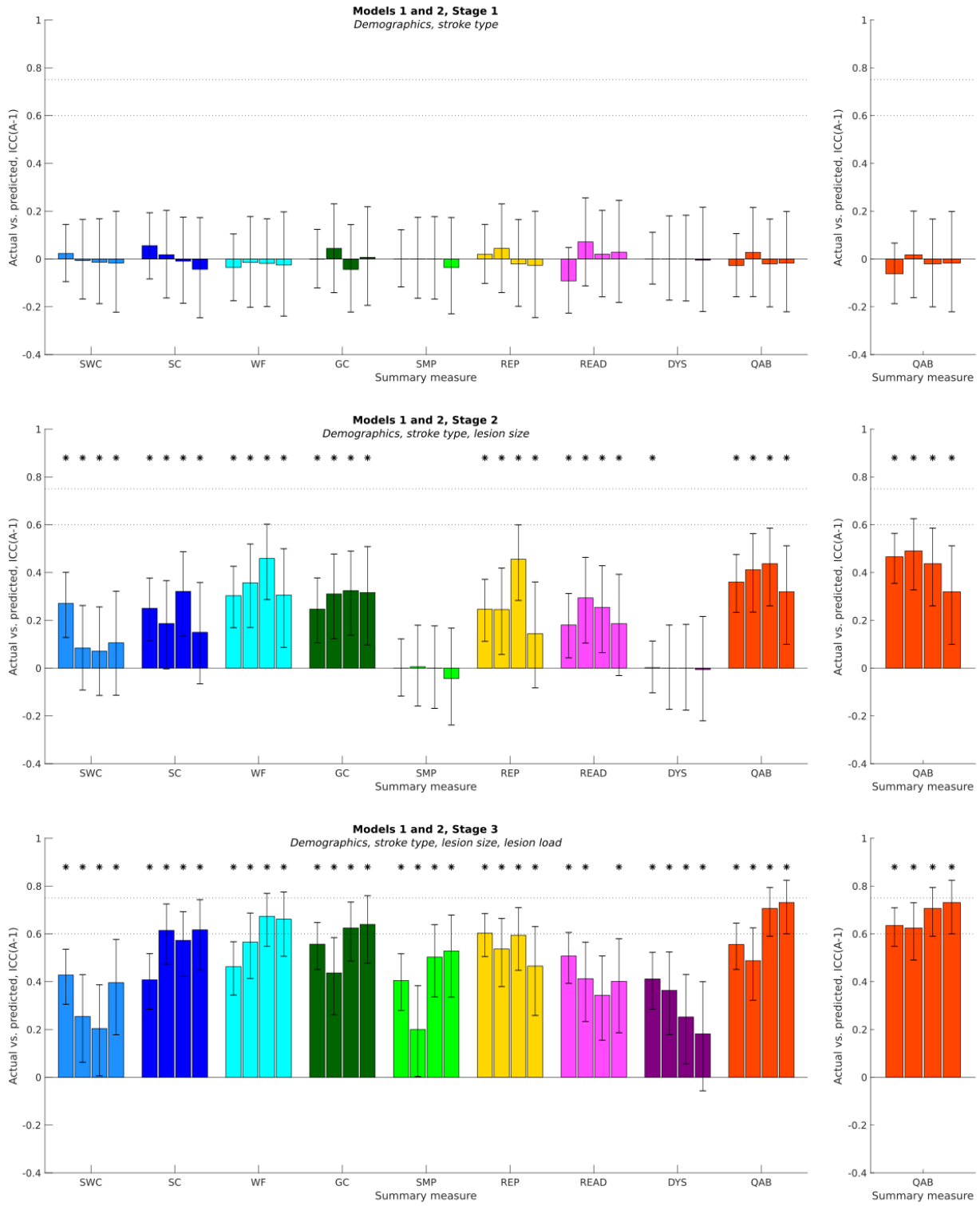

Figure 3.7: Plots of model performance (ICC) attained for Models 1 (left) and 2 (right) across domains (color) and times (bar within group; left = earliest) at different stages (rows). Dotted lines show "good" and "excellent" ICCs. Error bars show 95% CIs (parametric). Asterisks denote models on which new predictors significantly improved accuracy. SWC=single word comprehension, SC=sentence comprehension, WF=word finding, GC=grammatical construction, SMP=speech-motor programming, REP=repetition, READ=reading, DYS=dysarthria, QAB=overall.

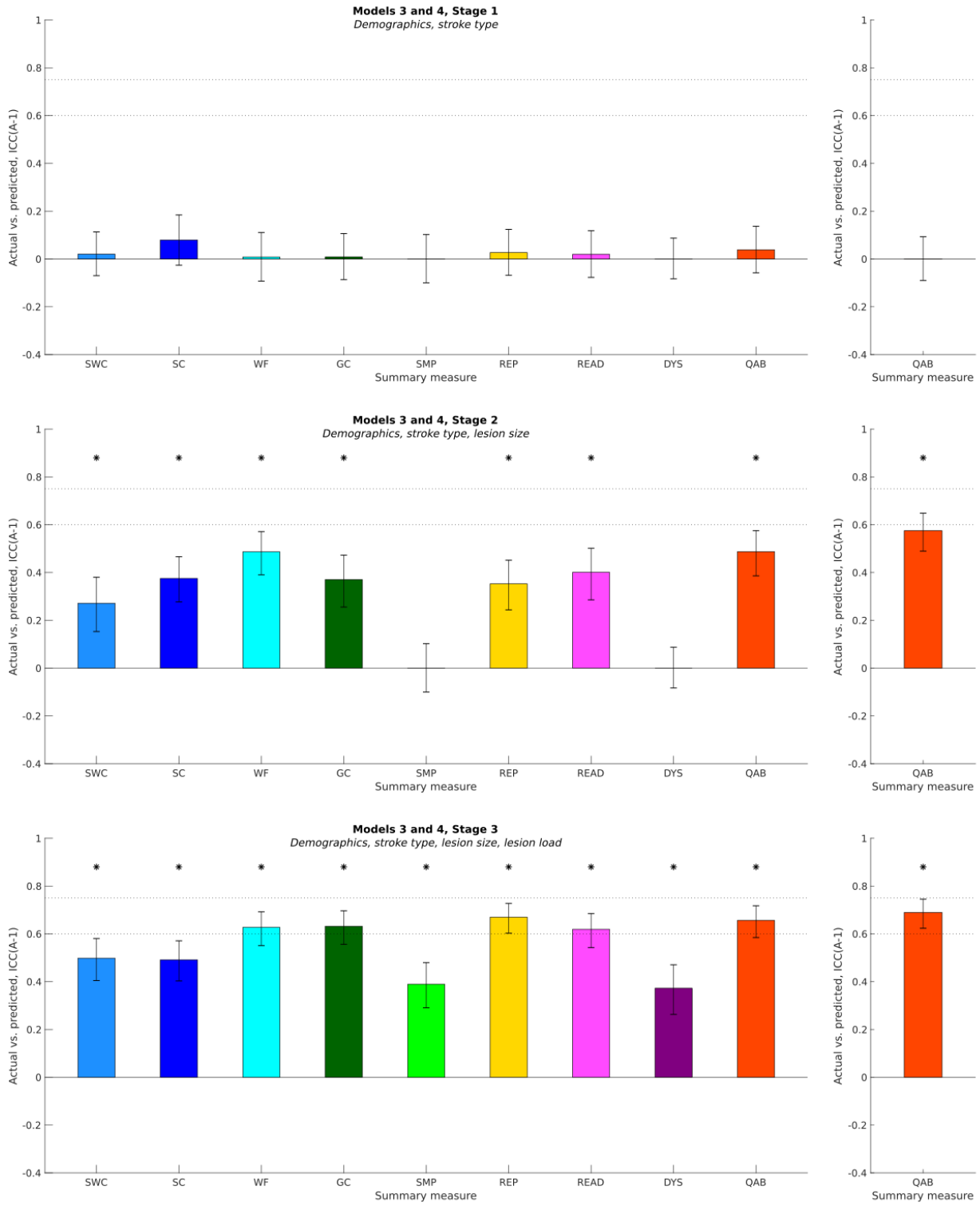

Figure 3.8: Plots of model performance (ICC) attained for Models 3 (left) and 4 (right) across domains (color) and times (bar within group; left = earliest) at different stages (rows). Dotted lines show "good" and "excellent" ICCs. Error bars show 95% CIs (parametric). Asterisks denote models on which new predictors significantly improved accuracy. SWC=single word comprehension, SC=sentence comprehension, WF=word finding, GC=grammatical construction, SMP=speech-motor programming, REP=repetition, READ=reading, DYS=dysarthria, QAB=overall.

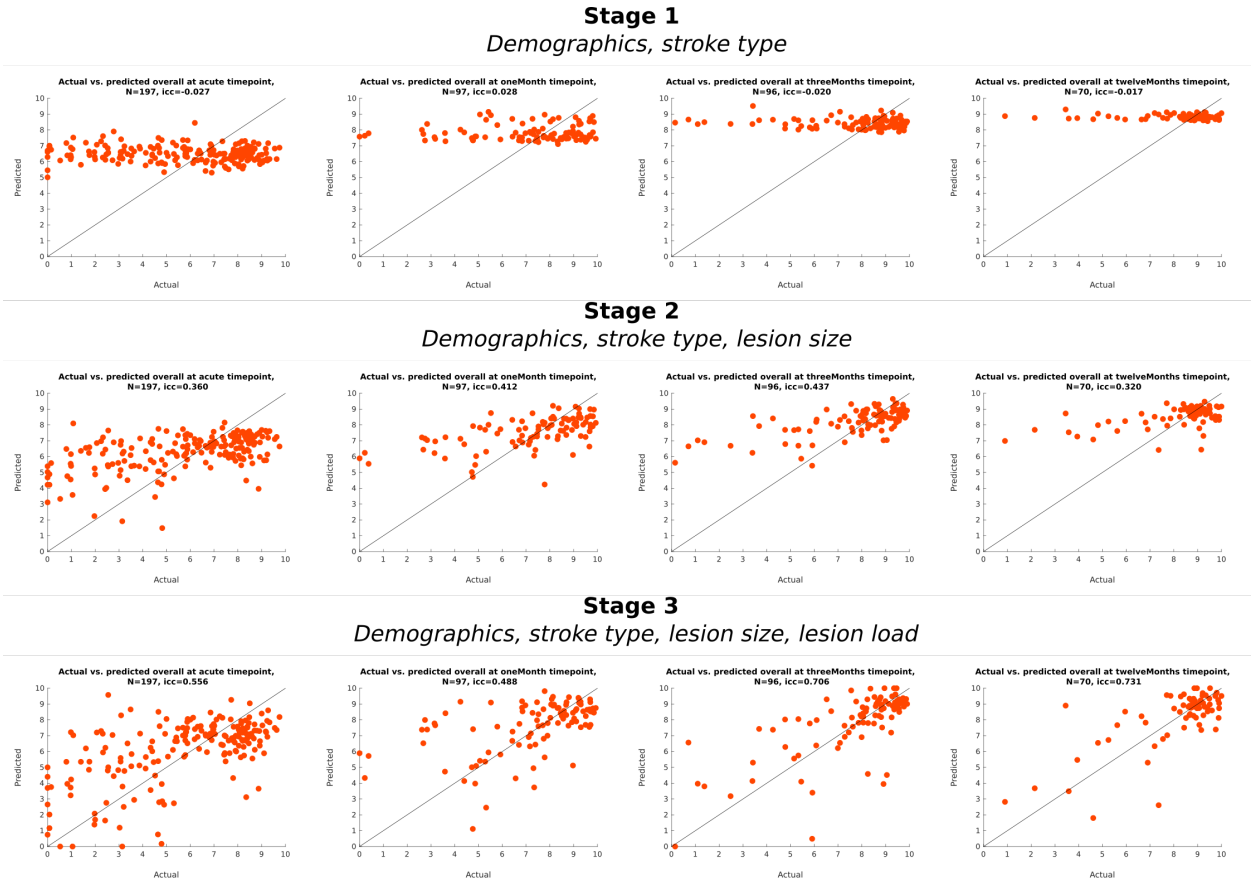

Figure 3.9: Representative scatter plots across model stages (rows) and times (columns) demonstrating increasing accuracy at predicting QAB overall score with lesion load included as a predictor. Scatters are drawn from Model 1, which excludes both patients without aphasia and untestable patients. Diagonals (black) correspond to the identity line (perfect correlation).

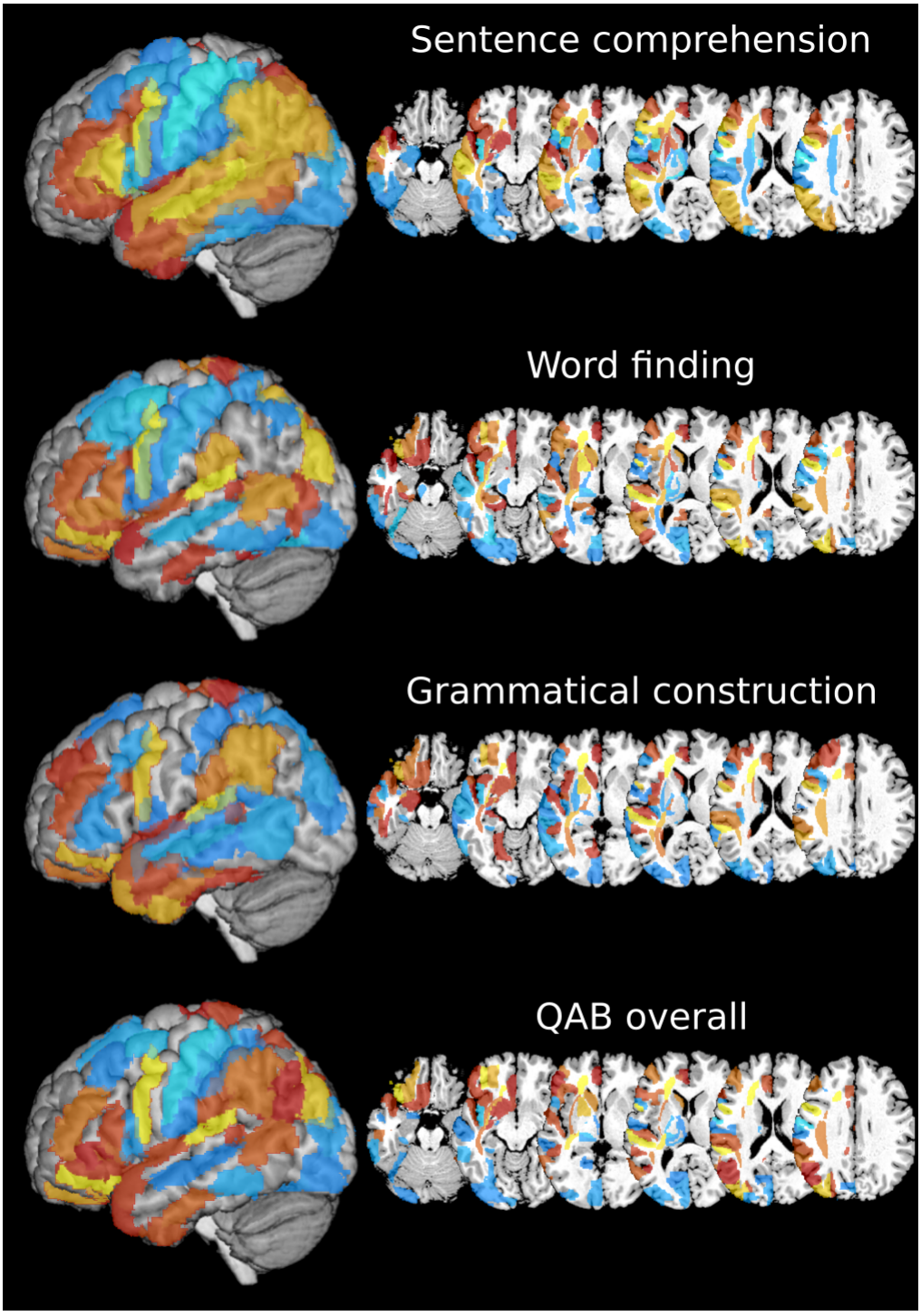

Figure 3.10: Regions of interest (ROIs) implicated for high-performing Model 1, Stage 3 models (all predictors except acute score, participants with aphasia only) at the one year time point. Hot colors reflect ROIs which were assigned negative beta weights (meaning damage was associated with worse than average scores); cool colors reflect ROIs which were assigned positive beta weights (meaning damage was associated with better than average scores). Maps are thresholded to show betas with values more extreme than 0.2 and capped at 1.0.

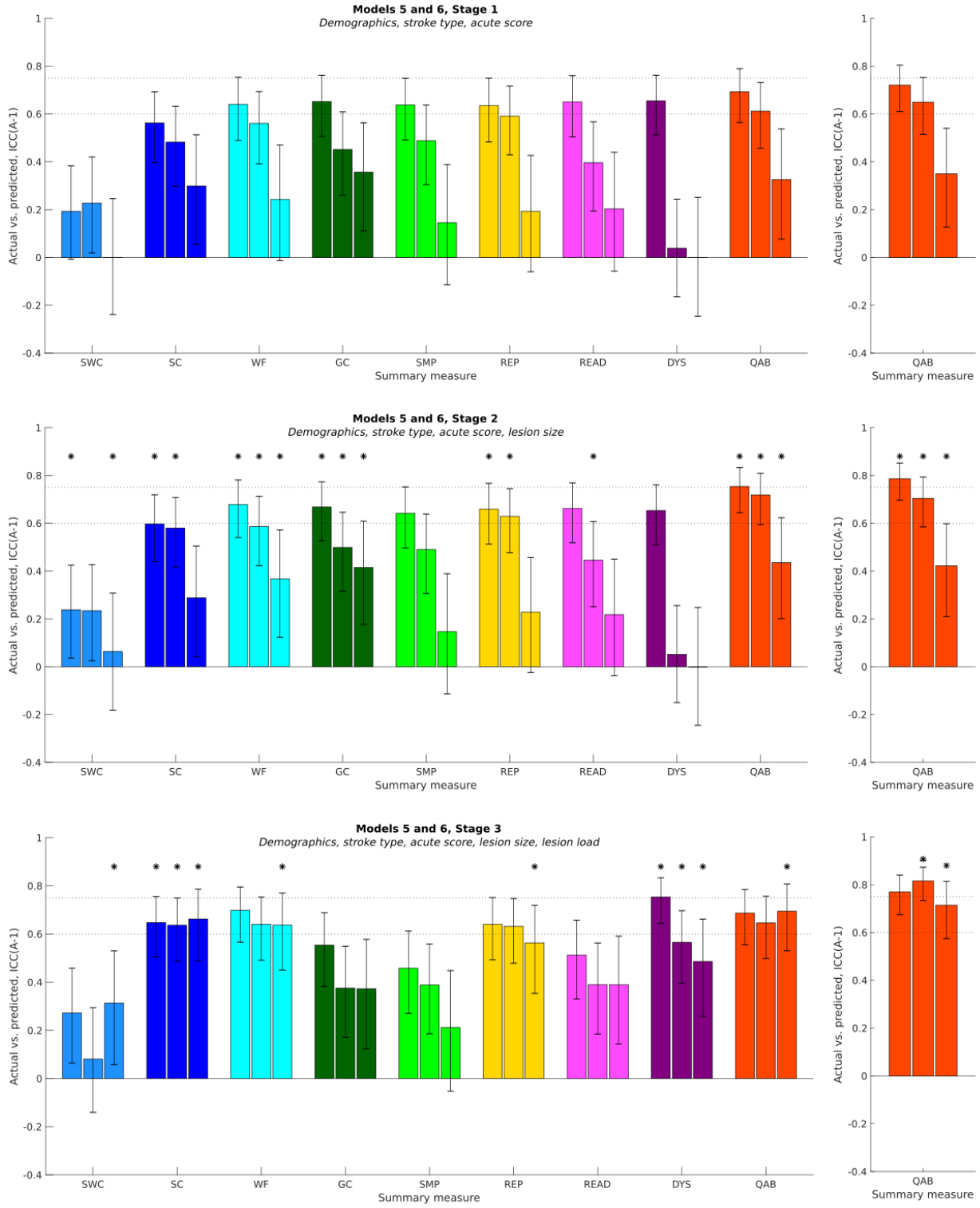

Figure 3.11: Plots of model performance (ICC) attained for Models 5 (left) and 6 (right) across domains (color) and times (bar within group; left = earliest) at different stages (rows). Dotted lines show "good" and "excellent" ICCs. Error bars show 95% CIs (parametric). Asterisks denote models on which new predictors significantly improved accuracy. SWC=single word comprehension, SC=sentence comprehension, WF=word finding, GC=grammatical construction, SMP=speech-motor programming, REP=repetition, READ=reading, DYS=dysarthria, QAB=overall.

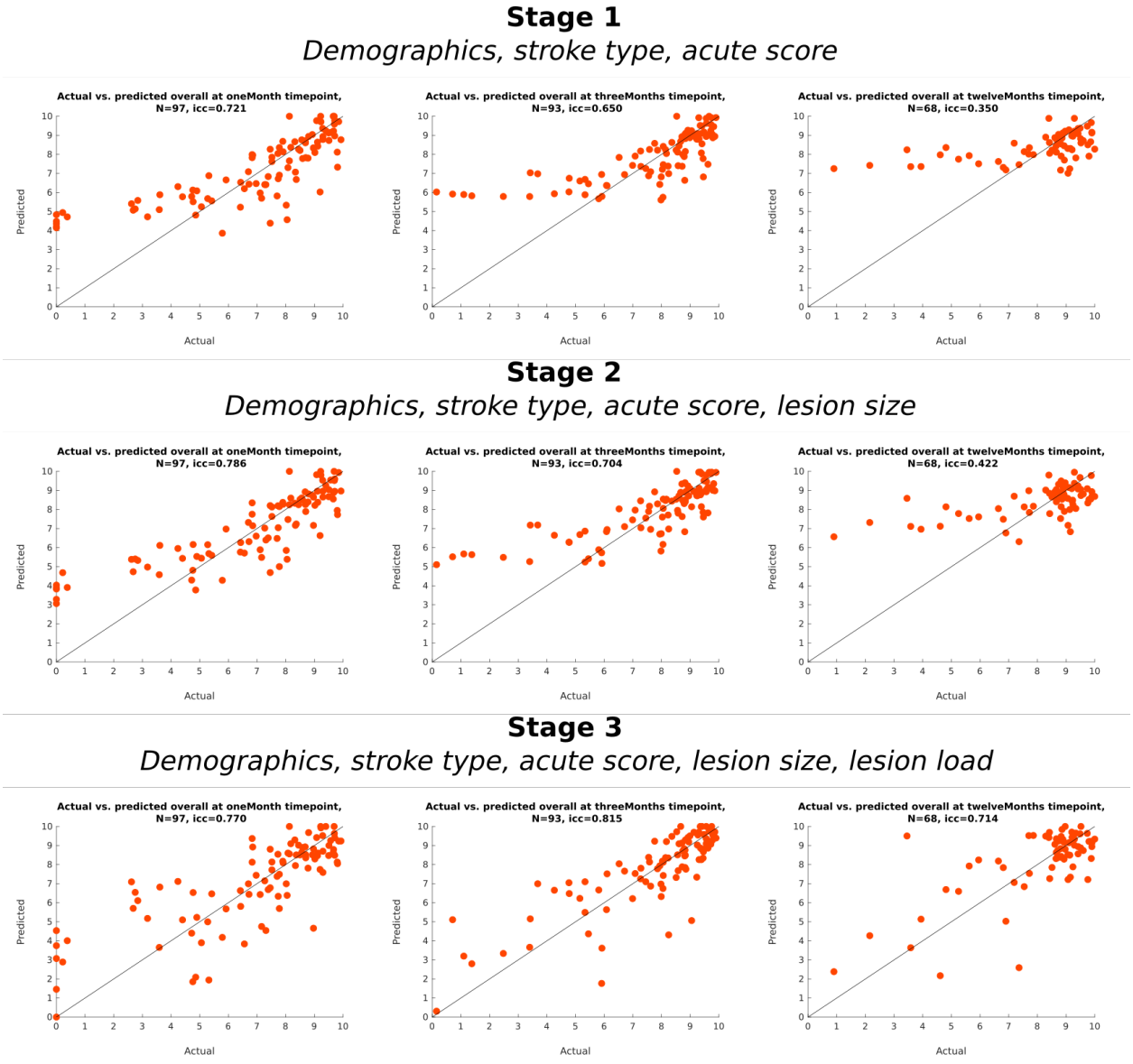

Figure 3.12: Representative scatter plots across model stages (rows) and times (columns) demonstrating generally high accuracy at predicting QAB overall score across model stages when acute score is included as a predictor, but with increased accuracy at later time points when lesion load information is added to the model. Scatters are drawn from Model 6, which includes untestable patients by treating them as globally aphasic. Diagonals (black) correspond to the identity line (perfect correlation).

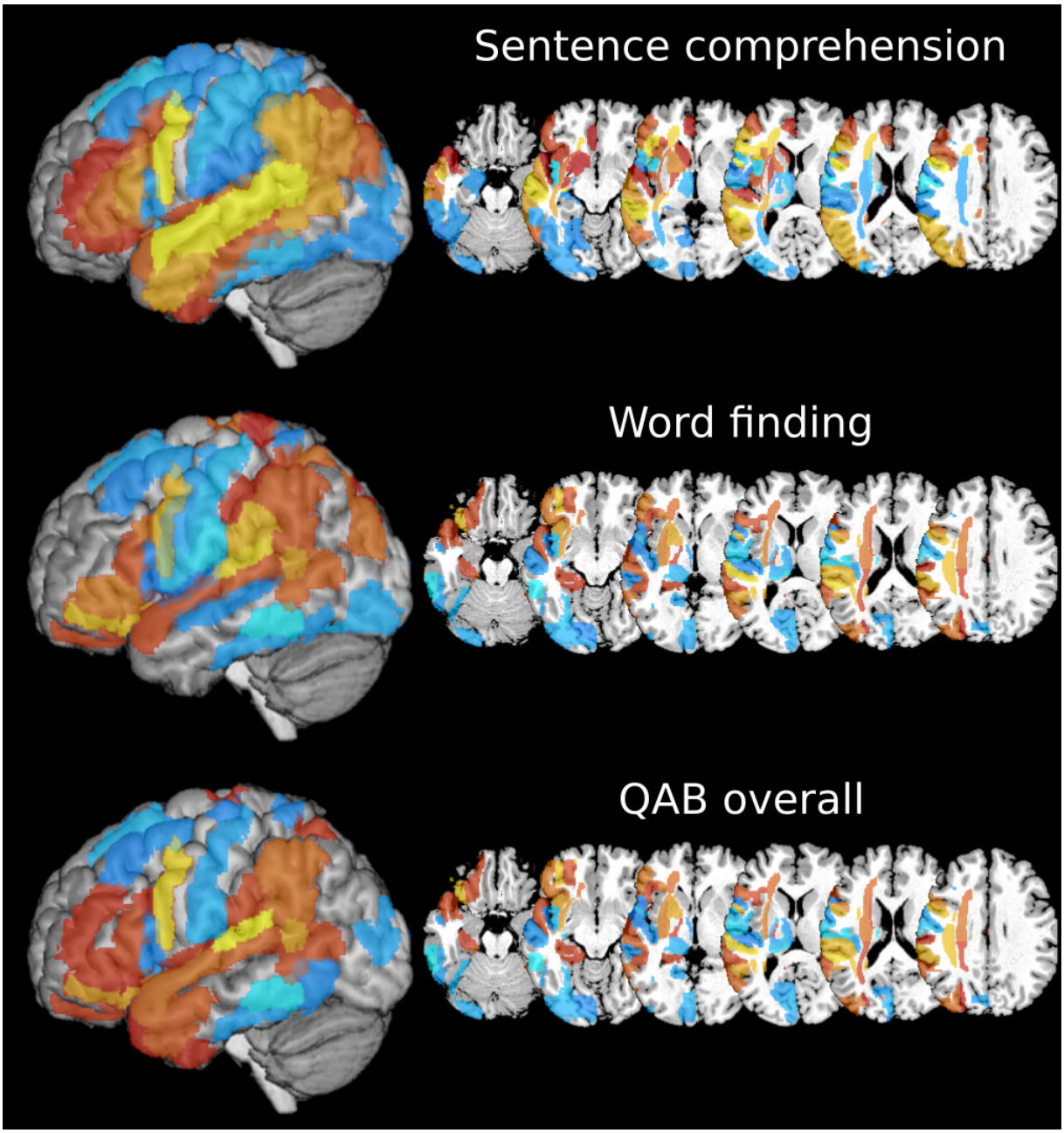

Figure 3.13: Regions of interest (ROIs) implicated for high-performing Model 5, Stage 3 models (all predictors including acute score, with untestable patients treated as missing) at the one year time point. Hot colors reflect ROIs which were assigned negative beta weights (meaning damage was associated with worse than average scores); cool colors reflect ROIs which were assigned positive beta weights (meaning damage was associated with better than average scores). Maps are thresholded to show betas with values more extreme than 0.2 and capped at 1.0.

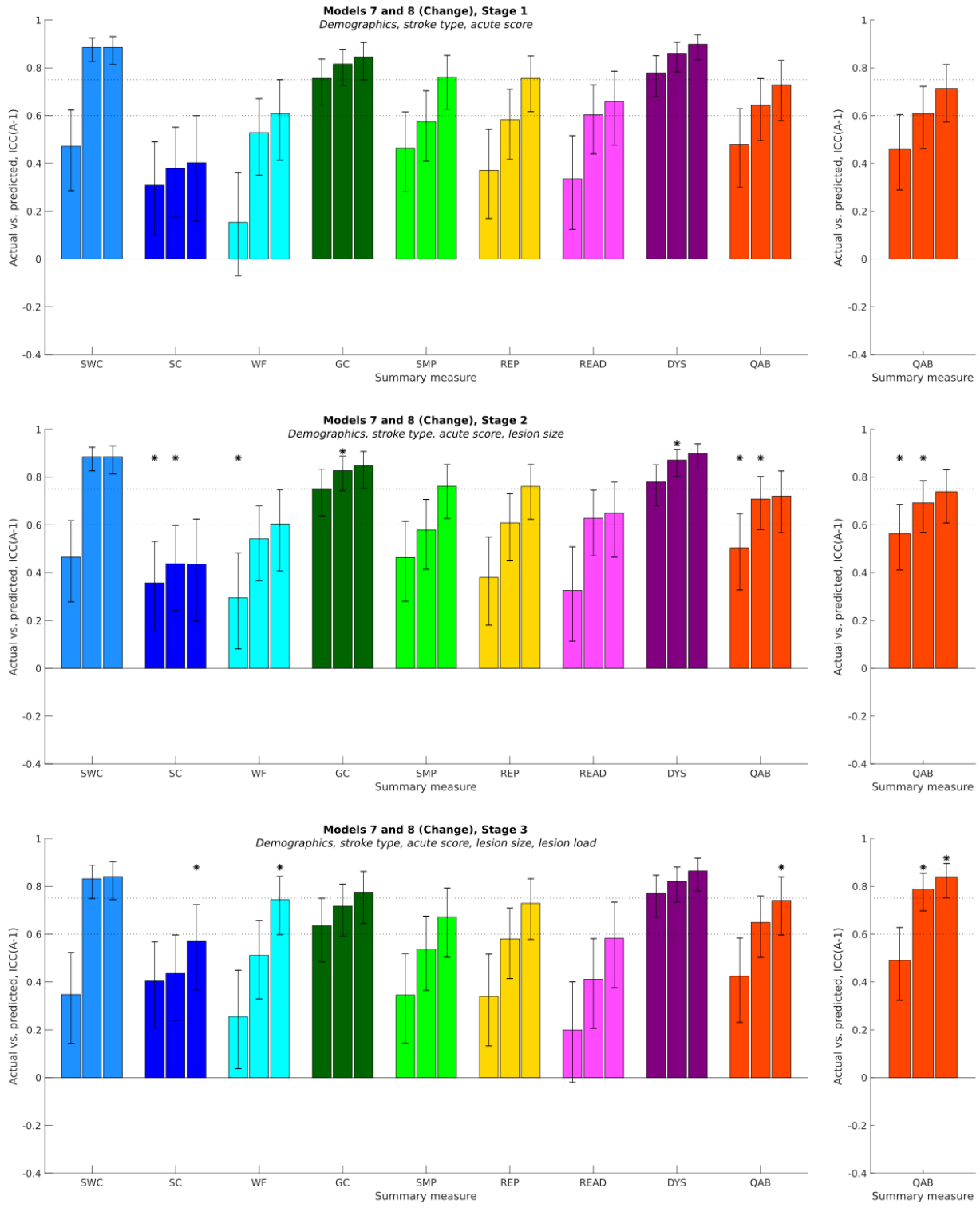

Figure 3.14: Plots of ICCs between actual change and predicted change for Models 7 (left) and 8 (right) across domains (color) and times (bar within group; left = earliest) at different stages (rows). Note validity concerns of this approach. Error bars show 95% CIs (parametric). Asterisks denote models on which new predictors significantly improved accuracy. SWC=single word comprehension, SC=sentence comprehension, WF=word finding, GC=grammatical construction, SMP=speech-motor programming, REP=repetition, READ=reading, DYS=dysarthria, QAB=overall.

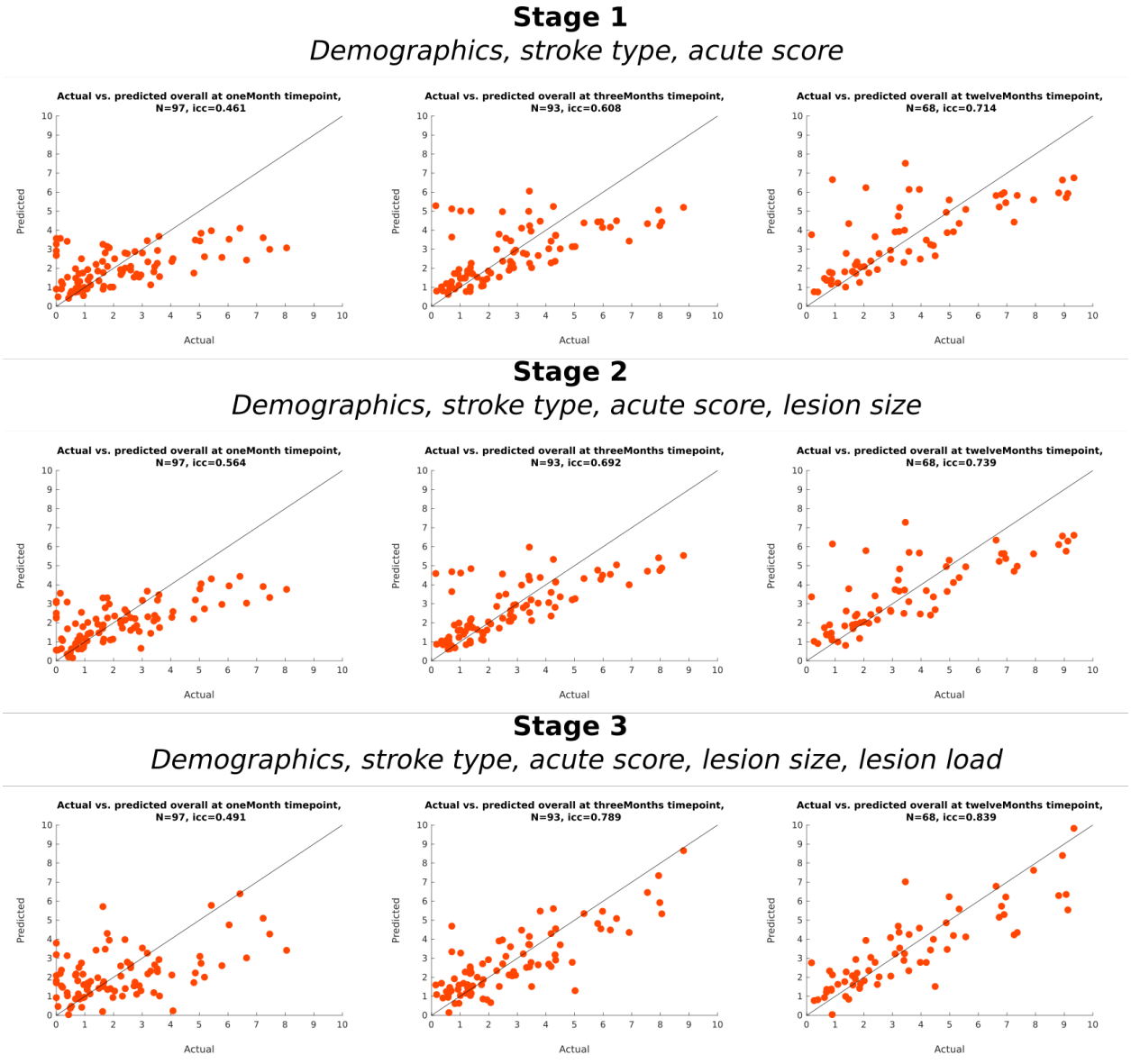

Figure 3.15: Representative scatter plots across model stages (rows) and times (columns) demonstrating generally high accuracy at predicting change in QAB overall score between acute and later time points across model stages when acute score is included as a predictor. Note validity concerns of this approach. Scatters are drawn from Model 8, which includes untestable patients by treating them as globally aphasic. Diagonals (black) correspond to the identity line (perfect correlation).

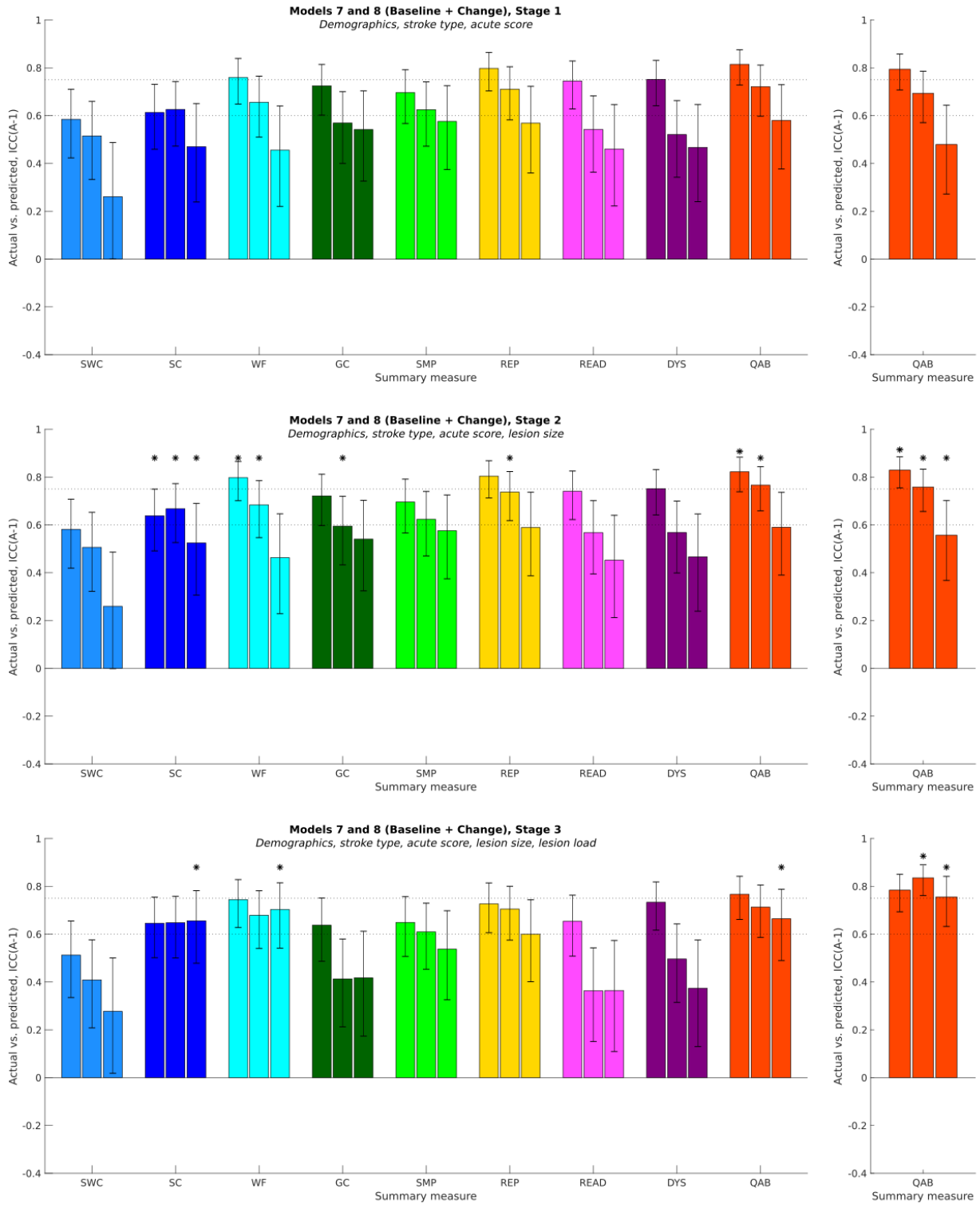

Figure 3.16: Plots of model performance (ICC) attained between actual outcome and *predicted outcome based on change* for Models 7 (left) and 8 (right) across language domains (color) and times (location within group; left = earliest) at different stages (rows), revealing the spuriousness of the results in Fig. 3.14. Asterisks show significance of new predictors. SWC=single word comprehension, SC=sentence comprehension, WF=word finding, GC=grammatical construction, SMP=speech-motor programming, REP=repetition, READ=reading, DYS=dysarthria, QAB=overall.

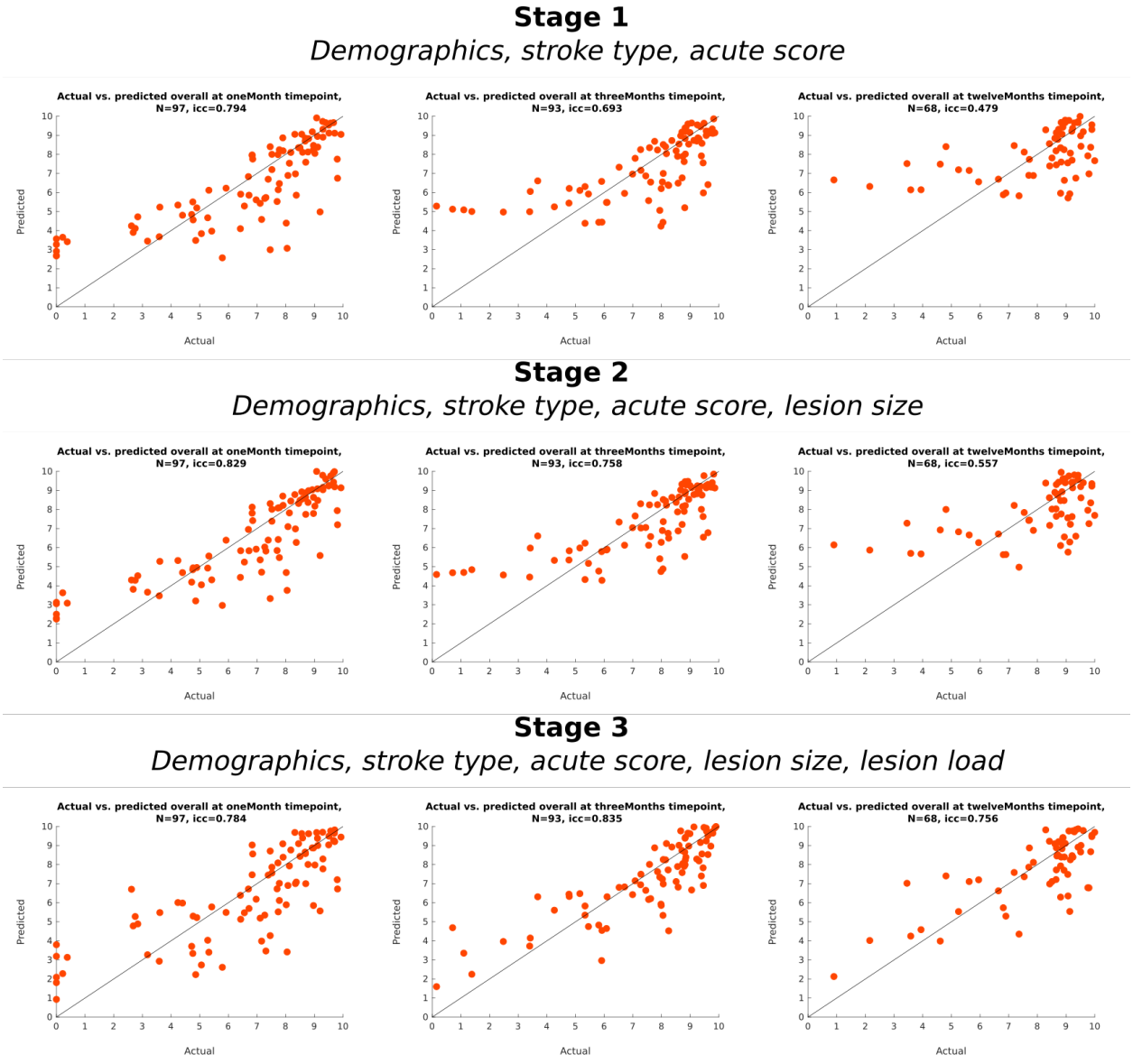

Figure 3.17: Representative scatter plots across model stages (rows) and times (columns) demonstrating generally high accuracy at predicting outcomes using predicted change scores with acute score as a predictor, but with different findings relative to predicting change alone (see Fig. 3.15). Scatters show actual outcomes versus predicted outcomes as generated by adding actual baselines to changes predicted by Model 8. Diagonals (black) correspond to the identity line (perfect correlation).

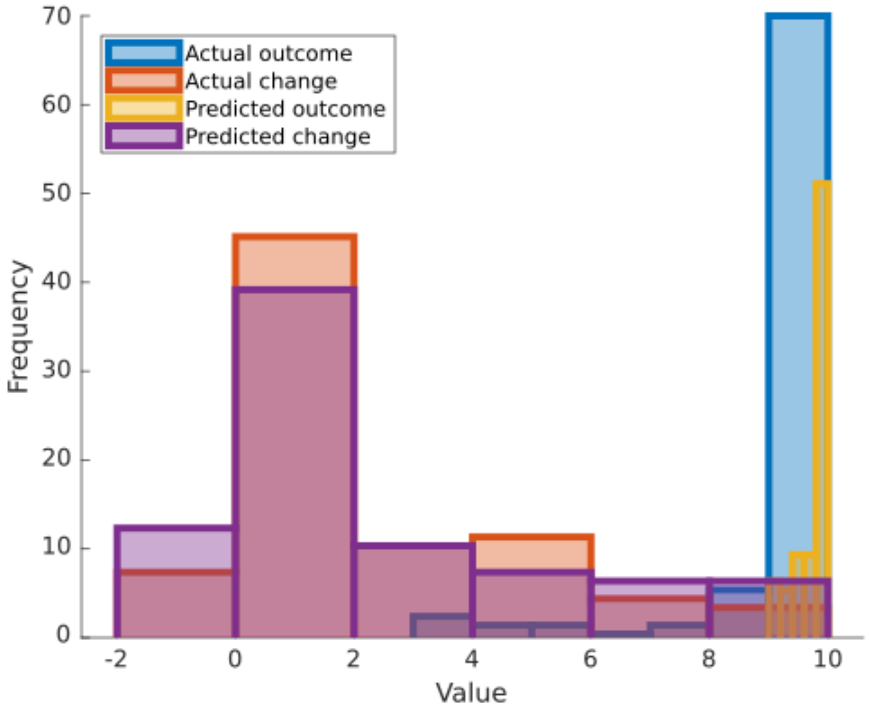

Figure 3.18: An illustration of why ICCs calculated between actual and predicted change might be inflated relative to those calculated between actual and predicted outcomes. Data reflect predictions of sentence comprehension at three months post-stroke from Model 5, Stage 1. Actual outcomes (blue) reflect ceiling effects in the true data, such that predicted outcomes (yellow) occupy a very small range between 9 and 10 ($ICC = 0.22$ for actual versus predicted outcomes). However, when the true baseline score is subtracted from actual and predicted outcomes (orange and purple, respectively), the invariance is obliterated and actual versus predicted change appear to be highly correlated ($ICC = 0.90$ for actual versus predicted change).

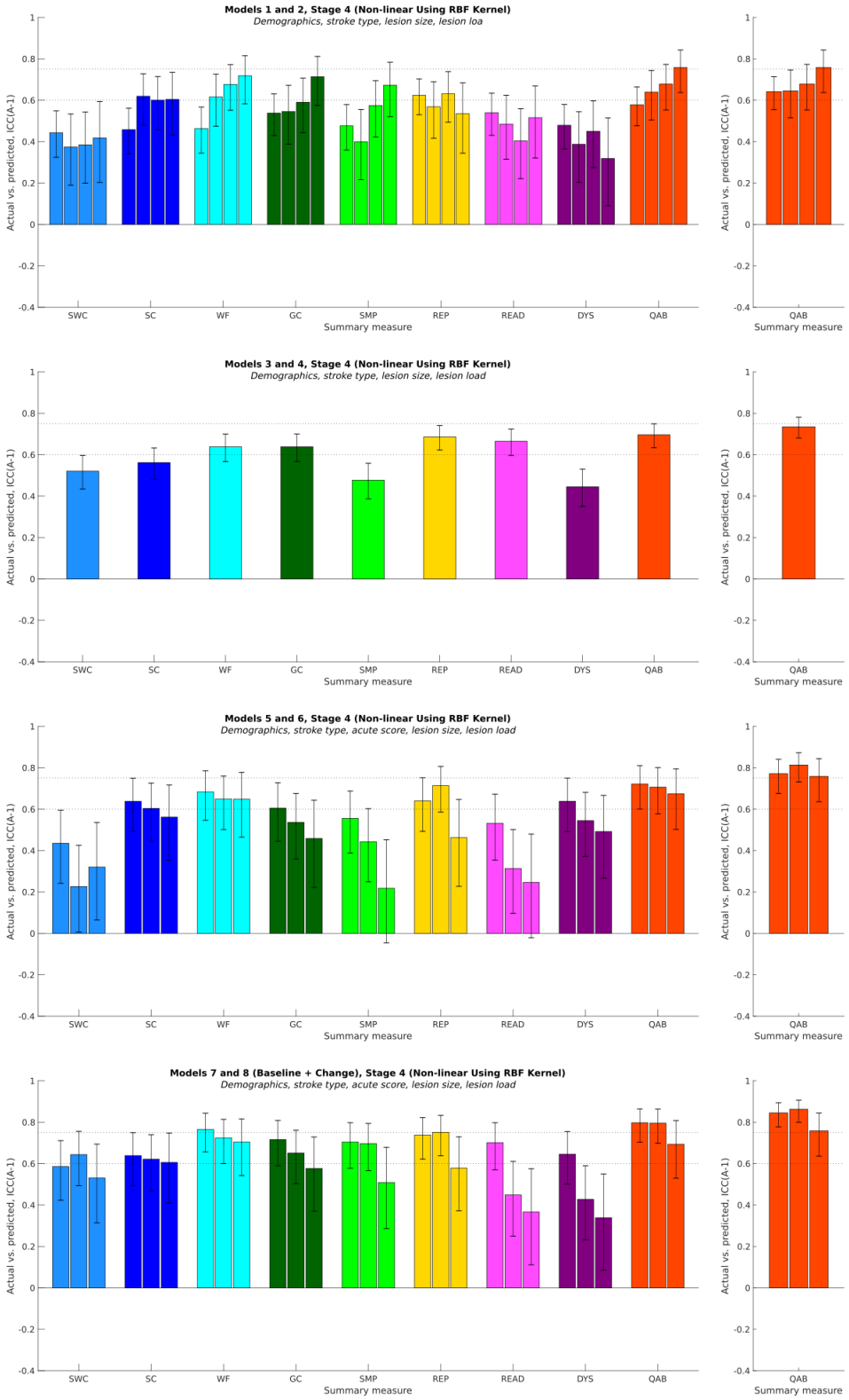

Figure 3.19: Model performance for final-stage non-linear RBF models. Error bars show 95% CIs (parametric). SWC=single word comprehension, SC=sentence comprehension, WF=word finding, GC=grammatical construction, SMP=speech-motor programming, REP=repetition, READ=reading, DYS=dysarthria, QAB=overall.

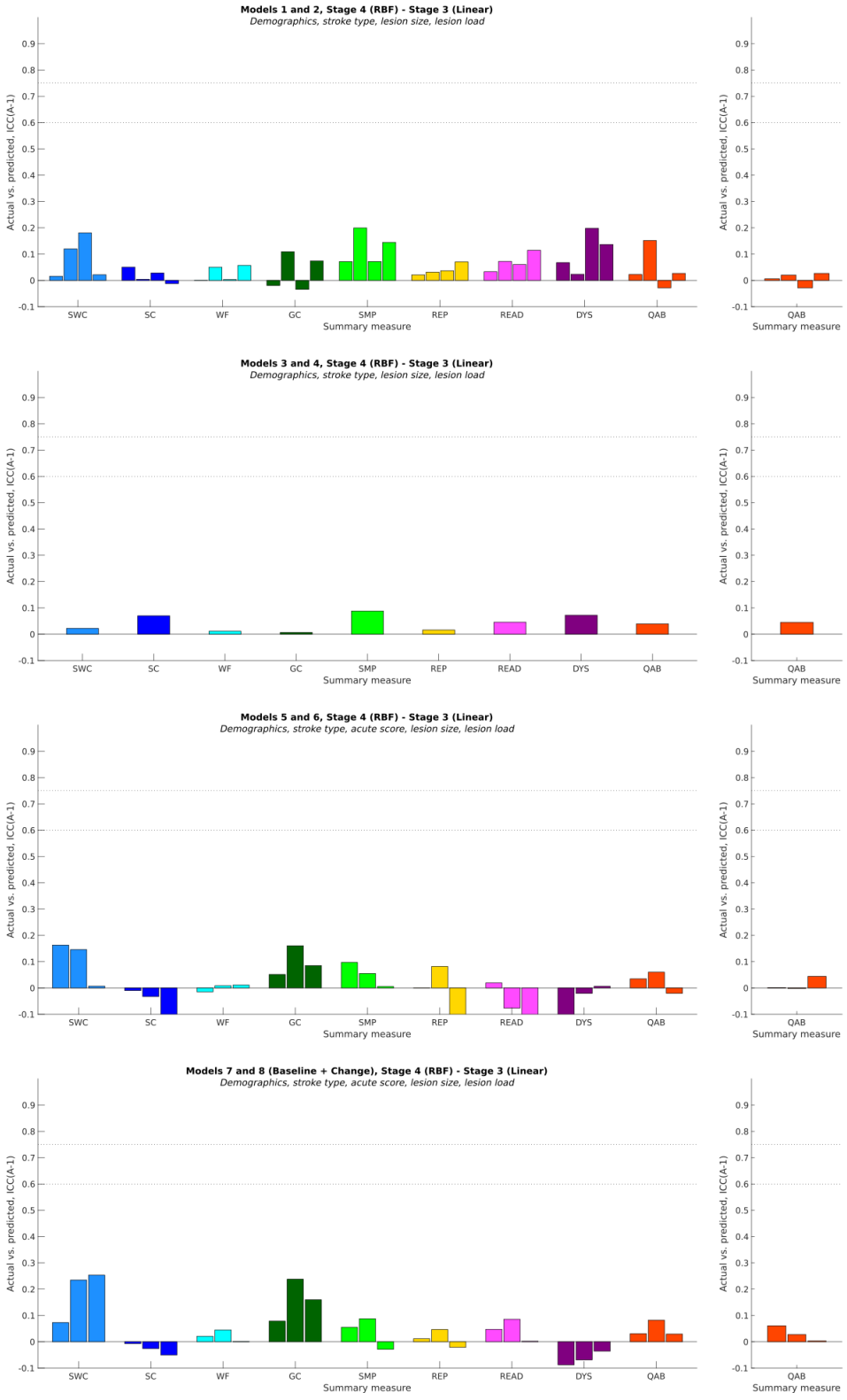

Figure 3.20: Differences in ICC between non-linear and linear models across domains (color) and times (location in group: left=earliest). No consistent benefits are readily apparent. SWC=single word comprehension, SC=sentence comprehension, WF=word finding, GC=grammatical construction, SMP=speech-motor, REP=repetition, READ=reading, DYS=dysarthria, QAB=overall.

CHAPTER 4

Discussion

The current study reveals that many aspects of language recovery in the first year following stroke can be predicted with good to excellent accuracy (ICCs up to 0.83 in linear models and 0.86 in RBF-transformed models, both for QAB overall at three months post-stroke) using SVR-based models that take demographic, language, and lesion-related variables as input. These values exceed many attained in the existing literature (e.g. Zhang et al., 2014; Del Gaizo et al., 2017; DeMarco and Turkeltaub, 2018; Kristinsson et al., 2021) despite being calculated in a more conservative manner in many cases (that is, using ICC(A,1) rather than Pearson correlation). Furthermore, our study demonstrates that information about the *location* of a lesion, beyond simply its size, is in many cases crucial for making these predictions, as made clear by the large number of significant increases in ICC observed with the addition of lesion load vectors into the models. This remains true even in cases when acute measures of language are included as predictors, particularly at later time points post-stroke. Finally, this study demonstrates differences in both predictive ability and the utility of different neural regions as predictors across different language domains, suggesting sub-specialization of particular regions for particular functions of language.

4.1 Language recovery is decelerating but continuous across most language domains

Figure 3.3 demonstrates a decelerating trajectory of recovery across the majority of language domains, in line with prior work (Kertesz and McCabe, 1977; Pedersen et al., 1995; Laska et al., 2001; Wilson, 2019). In all cases, the greatest gains appear to be made in the first month post-stroke, with slowing increases in function between one month and three months and three months and one year. Two notable exceptions to this rule in our data are speech-motor programming, which stays approximately stable after the one month time point, and word finding, which continues to show gains in function even in the three month to one year interval.

With regard to speech-motor programming, there appears to be very little research to date in-

vestigating the nature of recovery from specifically apraxia of speech (AOS) (Haley et al., 2016), perhaps due to the inherent difficulty in distinguishing apraxia of speech from aphasia (Basilakos et al., 2015). However, the trajectory of recovery for speech-motor programming observed here suggests that apraxia of speech may recover in a slightly different manner than aphasia does, perhaps reaching a plateau at an earlier time point. This would seem to be at odds with one of the only studies to longitudinally follow individuals with apraxia of speech, which showed similar patterns of recovery for both aphasia and apraxia (Hybbinette et al., 2021). Future work should examine whether the AOS-specific plateau observed here holds up in other longitudinal data sets. In general, scores related to motor speech in this data set are higher than scores on language domains, as demonstrated by the relatively high estimates for both speech-motor programming and dysarthria observed across time points. This may be an effect of scale; motor speech summary scores take on a more limited set of values than do language summary scores as calculated in the QAB (Wilson et al., 2018b), and thus may not be directly comparable to the language summary scores. However, prior work has suggested that while dysarthria may be more common than aphasia acutely (Ali et al., 2015; Mitchell et al., 2020), dysarthria may be more likely to have resolved completely by three months post-stroke than aphasia (Ali et al., 2015). The high scores reflecting lack of dysarthria at the one year time point and gradually decreasing correlation between measures of dysarthria and other language subscores observed in our data appear to support these findings.

With regard to word finding, the consistently steady gains observed after the three month time point are interesting, and could arise from a variety of causes. One theory might suggest that, because word finding is one of the most commonly impaired functions acutely following stroke (Conroy et al., 2018), the striking continued recovery may simply be due to the fact that there is more room for improvement available after the acute stage. However, compare the trajectory of word finding recovery to that of sentence comprehension, which starts at approximately the same severity acutely but does not exhibit the same steep recovery slope after three months; this suggests that a more severe initial presentation is not sufficient on its own to engender a markedly different course of language recovery. Another hypothesis is that, as speech-language therapy generally

focuses disproportionately on naming skills compared to other domains of language (Conroy et al., 2009), speech-language therapy is the cause of continuously increasing word finding abilities; however, as it is difficult to obtain reliable data about the nature and extent of speech-language therapy following stroke in a large cohort (Berthier, 2005; Xing et al., 2016; Price et al., 2017), support for this claim is lacking in this data set. It may be the case that, even independent of speech-language therapy, naming deficits are those that benefit the most from compensatory strategies developed over time (e.g. circumlocution, self-correction, pausing rather than moving forward with an incorrectly selected phoneme, etc.), and that it is the result of the formulation of those strategies that is observed here; indeed, the means by which word-finding is scored and calculated in the QAB may capture such effects, as delays are scored more leniently than paraphasias, for example. As with the motor speech findings, it will be interesting to see whether this effect replicates in future work, or when using other measures of language besides the QAB.

Another feature of interest in these findings is the striking difference in single word comprehension and sentence comprehension across time points. Though they both show similar contours of recovery, single word comprehension scores greatly exceed sentence comprehension scores at all time points, perhaps demonstrating a dissociation in the initial vulnerability of these two systems (that is, single word comprehension appears to be significantly less vulnerable to injury than does sentence comprehension). Sentence comprehension similarly shows a greater impairment overall than does grammatical construction, in line with conceptualizations of the language network that suggest distinct mechanisms to support comprehension and production of complex syntax (Matchin and Hickok, 2020). Note, as well, that correlations between both sentence comprehension and single word comprehension, and sentence comprehension and grammatical construction, are relatively low across all time points (see Fig. 3.4).

4.2 SVR can predict some language outcomes with excellent accuracy as measured by ICC

Good to excellent accuracy was attained given particular predictors for certain language domains and time points post-stroke.

Across models and time points, the most reliably well-predicted language domains were sentence comprehension, word finding, and QAB overall, while the least reliably well-predicted were single word comprehension, speech-motor programming, reading, and dysarthria.

The fact that QAB overall would be among the easiest subscores to predict is perhaps unsurprising, as what it captures is the overall severity of aphasia, a general measure likely to be influenced by the integrity of a variety of patient characteristics and neural regions simultaneously. Thus, as is indeed observed in Figures 3.10 and 3.13 (hot colors), SVR models predicting QAB overall seem to heavily utilize information about damage to left peri-Sylvian language regions, such as the inferior frontal gyrus, precentral gyrus, anterior temporal lobe, and angular and supramarginal gyri. Similarly, word finding correlates highly with QAB overall across time points (see Fig. 3.4), and predictions of word finding scores appear to be driven by a pared down but similar network of regions (see 3.10 and 3.13, second row). This suggests that measures of naming or word finding may be appropriate proxies for aphasia severity when time to conduct a full evaluation is limited (Wallace et al., 2014; Evans et al., 2020; Fridriksson, 2020). The ability to predict sentence comprehension, and perhaps particularly the extent to which lesion load information *aided* in predicting sentence comprehension even when acute scores were included in the model (most notable at the one year time point; see 3.11), was perhaps more interesting, as the comprehension of complex syntax is arguably a more distinct sub-function of language. Of note, the posterior superior temporal sulcus appeared to be a significant driver of sentence comprehension predictions across models (see Figs. 3.10 and 3.13, first row), in line with prior work that highlights the importance of this region for specifically syntactic comprehension (Pallier et al., 2011; Wilson et al., 2018a; Matchin et al., 2020; Matchin and Hickok, 2020).

It is interesting and perhaps unsurprising that recovery of single word comprehension and articulation are both more difficult to predict, at least when prediction accuracy is measured using ICC, given that these skills often recover relatively well (Selnes et al., 1984; Rogalsky et al., 2008; Ali et al., 2015; Wilson et al., 2018c), though this rule is certainly not without exceptions. It may be that single word comprehension is more bilateral in its underpinnings in healthy language function,

such that it is less vulnerable to injury via one-sided stroke; this notion would be in line with prior findings that damage to both auditory cortices is necessary to cause pure word deafness (Poeppel, 2001), that the bilateral superior temporal gyrus plays a large role in comprehending speech at the phoneme and word level (e.g. Mesgarani et al., 2014; Leonard et al., 2016), and that the anterior temporal lobes bilaterally are implicated in the word-level comprehension impairments of semantic dementia (though some aspects of this process appear to be more-left lateralized; Mesulam et al., 2013). Indeed, the current dominant model of language, Hickok and Poeppel (2007), purports that auditory word comprehension is bilateral in healthy language function (though cultural dominance is not necessarily a good index of scientific accuracy). Alternatively, a compensatory upregulation of the right hemisphere following injury may aid in the comprehension of single words, though recent findings suggest that right hemisphere reorganization may be a less prevalent mechanism of language recovery in stroke than is commonly believed (Wilson and Schneck, 2021). Dysarthria caused by upper motor neuron damage is often transient, and it has been theorized that this is due to the bilateral innervation to most cranial nerve nuclei (Enderby, 2013). However, it is important to note that some patients do continue to exhibit persistent deficits on both of these domains following left-hemisphere stroke (Palmer et al., 2007; Knollman-Porter et al., 2018), suggesting that, regardless of any potential underlying mechanism for its involvement, the isolated right hemisphere cannot always handle word comprehension or motor speech alone.

The lack of ability to predict recovery of reading from left hemisphere damage is somewhat surprising, as the neural bases of reading are somewhat well-established and appear to be primarily left-lateralized (Seghier and Price, 2011). Assessments of reading in the QAB are somewhat sparse compared to other language assessments, and writing is assessed only partially in this data set (not analyzed here); future work may need to probe more deeply into reading and writing abilities if predictions of recovery in these domains are to be accurately made.

4.3 Information about lesion location significantly improves predictions

The addition of lesion load vectors had a significant positive impact on predictive accuracy in some key scenarios. All Model 1 models except for one (reading at three months) were significantly improved by the presence of lesion load information, above and beyond what was predicted by lesion size (see Fig. 3.7). In line with prior work, handedness and gender did not appear to have any influence on long-term aphasia recovery, while age and years of education appeared to show relationships with outcomes in some cases (Watila and Balarabe, 2015; Gerstenecker and Lazar, 2019). Of particular note, speech-motor programming and dysarthria were not predictable at all from lesion size and demographic information alone, with predictive accuracy greatly increased by the addition of lesion load information. This suggests that speech-motor deficits in particular may result from focal, anatomically specific damage, rather than gross disruption to the language network (Enderby, 2013; Basilakos et al., 2015). Even when acute scores (which are able to predict outcomes quite well at one month without lesion information) were included in the models, the addition of lesion load information had a significant positive impact on predictions at the one year time point (see Figs. 3.11, 3.16). This suggests that the integrity of specific anatomical regions may be even more crucial for predicting recovery in the long-term than initial presentation, commonly believed to be the best predictor of final outcome (Lazar and Antoniello, 2008; Gerstenecker and Lazar, 2019). It has been theorized that language recovery in the first two weeks following stroke is dependent on re-normalization processes, e.g. reperfusion and resolution of diaschisis/swelling, while at later stages it is more dependent upon neuroplastic reorganization and recruitment of other regions and networks (Marsh and Hillis, 2006). The predictive power afforded by lesion load information at later stages of recovery may support this theory, as the extent to which remaining regions retain their integrity is likely to dictate their potential for recruitment or reintegration in the chronic stages of recovery.

4.4 Correlation-based accuracy on predictions of change scores should be interpreted with caution

A growing body of work has recently emerged suggesting that correlations between baseline scores and change scores (that is, some outcome score minus that baseline score) are statistically inflated due to mathematical coupling, or the correlation "of a variable with an expression containing that same variable" (Bowman et al., 2021, p. 1916). Due to ceiling effects, the general tendency of function following stroke to improve, and decreased variance in change scores compared to baselines, baseline-change correlations will often be trivially high, regardless of the biological mechanisms underlying recovery (Hope et al., 2019; Hawe et al., 2019; Bonkhoff et al., 2020; Bowman et al., 2021). The spuriousness of this relationship is quite eloquently put by Bowman et al. (2021): "...if the variability of X is substantially larger than the variability of Y, Y-X becomes close to -X+constant...As a result, the correlation of X with Y-X degenerates, approaching the correlation of X with -X+constant, which, of course, is minus one, what would be interpreted as maximum evidence for proportional recovery...To put it in the bluntest terms, if the variability of outcome scores is substantially smaller than initial scores, there really is no need to calculate the correlation between initial scores and change, we know exactly what it will be [namely, -1]...This raises the specter of tautology—in other words, one cannot help but find evidence for proportional recovery but that evidence is very often spurious" (p. 1916).

These statistical issues hold for correlations between actual and predicted change scores (Hope et al., 2019), as evidenced by the at times startlingly high prediction accuracies obtained in the original Models 7 and 8 (wherein acute scores were included as predictors to predict language *change*), particularly for those subscores and time points on which ceiling effects were most pronounced (sentence comprehension, grammatical construction, and dysarthria at the three and twelve month time points). Additionally, the disproportionate weighting of acute scores in these models may provide further reason to be suspicious of their generalizability; while cross-sectional models suggest that acute score has *decreasing* influence on prediction accuracies with increasing time (see Fig. 3.11), the change-specific models show the opposite pattern (see Fig. 3.14), with acute scores

being weighted by as much as six times more than other predictor variables when change is the response variable. It seems also worthwhile to note that the seminal paper on proportional recovery in aphasia is based on an extremely small sample size of 21 patients (Lazar et al., 2010), all who fell within a mild-to-moderate range of severity, with follow-up studies utilizing even smaller sample sizes (e.g. 14 patients with aphasia, with 4 excluded as non-fitters; Marchi et al., 2017).

This issue, additionally, warrants a conversation about ceiling effects in general—for example, whether ICC is really the appropriate metric for assessing accuracy of a model when ceiling effects are *expected* (as is perhaps the case with single word comprehension and dysarthria in our data set). It may well be the case that a measure such as mean squared error would be more appropriate in a case where a lack of variability in the true outcome scores is anticipated. However, it remains an open question whether measures of aphasia that result in ceiling effects are, truly, good measures of function. While people who attain a score of 10 might be "normal" in their language function, often this "normal" language is still markedly below their baseline. A professor who had a mild left hemisphere stroke, for example, may fall within the bounds of normal language function, but be unable to return to his or her work. Thus, when considering the deleterious effects ceiling effects can have on the interpretability of model accuracy in the presence of ceiling effects (at least when calculated using ICC), it is important to consider the origin of those ceiling effects in the first place, and whether that is a separate issue that needs to be addressed (see Hope et al., 2019; Bowman et al., 2021).

4.5 RBF-based SVR models do not appear to offer clear benefits over linear SVR models

Many studies to date using SVR methods have utilized kernel transformations in their analyses (e.g. Zhang et al., 2014; DeMarco and Turkeltaub, 2018; Hope et al., 2018; Kristinsson et al., 2021). These methods prevent clear interpretation of beta weights, require a number of arbitrary choices with regard to parameter selection, have a higher risk of overfitting, and rarely appear to have sound theoretical bases in terms of the kernel transform selected beyond it being "widely used" (Zhang et al., 2014). In the data reported here, some (but not all) models appeared to perform slightly

better under the use of the kernel trick with the field-specific recommended parameters (though this difference was not assessed statistically; see Fig. 3.20); however, it remains an open question as to whether the trade-off in interpretability when a non-linear kernel is introduced is worth the small increases in predictive accuracy it might afford. A linear SVR model assumes that features are additive, such that each increase in each feature value independently contributes some fixed amount (defined by the calculated beta weight) to the predicted response variable. While there are certainly good reasons to question whether the relationship between predictors and behavior is really that straightforward (one can, for example, easily imagine an interaction wherein pre-morbidly left versus right handed individuals with identical damage to the left hand-motor region might have different outcomes in their writing abilities), the optimal solution does not seem to be to pick an arbitrarily more complex relationship between all variables with no accompanying theoretical basis, particularly one in which the feature-prediction relationships cannot be easily recovered (Zhang et al., 2014). Rather than simply reporting the models with the highest accuracy, regardless of their complexity, results of all models, including the simplest and thereby most interpretable, are reported here.

4.6 Implications for treatment and ethical considerations

The ability to effectively predict outcomes for individuals with aphasia could have an extremely positive impact on clinical practice and living with aphasia.

First, a better baseline understanding of recovery from aphasia lays the groundwork for assessing the efficacy of treatment in clinical practice and/or clinical trials. Knowing what might be expected for a given patient's recovery at a given time based on lesion, language, and clinical characteristics alone could help to elucidate which speech and language treatments really do lead to better outcomes than would be expected naturally, essentially setting a threshold of recovery to exceed. Importantly, using this type of multidimensional and prediction-focused model, this threshold could be patient-specific, in line with recent work such as the Predicting Outcomes of Language Rehabilitation (POLAR) trial demonstrating that particular treatments may better ben-

efit particular individuals (Fridriksson and Hillis, 2021). Such a model could, in a similar manner, set goalposts for investigations into pharmacological and stimulation-based aids to behavioral treatment (e.g. Hillis, 2007; Crinion, 2016; Fridriksson et al., 2019). Additionally, due to the longitudinal aspects of the model described here, questions about the optimal timing of treatment (Holland and Fridriksson, 2001; Teasell et al., 2005; Marsh and Hillis, 2006; Godecke et al., 2012; Ali et al., 2021; Fridriksson and Hillis, 2021) could be more thoroughly investigated by providing time point-specific baselines for comparison. While this dissertation is agnostic as to what treatment approaches and timings are most effective, gaining clearer expectations for baseline recovery at particular time points will help to better define the bar for success of an intervention.

Second, the ability to provide a patient with a sense of what recovery is likely to look like *for them*, specifically, would help to set realistic expectations for the patient, their loved ones, and their clinical team alike, such that appropriate strategies for managing impairment and collaborative goal setting could be put into place (Haley et al., 2019). While this dissertation has taken a largely impairment-based perspective for the purposes of scientific clarity, a more social model of aphasia management could actually be aided by this work: providing individuals in the lives of people with aphasia (e.g. loved ones, clinicians, regular contacts) with a greater understanding of what are likely to be the person with aphasia's areas of strength could help those contacts to appropriately adjust their *own* behavior to better meet their loved one where they are. To quote Byng and Duchan (2005) in their paper on the social model of disability in aphasia, "...if other people behaved differently and if environments were changed, then many of the challenges associated with an impairment would be considerably reduced" (p. 907). Patient-specific models of expected relative strengths could help inform individuals *around* the patient of what is likely to be easiest or hardest for them, such that those proximal individuals could learn to better create conversational and environmental contexts to reveal competence in the person with aphasia.

A final thought is that, while students of speech-language pathology tend to recognize the importance of neuroanatomical awareness in clinical practice (Martin et al., 2014; Barros et al., 2017), neuroanatomical information is often poorly retained (Barros et al., 2018). Many student clinicians

exhibit "neurophobia" (Javaid et al., 2018) based on the perceived difficulty of neuroscience-related content compared to other coursework, which may prevent them from engaging deeply with neurological information which could be illuminating when available in clinical practice. Thus, the creation of an automatic tool by which to "interpret" neuroimaging data, such as the one described here, might help make the ability to make neuroanatomically informed predictions for patients more accessible to clinicians across the spectrum of care.

However, despite all of these potential benefits, it is crucial to note that statistical models are just that—models—and can only be as good as the data that goes into them and the validity of the methods that they employ. While great care has been taken to collect a high-quality patient sample and ensure the integrity of the data and analyses presented herein, machine learning approaches should always be considered as a *supplement* to, not a replacement for, clinical expertise. As machine learning models become more and more common in clinical practice, blind trust in algorithms that purport to, for example, predict when a given patient will cease to benefit from therapy could have detrimental effects on patients' quality of care or recovery (Challen et al., 2019; Thomas, 2020). Indeed, important questions have been raised as to precisely *how* neuroimaging-based models for predicting recovery from aphasia will benefit clinical practice, given common concerns about a lack of regard for individual differences, poor validation on independent data sets, inaccessibility of scanner environments for certain patients, and inattention to predictors that do not relate directly to the academic hypotheses in question (Shuster, 2018). It is the belief of the author that many of these concerns are addressed in this dissertation (e.g., individual differences are accounted for via the positioning of patients in a multidimensional symptom space; leave-one-out cross-validation at least partially handles a risk of overfitting; patients who were not MRI-safe are included via drawing lesions on CTs; demographic and non-lesion based predictors are already included, with even more predictors planned for inclusion in the future); nevertheless, this work should simply be considered an early step towards a better understanding of the myriad factors that can influence language recovery, in tandem with individual patient characteristics, therapeutic intervention, changes in neural function, and stochastic processes beyond our current

understanding—a tool, rather than a solution. It is encouraging that similar work is currently being investigated for a directly applied purpose: to determine who is most likely to benefit from which aphasia treatments (Fridriksson, 2018; Spell et al., 2020; Fridriksson and Hillis, 2021).

4.7 Limitations and future directions

This study has several notable limitations.

The participant sample herein was limited to individuals with primarily left hemisphere damage, with right hemisphere regions excluded entirely from analysis. While this decision is justified in order to avoid the curse of dimensionality (that is, having significantly more predictors than observations, particularly when those predictors are sparse in nature), it precludes the possibility of better understanding any structural contributions of the right hemisphere to language recovery and function. Perhaps similarly, the treatment of white matter regions in this analysis was somewhat crude compared to the treatment of gray matter regions, with 123 gray matter regions considered compared to only 21 white matter regions. Future work could make use of the DTI available for the majority of patients in this data set to trace white matter tracts within individuals, rather than relying on average white matter atlases.

In terms of statistical validity, the reporting of beta weights to ascribe importance to particular predictors was quite experimental, without accompanying metrics of significance. At this time, there do not appear to be commonly agreed upon guidelines for assessing the statistical significance of SVR-based beta weights, and the extent to which doing so is valid, at least in the case of functional neuroimaging data, is debated (Haufe et al., 2014); thus, it is important to note that these reported values should be interpreted with caution. Similarly, the statistical significance of differences between linear and non-linear models was not assessed due to a lack of a clear path forward for conducting such an analysis. Though there do not, on the surface, appear to be clear benefits of using non-linear over linear SVR models in the analyses herein, the jury appears to still be out on the matter, with some work suggesting that nonlinear SVM models provide no benefit over linear ones (Misaki et al., 2010), and other work suggesting non-linear models are

superior (Hope et al., 2018). The question of whether linear or non-linear models perform better for predictive modeling using neuroimaging data thus remains an empirical question to be examined by future studies.

Recent work has introduced the possibility of longitudinal SVR—that is, SVR analyses that can account for the fact that the same patients are evaluated multiple times longitudinally, and thereby exploit within-subject dependence (Chen and Bowman, 2011; Du et al., 2015). However, as analytical tools for this purpose are not yet available in standard software packages, longitudinal SVR was not attempted here; such an approach may be attempted as methods for longitudinal SVR become more accessible and user-friendly.

Finally, leave-one-out cross-validation was used in this analysis in order to maximally utilize the available data while maintaining some metric of generalizability. However, concerns have been raised about high variability of predictions using leave-one-out cross-validation compared to less noisy k-fold procedures (Poldrack et al., 2020), with some authors claiming leave-one-out cross-validation can still result in overfitting (Halai et al., 2020). Additionally, although the training and validation data used in our cross-validation procedure were fully independent, we were not able to hold out a true independent test set to evaluate final model performance without sacrificing our powerful sample size. As data from future patients is collected, this new data will become the test set upon which the true generalizability of our models can be assessed. Of course, it is our hope that these models will be equally effective for patient data acquired outside of Vanderbilt University Medical Center, which remains to be seen (Price et al., 2017; Loughnan et al., 2019).

Though creating an effective model for structurally based predictions of language recovery is exciting in and of itself, it is perhaps most exciting in its ability to serve as a baseline upon which other, more functionally oriented predictors can be assessed for their utility in predicting language recovery. Future work will examine the extent to which maps of the language network acquired using fMRI (Wilson et al., 2018c; Yen et al., 2019), along with other measures of brain health or structure such as leukoaraiosis and tractography and information about provision of speech and language therapy, can account for variance unexplained by these models. A better understanding

of patterns of reorganization following injury and their functional consequences could further clarify what is different when, structural damage being equal, recovery is more successful in some individuals than others, with the long-term aim of finding methods to induce such positive change in language recovery.

CHAPTER 5

Conclusion

This study is the first to systematically predict language outcomes for multiple pre-defined time points and on multiple language domains post-stroke, reliably doing so with good to excellent accuracy. Its findings demonstrate that information about lesion location is crucial for making the majority of these predictions, particularly at later time points post-stroke, suggesting that language recovery in the long-term is supported by mechanisms specific to particular neural regions, varying by language domain, rather than simply global processes of recovery. This work provides a valuable structural baseline upon which to build further, more functionally-oriented models of language recovery incorporating functional maps of language organization and effects of speech-language therapy in individual patients. Taken together, these scientific endeavors will help to elucidate not only *who* we expect to show successful recovery from aphasia, but also *how* and *why* that recovery might occur. This will help the field of communication sciences to better design personalized treatment plans for individual people with aphasia, more effectively manage expectations of these individuals and their loved ones, identify potential targets for stimulation-based or pharmacological treatments, and to better understand the neural bases of one of our fundamental human abilities—the capacity to communicate using language.

References

Ali, M., Lyden, P., and Brady, M. (2015). Aphasia and dysarthria in acute stroke: Recovery and functional outcome. *International Journal of Stroke*, 10(3):400–406.

Ali, M., VandenBerg, K., Williams, L. J., Williams, L. R., Abo, M., Becker, F., Bowen, A., Brandenburg, C., Breitenstein, C., Bruehl, S., Copland, D. A., Cranfill, T. B., Pietro-Bachmann, M. d., Enderby, P., Fillingham, J., Lucia Galli, F., Gandolfi, M., Glize, B., Godecke, E., Hawkins, N., Hilari, K., Hinckley, J., Horton, S., Howard, D., Jaecks, P., Jefferies, E., Jesus, L. M., Kambanaros, M., Kyoung Kang, E., Khedr, E. M., Pak-Hin Kong, A., Kukkonen, T., Laganaro, M., Lambon Ralph, M. A., Charlotte Laska, A., Leemann, B., Leff, A. P., Lima, R. R., Lorenz, A., Mac Whinney, B., Shisler Marshall, R., Mattioli, F., Mavis, I., Meinzer, M., Nilipour, R., Noé, E., Paik, N.-J., Palmer, R., Papathanasiou, I., Patricio, B. F., Pavão Martins, I., Price, C., Prizl Jakovac, T., Rochon, E., Rose, M. L., Rosso, C., Rubi-Fessen, I., Ruiter, M. B., Snell, C., Stahl, B., Szaflarski, J. P., Thomas, S. A., van de Sandt-Koenderman, M., van der Meulen, I., Visch-Brink, E., Worrall, L., Harris Wright, H., and Brady, M. C. (2021). Predictors of poststroke aphasia recovery. *Stroke*, 52(5):1778–1787.

Allen, L. M., Hasso, A. N., Handwerker, J., and Farid, H. (2012). Sequence-specific MR imaging findings that are useful in dating ischemic stroke. *RadioGraphics*, 32(5):1285–1297.

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1):95–113.

Ashburner, J. and Friston, K. J. (2000). Voxel-based morphometry–the methods. *NeuroImage*, 11(6):805–821.

Ashburner, J. and Friston, K. J. (2005). Unified segmentation. *NeuroImage*, 26(3):839–851.

Bakheit, A., Shaw, S., Carrington, S., and Griffiths, S. (2007). The rate and extent of improve-

ment with therapy from the different types of aphasia in the first year after stroke. *Clinical Rehabilitation*, 21(10):941–949.

Baliyan, V., Das, C. J., Sharma, R., and Gupta, A. K. (2016). Diffusion weighted imaging: Technique and applications. *World Journal of Radiology*, 8(9):785–798.

Barros, M. D., Silva, V. A., and Liquidato, B. M. (2017). Is anatomy important for speech language pathology (SLP) undergraduate students? *The FASEB Journal*, 31(S1):732.14–732.14.

Barros, M. D., Silva, V. A., Mendes, C. J. L., and Liquidato, B. M. (2018). Retention of anatomic knowledge in speech-language pathology undergraduate students. *The FASEB Journal*, 32(S1):508.1–508.1.

Basilakos, A., Rorden, C., Bonilha, L., Moser, D., and Fridriksson, J. (2015). Patterns of poststroke brain damage that predict speech production errors in apraxia of speech and aphasia dissociate. *Stroke*, 46(6):1561–1566.

Basso, A., Lecours, A. R., Moraschini, S., and Vanier, M. (1985). Anatomoclinical correlations of the aphasias as defined through computerized tomography: Exceptions. *Brain and Language*, 26(2):201–229.

Bates, E., Wilson, S. M., Saygin, A. P., Dick, F., Sereno, M. I., Knight, R. T., and Dronkers, N. F. (2003). Voxel-based lesion-symptom mapping. *Nature Neuroscience*, 6(5):448–450.

Benjamin, E. J., Blaha, M. J., Chiuve, S. E., Cushman, M., Das, S. R., Deo, R., de Ferranti, S. D., Floyd, J., Fornage, M., Gillespie, C., Isasi, C. R., Jiménez, M. C., Jordan, L. C., Judd, S. E., Lackland, D., Lichtman, J. H., Lisabeth, L., Liu, S., Longenecker, C. T., Mackey, R. H., Matsushita, K., Mozaffarian, D., Mussolino, M. E., Nasir, K., Neumar, R. W., Palaniappan, L., Pandey, D. K., Thiagarajan, R. R., Reeves, M. J., Ritchey, M., Rodriguez, C. J., Roth, G. A., Rosamond, W. D., Sasson, C., Towfighi, A., Tsao, C. W., Turner, M. B., Virani, S. S., Voeks, J. H., Willey, J. Z., Wilkins, J. T., Wu, J. H., Alger, H. M., Wong, S. S., Muntner, P., and

American Heart Association Statistics Committee and Stroke Statistics Subcommittee (2017). Heart disease and stroke statistics 2017 update: A report from the American Heart Association. *Circulation*, 135(10):e146–e603.

Berker, E. A., Berker, A. H., and Smith, A. (1986). Translation of Broca's 1865 report: Localization of speech in the third left frontal convolution. *Archives of Neurology*, 43(10):1065–1072.

Berthier, M. L. (2001). Unexpected brain–language relationships in aphasia: Evidence from transcortical sensory aphasia associated with frontal lobe lesions. *Aphasiology*, 15(2):99–130.

Berthier, M. L. (2005). Poststroke aphasia: Epidemiology, pathophysiology and treatment. *Drugs & Aging*, 22(2):163–182.

Bhogal, S. K., Teasell, R., and Speechley, M. (2003). Intensity of aphasia therapy, impact on recovery. *Stroke*, 34(4):987–993.

Bonkhoff, A. K., Hope, T., Bzdok, D., Guggisberg, A. G., Hawe, R. L., Dukelow, S. P., Rehme, A. K., Fink, G. R., Grefkes, C., and Bowman, H. (2020). Bringing proportional recovery into proportion: Bayesian modelling of post-stroke motor impairment. *Brain*, 143(7):2189–2206.

Bowman, H., Bonkhoff, A., Hope, T., Grefkes, C., and Price, C. (2021). Inflated estimates of proportional recovery from stroke. *Stroke*, 52(5):1915–1920.

Breitenstein, C., Grewe, T., Flöel, A., Ziegler, W., Springer, L., Martus, P., Huber, W., Willmes, K., Ringelstein, E. B., Haeusler, K. G., Abel, S., Glindemann, R., Domahs, F., Regenbrecht, F., Schlenck, K.-J., Thomas, M., Obrig, H., de Langen, E., Rocker, R., Wigbers, F., Rühmkorf, C., Hempen, I., List, J., Baumgaertner, A., and FCET2EC study group (2017). Intensive speech and language therapy in patients with chronic aphasia after stroke: A randomised, open-label, blinded-endpoint, controlled trial in a health-care setting. *Lancet*, 389(10078):1528–1538.

Bright, F. A. S., Kayes, N. M., McCann, C. M., and McPherson, K. M. (2013). Hope in people with aphasia. *Aphasiology*, 27(1):41–58.

Byng, S. and Duchan, J. F. (2005). Social model philosophies and principles: Their applications to therapies for aphasia. *Aphasiology*, 19(10/11):906–922.

Caplan, D., Hildebrandt, N., and Makris, N. (1996). Location of lesions in stroke patients with deficits in syntactic processing in sentence comprehension. *Brain*, 119(3):933–949.

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., and Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, 28(3):231–237.

Chen, S. and Bowman, F. (2011). A novel support vector classifier for longitudinal high dimensional data and its application to neuroimaging data. *Statistical Analysis and Data Mining*, 4:604–611.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4):284–290.

Connor, L. T., Obler, L. K., Tocco, M., Fitzpatrick, P. M., and Albert, M. L. (2001). Effect of socioeconomic status on aphasia severity and recovery. *Brain and Language*, 78(2):254–257.

Conroy, P., Sage, K., and Ralph, M. L. (2009). Improved vocabulary production after naming therapy in aphasia: Can gains in picture naming generalize to connected speech? *International Journal of Language & Communication Disorders*, 44(6):1036–1062.

Conroy, P., Sotiropoulou Drosopoulou, C., Humphreys, G. F., Halai, A. D., and Lambon Ralph, M. A. (2018). Time for a quick word? The striking benefits of training speed and accuracy of word retrieval in post-stroke aphasia. *Brain*, 141(6):1815–1827.

Corbetta, M., Ramsey, L., Callejas, A., Baldassarre, A., Hacker, C. D., Siegel, J. S., Astafiev, S. V., Rengachary, J., Zinn, K., Lang, C. E., Connor, L. T., Fucetola, R., Strube, M., Carter, A. R., and Shulman, G. L. (2015). Common behavioral clusters and subcortical anatomy in stroke. *Neuron*, 85(5):927–941.

Cox, R. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, 29(3).

Crary, M. A. and Gonzalez Rothi, L. J. (1989). Predicting the Western Aphasia Battery Aphasia Quotient. *The Journal of Speech and Hearing Disorders*, 54(2):163–166.

Crinion, J., Holland, A., Copland, D., Thompson, C., and Hillis, A. (2013). Neuroimaging in aphasia treatment research: Quantifying brain lesions after stroke. *NeuroImage*, 73:208–214.

Crinion, J. T. (2016). Transcranial direct current stimulation as a novel method for enhancing aphasia treatment effects. *European Psychologist*, 21(1):65–77.

Crockford, C. and Lesser, R. (1994). Assessing functional communication in aphasia: Clinical utility and time demands of three methods. *International Journal of Language & Communication Disorders*, 29(2):165–182.

Cunnings, I. and Finlayson, I. (2015). Mixed effects modeling and longitudinal data analysis. In *Advancing Quantitative Methods in Second Language Research*. Routledge.

Dehkharghani, S. and Andre, J. (2017). Imaging approaches to stroke and neurovascular disease. *Neurosurgery*, 80(5):681–700.

Del Gaizo, J., Fridriksson, J., Yourganov, G., Hillis, A. E., Hickok, G., Misic, B., Rorden, C., and Bonilha, L. (2017). Mapping language networks using the structural and dynamic brain connectomes. *eNeuro*, 4(5).

DeMarco, A. T. and Turkeltaub, P. E. (2018). A multivariate lesion symptom mapping toolbox and examination of lesion-volume biases and correction methods in lesion-symptom mapping. *Human Brain Mapping*, 39(11):4169–4182.

Doogan, C., Dignam, J., Copland, D., and Leff, A. (2018). Aphasia recovery: When, how and who to treat? *Current Neurology and Neuroscience Reports*, 18(12):90.

Doyle, P., McNeil, M., Hula, W., and Mikolic, J. (2003). The Burden of Stroke Scale (BOSS): Validating patient-reported communication difficulty and associated psychological distress in stroke survivors. *Aphasiology*, 17(3):291–304.

Dronkers, N. F., Plaisant, O., Iba-Zizen, M. T., and Cabanis, E. A. (2007). Paul Broca's historic cases: High resolution MR imaging of the brains of Leborgne and Lelong. *Brain*, 130(5):1432–1441.

Du, W., Cheung, H., Goldberg, I., Thambisetty, M., Becker, K., and Johnson, C. (2015). A longitudinal support vector reression for prediction of ALS score. *IEEE International Conference on Bioinformatics and Biomedicine*, 2015:1586–1590.

Dunn, L. E., Schweber, A. B., Manson, D. K., Lendaris, A., Herber, C., Marshall, R. S., and Lazar, R. M. (2016). Variability in motor and language recovery during the acute stroke period. *Cerebrovascular Diseases Extra*, 6(1):12–21.

Ellis, C. and Urban, S. (2016). Age and aphasia: A review of presence, type, recovery and clinical outcomes. *Topics in Stroke Rehabilitation*, 23(6):430–439.

Elman, R. J. (2016). Aphasia centers and the life participation approach to aphasia. *Topics in Language Disorders*, 36(2):154–167.

Enderby, P. (2013). Chapter 22: Disorders of communication: Dysarthria. In Barnes, M. P. and Good, D. C., editors, *Handbook of Clinical Neurology*, volume 110 of *Neurological Rehabilitation*, pages 273–281. Elsevier.

Evans, W. S., Hula, W. D., Quique, Y., and Starns, J. J. (2020). How much time do people with aphasia need to respond during picture naming? Estimating optimal response time cutoffs using a multinomial ex-Gaussian approach. *Journal of Speech, Language, and Hearing Research*, 63(2):599–614.

Fama, M. E., Baron, C. R., Hatfield, B., and Turkeltaub, P. E. (2016). Group therapy as a social context for aphasia recovery: A pilot, observational study in an acute rehabilitation hospital. *Topics in Stroke Rehabilitation*, 23(4):276–283.

Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S., Laird, A. R., Fox, P. T., Eickhoff, S. B., Yu, C., and Jiang, T. (2016). The Human Brainnetome Atlas: A new brain atlas based on connectional architecture. *Cerebral Cortex*, 26(8):3508–3526.

Flowers, H. L., Skoretz, S. A., Silver, F. L., Rochon, E., Fang, J., Flamand-Roze, C., and Martino, R. (2016). Poststroke aphasia frequency, recovery, and outcomes: A systematic review and meta-analysis. *Archives of Physical Medicine and Rehabilitation*, 97(12):2188–2201.e8.

Forkel, S. J. and Catani, M. (2018). Lesion mapping in acute stroke aphasia and its implications for recovery. *Neuropsychologia*, 115:88–100.

Fridriksson, J. (2018). Modeling treated recovery from aphasia (Clinicaltrials.gov Identifier NCT03416738). Retrieved from https://clinicaltrials.gov/ct2/show/NCT03416738.

Fridriksson, J. (2020). In the defense of 'naming' as an outcome measure in aphasia therapy studies. https://cstar.sc.edu/in-the-defense-of-naming-as-an-outcome-measure-in-aphasia-therapy-studies/.

Fridriksson, J., Basilakos, A., Stark, B. C., Rorden, C., Elm, J., Gottfried, M., George, M. S., Sen, S., and Bonilha, L. (2019). Transcranial direct current stimulation to treat aphasia: Longitudinal analysis of a randomized controlled trial. *Brain Stimulation*, 12(1):190–191.

Fridriksson, J., den Ouden, D.-B., Hillis, A. E., Hickok, G., Rorden, C., Basilakos, A., Yourganov, G., and Bonilha, L. (2018). Anatomy of aphasia revisited. *Brain*, 141(3):848–862.

Fridriksson, J. and Hillis, A. E. (2021). Current approaches to the treatment of post-stroke aphasia. *Journal of Stroke*, 23(2):183–201.

Galletta, E. E. and Barrett, A. M. (2014). Impairment and functional interventions for aphasia: Having it all. *Current Physical Medicine and Rehabilitation Reports*, 2(2):114–120.

Gerstenecker, A. and Lazar, R. M. (2019). Language recovery following stroke. *The Clinical Neuropsychologist*, 33(5):928–947.

Geschwind, N. (1965). Disconnexion syndromes in animals and man. *Brain*, 88(2):237.

Geva, S., Baron, J.-C., Jones, P. S., Price, C. J., and Warburton, E. A. (2012). A comparison of VLSM and VBM in a cohort of patients with post-stroke aphasia. *NeuroImage : Clinical*, 1(1):37–47.

Godecke, E., Hird, K., Lalor, E. E., Rai, T., and Phillips, M. R. (2012). Very early poststroke aphasia therapy: a pilot randomized controlled efficacy trial. *International Journal of Stroke: Official Journal of the International Stroke Society*, 7(8):635–644.

González-Fernández, M., Davis, C., Molitoris, J. J., Newhart, M., Leigh, R., and Hillis, A. E. (2011). Formal education, socioeconomic status, and the severity of aphasia after stroke. *Archives of Physical Medicine and Rehabilitation*, 92(11):1809–1813.

Goodglass, H., Kaplan, E., Weintraub, S., and Barresi, B. (2001). *Boston Diagnostic Aphasia Examination*. Lippincott Williams & Wilkins, Philadephia.

Halai, A. D., Woollams, A. M., and Lambon Ralph, M. A. (2020). Investigating the effect of changing parameters when building prediction models for post-stroke aphasia. *Nature Human Behaviour*, 4(7):725–735.

Haley, K. L., Cunningham, K. T., Barry, J., and de Riesthal, M. (2019). Collaborative goals for communicative life participation in aphasia: The FOURC model. *American Journal of Speech-Language Pathology*, 28(1):1–13.

Haley, K. L., Shafer, J. N., Harmon, T. G., and Jacks, A. (2016). Recovering with acquired apraxia of speech: The first 2 years. *American Journal of Speech-Language Pathology*, 25(4S).

Hanson, W. R., Riege, W. H., Metter, E. J., and Inman, V. W. (1982). Factor-derived categories of chronic aphasia. *Brain and Language*, 15(2):369–380.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., and Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96–110.

Hawe, R. L., Scott, S. H., and Dukelow, S. P. (2019). Taking proportional out of stroke recovery. *Stroke*, 50(1):204–211.

Head, H. (1926). *Aphasia and kindred disorders of speech*. Cambridge University Press.

Heilman, K. M. (2015). *Aphasia Syndromes and Information Processing Models: A Historical Perspective*, volume 1. In *The Oxford Handbook of Aphasia and Language Disorders*, Oxford University Press.

Heit, J. J., Iv, M., and Wintermark, M. (2017). Imaging of intracranial hemorrhage. *Journal of Stroke*, 19(1):11–27.

Herbet, G., Lafargue, G., and Duffau, H. (2015). Rethinking voxel-wise lesion-deficit analysis: A new challenge for computational neuropsychology. *Cortex*, 64:413–416.

Hersh, D. (1998). Beyond the 'plateau': Discharge dilemmas in chronic aphasia. *Aphasiology*, 12(3):207–218.

Hickok, G. and Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402.

Hidalgo, B. and Goodman, M. (2013). Multivariate or multivariable regression? *American Journal of Public Health*, 103(1):39–40.

Hillis, A. E. (2007). Pharmacological, surgical, and neurovascular interventions to augment acute aphasia recovery. *American Journal of Physical Medicine & Rehabilitation*, 86(6):426–434.

Hilton, R., Leenhouts, S., Webster, J., and Morris, J. (2014). Information, support and training needs of relatives of people with aphasia: Evidence from the literature. *Aphasiology*, 28(7):797–822.

Holland, A. and Fridriksson, J. (2001). Aphasia management during the early phases of recovery following stroke. *American Journal of Speech-Language Pathology*, 10(1):19–28. Publisher: American Speech-Language-Hearing Association.

Holland, A. L., Greenhouse, J. B., Fromm, D., and Swindell, C. S. (1989). Predictors of language restitution following stroke: A multivariate analysis. *Journal of Speech, Language, and Hearing Research*, 32(2):232–238.

Hope, T. M., Leff, A. P., and Price, C. J. (2018). Predicting language outcomes after stroke: Is structural disconnection a useful predictor? *NeuroImage : Clinical*, 19:22–29.

Hope, T. M., Seghier, M. L., Leff, A. P., and Price, C. J. (2013). Predicting outcome and recovery after stroke with lesions extracted from MRI images. *NeuroImage : Clinical*, 2:424–433.

Hope, T. M. H., Friston, K., Price, C. J., Leff, A. P., Rotshtein, P., and Bowman, H. (2019). Recovery after stroke: Not so proportional after all? *Brain*, 142(1):15–22.

Hope, T. M. H., Leff, A. P., Prejawa, S., Bruce, R., Haigh, Z., Lim, L., Ramsden, S., Oberhuber, M., Ludersdorfer, P., Crinion, J., Seghier, M. L., and Price, C. J. (2017). Right hemisphere structural adaptation and changing language skills years after left hemisphere stroke. *Brain*, 140(6):1718–1728.

Horner, R. D., Swanson, J. W., Bosworth, H. B., and Matchar, D. B. (2003). Effects of race and poverty on the process and outcome of inpatient rehabilitation services among stroke patients. *Stroke*, 34(4):1027–1031.

Hosseini, M., Powell, M., Collins, J., Callahan-Flintoft, C., Jones, W., Bowman, H., and Wyble,

B. (2020). I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. *Neuroscience & Biobehavioral Reviews*, 119:456–467.

Hula, W., Donovan, N. J., Kendall, D. L., and Gonzalez-Rothi, L. J. (2010). Item response theory analysis of the Western Aphasia Battery. *Aphasiology*, 24(11):1326–1341.

Hybbinette, H., Schalling, E., Plantin, J., Nygren-Deboussard, C., Schütz, M., Östberg, P., and Lindberg, P. G. (2021). Recovery of apraxia of speech and aphasia in patients with hand motor impairment after stroke. *Frontiers in Neurology*, 12.

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

Ivanova, M. V. and Hallowell, B. (2013). A tutorial on aphasia test development in any language: Key substantive and psychometric considerations. *Aphasiology*, 27(8):891–920.

Ivanova, M. V., Herron, T. J., Dronkers, N. F., and Baldo, J. V. (2021). An empirical comparison of univariate versus multivariate methods for the analysis of brain–behavior mapping. *Human Brain Mapping*, 42(4):1070–1101.

Javaid, M. A., Chakraborty, S., Cryan, J. F., Schellekens, H., and Toulouse, A. (2018). Understanding neurophobia: Reasons behind impaired understanding and learning of neuroanatomy in cross-disciplinary healthcare students. *Anatomical Sciences Education*, 11(1):81–93.

Johnson, L., Basilakos, A., Yourganov, G., Cai, B., Bonilha, L., Rorden, C., and Fridriksson, J. (2019). Progression of aphasia severity in the chronic stages of stroke. *American Journal of Speech-Language Pathology*, 28(2):639–649.

Johnson, N., Marquine, M., Flores, I., Umlauf, A., Baum, C., Wong, A., Young, A., Manly, J., Heinemann, A., Magasi, S., and Heaton, R. (2017). Racial differences in neurocognitive outcomes post-stroke: The impact of healthcare variables. *Journal of the International Neuropsychological Society*, 23(8):640–652.

Joynt, R. J. M. D. and Benton, A. L. (1964). The memoir of Marc Dax on aphasia. *Neurology*, 14(9):851–854.

Jung, I. Y., Lim, J. Y., Kang, E. K., Sohn, H. M., and Paik, N. J. (2011). The factors associated with good responses to speech therapy combined with transcranial direct current stimulation in post-stroke aphasic patients. *Annals of Rehabilitation Medicine*, 35(4):460–469.

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

Kasdan, A., Levy, D., Herrington, D., Wilson, S., and Gordon, R. (2021). A music and arts program for individuals with aphasia. Neuromusic Conference, Virtual.

Kertesz, A. (2007). *The Western Aphasia Battery Revised*. Pearson, San Antonio, TX.

Kertesz, A., Harlock, W., and Coates, R. (1979). Computer tomographic localization, lesion size, and prognosis in aphasia and nonverbal impairment. *Brain and Language*, 8(1):34–50.

Kertesz, A., Lesk, D., and McCabe, P. (1977). Isotope localization of infarcts in aphasia. *Archives of Neurology*, 34(10):590–601.

Kertesz, A. and McCabe, P. (1977). Recovery patterns and prognosis in aphasia. *Brain*, 100(1):1–18.

Klein, A., Andersson, J., Ardekani, B. A., Ashburner, J., Avants, B., Chiang, M.-C., Christensen, G. E., Collins, D. L., Gee, J., Hellier, P., Song, J. H., Jenkinson, M., Lepage, C., Rueckert, D., Thompson, P., Vercauteren, T., Woods, R. P., Mann, J. J., and Parsey, R. V. (2009). Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage*, 46(3):786–802.

Knollman-Porter, K., Dietz, A., and Dahlem, K. (2018). Intensive auditory comprehension treatment for severe aphasia: A feasibility study. *American Journal of Speech-Language Pathology*, 27(3):936–949.

Kodumuri, N., Sebastian, R., Davis, C., Posner, J., Kim, E. H., Tippett, D. C., Wright, A., and Hillis, A. E. (2016). The association of insular stroke with lesion volume. *NeuroImage : Clinical*, 11:41–45.

Kristinsson, S., Zhang, W., Rorden, C., Newman-Norlund, R., Basilakos, A., Bonilha, L., Yourganov, G., Xiao, F., Hillis, A., and Fridriksson, J. (2021). Machine learning-based multimodal prediction of language outcomes in chronic aphasia. *Human Brain Mapping*, 42(6):1682–1698.

Kundert, R., Goldsmith, J., Veerbeek, J. M., Krakauer, J. W., and Luft, A. R. (2019). What the proportional recovery rule is (and is not): Methodological and statistical considerations. *Neurorehabilitation and Neural Repair*, 33(11):876–887.

Laska, A. C., Hellblom, A., Murray, V., Kahan, T., and Von Arbin, M. (2001). Aphasia in acute stroke and relation to outcome. *Journal of Internal Medicine*, 249(5):413–422.

Lazar, R. and Antoniello, D. (2008). Variability in recovery from aphasia. *Current Neurology and Neuroscience Reports*, 8(6):497–502.

Lazar, R., Minzer, B., Antoniello, D., Festa, J., Krakauer, J., and Marshall, R. (2010). Improvement in aphasia scores after stroke is well predicted by initial severity. *Stroke*, 41(7):1485–1488.

Leblanc, R. (2019). A Parisian spring: The debate on language localization at the Imperial Academy of Medicine, Paris, April 4–June 13, 1865. *Neurosurgical Focus*, 47(3):E3.

Leff, A. P., Schofield, T. M., Crinion, J. T., Seghier, M. L., Grogan, A., Green, D. W., and Price, C. J. (2009). The left superior temporal gyrus is a shared substrate for auditory short-term memory and speech comprehension: Evidence from 210 patients with stroke. *Brain*, 132(Pt 12):3401–3410.

Lendrem, W. and Lincoln, N. B. (1985). Spontaneous recovery of language in patients with aphasia

between 4 and 34 weeks after stroke. *Journal of Neurology, Neurosurgery, and Psychiatry*, 48(8):743–748.

Leonard, M. K., Baud, M. O., Sjerps, M. J., and Chang, E. F. (2016). Perceptual restoration of masked speech in human cortex. *Nature Communications*, 7:13619.

Levelt, W. (2013). *A history of psycholinguistics: The pre-Chomskyan era*. OUP Oxford.

Lichtheim, L. (1885). On aphasia. *Brain*, 7:433–484.

Liew, S. L., Anglin, J. M., Banks, N. W., Sondag, M., Ito, K. L., Kim, H., Chan, J., Ito, J., Jung, C., Khoshab, N., Lefebvre, S., Nakamura, W., Saldana, D., Schmiesing, A., Tran, C., Vo, D., Ard, T., Heydari, P., Kim, B., Aziz-Zadeh, L., Cramer, S. C., Liu, J., Soekadar, S., Nordvik, J. E., Westlye, L. T., Wang, J., Winstein, C., Yu, C., Ai, L., Koo, B., Craddock, R. C., Milham, M., Lakich, M., Pienta, A., and Stroud, A. (2018). A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. *Scientific Data*, 5(1):180011.

Liljequist, D., Elfving, B., and Roaldsen, K. S. (2019). Intraclass correlation: A discussion and demonstration of basic features. *PLoS One*, 14(7).

Lin, M. P. and Liebeskind, D. S. (2016). Imaging of ischemic stroke. *Continuum*, 22(5):1399–1423.

Lincoln, N. B., McGuirk, E., Mulley, G. P., Lendrem, W., Jones, A. C., and Mitchell, J. R. (1984). Effectiveness of speech therapy for aphasic stroke patients: A randomised controlled trial. *Lancet*, 323(8388):1197–1200.

Lincoln, N. B., Pickersgill, M. J., and Valentine, J. D. (1981). Is the Porch Index of Communicative Ability an equal interval scale? *International Journal of Language & Communication Disorders*, 16(3):185–191.

Loughnan, R., Lorca-Puls, D. L., Gajardo-Vidal, A., Espejo-Videla, V., Gillebert, C. R., Mantini,

D., Price, C. J., and Hope, T. M. H. (2019). Generalizing post-stroke prognoses from research data to clinical data. *NeuroImage: Clinical*, 24:102005.

Mah, Y.-H., Husain, M., Rees, G., and Nachev, P. (2014). Human brain lesion-deficit inference remapped. *Brain*, 137(Pt 9):2522–2531.

Marchi, N. A., Ptak, R., Di Pietro, M., Schnider, A., and Guggisberg, A. G. (2017). Principles of proportional recovery after stroke generalize to neglect and aphasia. *European Journal of Neurology*, 24(8):1084–1087.

Marsh, E. and Hillis, A. (2006). Recovery from aphasia following brain injury: The role of reorganization. *Progress in Brain Research*, 157:143–156.

Martin, K., Bessell, N. J., and Scholten, I. (2014). The perceived importance of anatomy and neuroanatomy in the practice of speech-language pathology. *Anatomical Sciences Education*, 7(1):28–37.

Matchin, W., Basilakos, A., Stark, B. C., den Ouden, D. B., Fridriksson, J., and Hickok, G. (2020). Agrammatism and paragrammatism: A cortical double dissociation revealed by lesion-symptom mapping. *Neurobiology of Language*, 1(2):208–225.

Matchin, W. and Hickok, G. (2020). The cortical organization of syntax. *Cerebral Cortex*, 30(3):1481–1498.

Mathworks, Inc. (2019). MATLAB (R2019a). Natick, Massachusetts: The MathWorks Inc.

Mauri, M., Elli, T., Caviglia, G., Uboldi, G., and Azzi, M. (2017). RAWGraphs: A visualisation platform to create open outputs. In *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter*, CHItaly 2017, pages 1–5, New York, NY, USA. Association for Computing Machinery.

Mayo Clinic Staff (2020). Aphasia: Symptoms and causes. https://www.mayoclinic.org/diseases-conditions/aphasia/symptoms-causes/syc-20369518?p=1.

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010.

Mesulam, M. M., Wieneke, C., Hurley, R., Rademaker, A., Thompson, C. K., Weintraub, S., and Rogalski, E. J. (2013). Words and objects at the tip of the left temporal lobe in primary progressive aphasia. *Brain*, 136:601–618.

Metter, E. J., Kempler, D., Jackson, C., Hanson, W. R., Mazziotta, J. C., and Phelps, M. E. (1989). Cerebral glucose metabolism in Wernicke's, Broca's, and conduction aphasia. *Archives of Neurology*, 46(1):27–34.

Misaki, M., Kim, Y., Bandettini, P. A., and Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage*, 53(1):103–118.

Mitchell, C., Gittins, M., Tyson, S., Vail, A., Conroy, P., Paley, L., and Bowen, A. (2020). Prevalence of aphasia and dysarthria among inpatient stroke survivors: Describing the population, therapy provision and outcomes on discharge. *Aphasiology*.

Mohr, J. (1976). Broca's area and Broca's aphasia (1976). In Grodzinsky, Y. and Amunts, K., editors, *Broca's Region*, pages 384–394.

Mori, S., Wakana, S., Zijl, P. C. V., and Nagae-Poetscher, L. M. (2005). *MRI Atlas of Human White Matter*. Elsevier.

Musser, B., Wilkinson, J., Gilbert, T., and Bokhour, B. G. (2015). Changes in identity after aphasic stroke: Implications for primary care. *International Journal of Family Medicine*, 2015.

Nachev, P. (2015). The first step in modern lesion-deficit analysis. *Brain*, 138(6):e354–e354.

Naeser, M. A. and Hayward, R. W. (1978). Lesion localization in aphasia with cranial computed tomography and the Boston Diagnostic Aphasia Exam. *Neurology*, 28(6):545–551.

Naeser, M. A. and Palumbo, C. L. (1994). Neuroimaging and language recovery in stroke. *Journal of Clinical Neurophysiology*, 11(2):150–174.

NIDCD (2015). NIDCD fact sheet on voice, speech, and language: Aphasia. https://www.nidcd.nih.gov/sites/default/files/Documents/health/voice/Aphasia.pdf.

Olsen, T. S., Bruhn, P., and Oberg, R. G. (1986). Cortical hypoperfusion as a possible cause of 'subcortical aphasia'. *Brain*, 109 (Pt 3):393–410.

Osa García, A., Brambati, S. M., Brisebois, A., Désilets-Barnabé, M., Houzé, B., Bedetti, C., Rochon, E., Leonard, C., Desautels, A., and Marcotte, K. (2020). Predicting early post-stroke aphasia outcome from initial aphasia severity. *Frontiers in Neurology*, 11.

Pallier, C., Devauchelle, A., and Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, 108(6):2522–2527.

Palmer, R., Enderby, P., and Hawley, M. (2007). Addressing the needs of speakers with long-standing dysarthria: Computerized and traditional therapy compared. *International Journal of Language & Communication Disorders*, 42 Suppl 1:61–79.

Paolucci, S., Antonucci, G., Grasso, M. G., Bragoni, M., Coiro, P., De Angelis, D., Fusco, F. R., Morelli, D., Venturiero, V., Troisi, E., and Pratesi, L. (2003). Functional outcome of ischemic and hemorrhagic stroke patients after inpatient rehabilitation: A matched comparison. *Stroke*, 34(12):2861–2865.

Patterson, J. (2015). *Aphasia Assessment*, volume 1. In *The Oxford Handbook of Aphasia and Language Disorders*, Oxford University Press.

Pedersen, P., Jørgensen, H., Nakayama, H., Raaschou, H., and Olsen, T. (1995). Aphasia in acute stroke: Incidence, determinants, and recovery. *Annals of neurology*, 38:659–66.

Pedersen, P. M., Vinter, K., and Olsen, T. S. (2004). Aphasia after stroke: Type, severity and prognosis. *Cerebrovascular Diseases*, 17(1):35–43.

Pickersgill, M. J. and Lincoln, N. B. (1983). Prognostic indicators and the pattern of recovery of communication in aphasic stroke patients. *Journal of Neurology, Neurosurgery, and Psychiatry*, 46(2):130–139.

Pizzamiglio, L., Mammucari, A., and Razzano, C. (1985). Evidence for sex differences in brain organization in recovery in aphasia. *Brain and Language*, 25(2):213–223.

Poeppel, D. (2001). Pure word deafness and the bilateral processing of the speech code. *Cognitive Science*, 25(5):679–693.

Poldrack, R. A., Huckins, G., and Varoquaux, G. (2020). Establishment of best practices for evidence for prediction: A review. *JAMA Psychiatry*, 77(5):534–540.

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

Pustina, D., Avants, B., Faseyitan, O. K., Medaglia, J. D., and Coslett, H. B. (2018). Improved accuracy of lesion to symptom mapping with multivariate sparse canonical correlations. *Neuropsychologia*, 115:154–166.

Pustina, D., Coslett, H. B., Ungar, L., Faseyitan, O. K., Medaglia, J. D., Avants, B., and Schwartz, M. F. (2017). Enhanced estimations of post-stroke aphasia severity using stacked multimodal predictions. *Human Brain Mapping*, 38(11):5603–5615.

Rogalsky, C., Pitz, E., Hillis, A. E., and Hickok, G. (2008). Auditory word comprehension impairment in acute stroke: Relative contribution of phonemic versus semantic factors. *Brain and Language*, 107(2):167–169.

Rorden, C., Fridriksson, J., and Karnath, H. O. (2009). An evaluation of traditional and novel tools for lesion behavior mapping. *NeuroImage*, 44(4):1355–1362.

Salarian, A. (2021). Intraclass correlation coefficient (ICC). MATLAB Central File Exchange, https://www.mathworks.com/matlabcentral/fileexchange/22099-intraclass-correlation-coefficient-icc.

Salvadori, E., Papi, G., Insalata, G., Rinnoci, V., Donnini, I., Martini, M., Falsini, C., Hakiki, B., Romoli, A., Barbato, C., Polcaro, P., Casamorata, F., Macchi, C., Cecchi, F., and Poggesi, A. (2020). Comparison between ischemic and hemorrhagic strokes in functional outcome at discharge from an intensive rehabilitation hospital. *Diagnostics*, 11(1):38.

Schneck, S. M., Entrup, J. L., Duff, M. C., and Wilson, S. M. (2021). Unexpected absence of aphasia following left temporal hemorrhage: a case study with functional neuroimaging to characterize the nature of atypical language localization. *Neurocase*, 27(1):97–105.

Schwartz, M. F., Kimberg, D. Y., Walker, G. M., Faseyitan, O., Brecher, A., Dell, G. S., and Coslett, H. B. (2009). Anterior temporal involvement in semantic word retrieval: Voxel-based lesion-symptom mapping evidence from aphasia. *Brain*, 132(12):3411–3427.

Seghier, M. L. and Price, C. J. (2011). Explaining left lateralization for words in the ventral occipitotemporal cortex. *Journal of Neuroscience*, 31(41):14745–14753.

Selnes, O. A., Niccum, N., Knopman, D. S., and Rubens, A. B. (1984). Recovery of single word comprehension: CT-scan correlates. 21(1):72–84.

Shuster, L. (2018). Considerations for the use of neuroimaging technologies for predicting recovery of speech and language in aphasia. *American Journal of Speech-Language Pathology*, 27(1S):291–305.

Siegel, J. S., Ramsey, L. E., Snyder, A. Z., Metcalf, N. V., Chacko, R. V., Weinberger, K., Baldassarre, A., Hacker, C. D., Shulman, G. L., and Corbetta, M. (2016). Disruptions of network connectivity predict impairment in multiple behavioral domains after stroke. *Proceedings of the National Academy of Sciences*, 113(30):E4367–E4376.

Spaccavento, S., Craca, A., Del Prete, M., Falcone, R., Colucci, A., Di Palma, A., and Loverre, A. (2013). Quality of life measurement and outcome in aphasia. *Neuropsychiatric Disease and Treatment*, 10:27–37.

Spell, L. A., Richardson, J. D., Basilakos, A., Stark, B. C., Teklehaimanot, A., Hillis, A. E., and Fridriksson, J. (2020). Developing, implementing, and improving assessment and treatment fidelity in clinical aphasia research. *American Journal of Speech-Language Pathology*, 29(1):286–298.

Sperber, C., Wiesen, D., and Karnath, H. O. (2019). An empirical evaluation of multivariate lesion behaviour mapping using support vector regression. 40(5):1381–1390.

Spreen, O. and Risser, A. H. (2003). *Assessment of Aphasia*. Oxford University Press, USA.

Teasell, R., Bitensky, J., Salter, K., and Bayona, N. A. (2005). The role of timing and intensity of rehabilitation therapies. *Topics in Stroke Rehabilitation*, 12(3):46–57.

Tesak, J. and Code, C. (2008). *Milestones in the History of Aphasia: Theories and Protagonists*. Psychology Press.

Thomas, R. (2020). Medicine's machine learning problem. *Boston Review*. http://bostonreview.net/science-nature/rachel-thomas-medicines-machine-learning-problem.

Turkheimer, E., Yeo, R. A., Jones, C. L., and Bigler, E. D. (1990). Quantitative assessment of covariation between neuropsychological function and location of naturally occurring lesions in humans. *Journal of Clinical and Experimental Neuropsychology*, 12(4):549–565.

Umarova, R. M., Sperber, C., Kaller, C. P., Schmidt, C. S. M., Urbach, H., Klöppel, S., Weiller, C., and Karnath, H. O. (2019). Cognitive reserve impacts on disability and cognitive deficits in acute stroke. *Journal of Neurology*, 266(10):2495–2504.

Vapnik, V. N. (1998). *Statistical learning theory*. Adaptive and learning systems for signal processing, communications, and control. Wiley, New York.

Wallace, S. J., Worrall, L., Rose, T., and Dorze, G. L. (2014). Measuring outcomes in aphasia research: A review of current practice and an agenda for standardisation. *Aphasiology*, 28(11):1364–1384.

Wallentin, M. (2018). Sex differences in post-stroke aphasia rates are caused by age: A meta-analysis and database query. *PLoS One*, 13(12).

Watila, M. M. and Balarabe, S. A. (2015). Factors predicting post-stroke aphasia recovery. *Journal of the Neurological Sciences*, 352(1-2):12–18.

Wernicke, C. (1875). The aphasia symptom-complex: A psychological study on an anatomical basis (1875). In Eling, P., editor, *Reader in the history of aphasia: From Franz Gall to Norman Geschwind*, page 90. John Benjamins Publishing.

Wernicke, C. (1886). Some new studies on aphasia (1886). In Eling, P., editor, *Reader in the History of Aphasia: From Franz Gall to Norman Geschwind*, page 69. John Benjamins Publishing.

Wertz, R. T. and Dronkers, N. F. (1990). Effects of age on aphasia. In *Proceedings of Research Symposium on Communication Sciences and Disorders of Aging*, pages 88–98. American Speech Language Hearing Association.

Willmes, K. and Poeck, K. (1993). To what extent can aphasic syndromes be localized? *Brain*, 116(6):1527–1540.

Wilson, S. (2019). Neurolinguistic studies of patients with acquired aphasias. In Zubicaray, G. and Schiller, N. O., editors, *The Oxford Handbook of Neurolinguistics*. Oxford University Press.

Wilson, S. M. (2017). Lesion-symptom mapping in the study of spoken language understanding. *Language, Cognition and Neuroscience*, 32(7):891–899.

Wilson, S. M., Bautista, A., and McCarron, A. (2018a). Convergence of spoken and written language processing in the superior temporal sulcus. *NeuroImage*, 171:62–74.

Wilson, S. M., Eriksson, D. K., Schneck, S. M., and Lucanie, J. M. (2018b). A quick aphasia battery for efficient, reliable, and multidimensional assessment of language function. *PLoS One*, 13(2):e0192773.

Wilson, S. M., Henry, M. L., Besbris, M., Ogar, J. M., Dronkers, N. F., Jarrold, W., Miller, B. L., and Gorno-Tempini, M. L. (2010). Connected speech production in three variants of primary progressive aphasia. *Brain*, 133(7):2069–2088.

Wilson, S. M. and Hula, W. D. (2019). Multivariate approaches to understanding aphasia and its neural substrates. *Current Neurology and Neuroscience Reports*, 19(8):53.

Wilson, S. M., Lam, D., Babiak, M. C., Perry, D. W., Shih, T., Hess, C. P., Berger, M. S., and Chang, E. F. (2015). Transient aphasias after left hemisphere resective surgery. *Journal of Neurosurgery*, 123(3):581–593.

Wilson, S. M., Ogar, J. M., Laluz, V., Growdon, M., Jang, J., Glenn, S., Miller, B. L., Weiner, M. W., and Gorno-Tempini, M. L. (2009). Automated MRI-based classification of primary progressive aphasia variants. *NeuroImage*, 47(4):1558–1567.

Wilson, S. M. and Schneck, S. M. (2021). Neuroplasticity in post-stroke aphasia: A systematic review and meta-analysis of functional imaging studies of reorganization of language processing. *Neurobiology of Language*, 2(1):22–82.

Wilson, S. M., Yen, M., and Eriksson, D. K. (2018c). An adaptive semantic matching paradigm for reliable and valid language mapping in individuals with aphasia. *Human Brain Mapping*, 39(8):3285–3307.

Worrall, L., Sherratt, S., Rogers, P., Howe, T., Hersh, D., Ferguson, A., and Davidson, B. (2011). What people with aphasia want: Their goals according to the ICF. *Aphasiology*, 25(3):309–322.

Xing, C., Arai, K., Lo, E. H., and Hommel, M. (2012). Pathophysiologic cascades in ischemic stroke. *International Journal of Stroke*, 7(5):378–385.

Xing, S., Lacey, E. H., Skipper-Kallal, L. M., Jiang, X., Harris-Love, M. L., Zeng, J., and Turkeltaub, P. E. (2016). Right hemisphere grey matter structure and language outcomes in chronic left hemisphere stroke. *Brain*, 139(1):227–241.

Xu, T., Jha, A., and Nachev, P. (2018). The dimensionalities of lesion-deficit mapping. *Neuropsychologia*, 115:134–141.

Yen, M., DeMarco, A. T., and Wilson, S. M. (2019). Adaptive paradigms for mapping phonological regions in individual participants. *NeuroImage*, 189:368–379.

Yourganov, G., Fridriksson, J., Rorden, C., Gleichgerrcht, E., and Bonilha, L. (2016). Multivariate connectome-based symptom mapping in post-stroke patients: Networks supporting language and speech. *Journal of Neuroscience*, 36(25):6668–6679.

Yourganov, G., Smith, K. G., Fridriksson, J., and Rorden, C. (2015). Predicting aphasia type from brain damage measured with structural MRI. *Cortex*, 73:203–215.

Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., and Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *NeuroImage*, 31(3):1116–1128.

Zhang, Y., Kimberg, D. Y., Coslett, H. B., Schwartz, M. F., and Wang, Z. (2014). Multivariate lesion-symptom mapping using support vector regression. *Human Brain Mapping*, 35(12):5861–5876.