
Post Stroke Motor Recovery Genome Wide Association Study: A *Domain-Specific Approach*

Chad M. Aldridge^{1,*}, Robynne, Braun², Keith L. Keene^{3,4}, Fang-Chi Hsu⁵, Michele M. Sale⁶,
Bradford B. Worrall^{1,6}

¹Department of Neurology, University of Virginia, Charlottesville, VA, USA

²Department of Neurology, University of Maryland, Baltimore, MD, USA

³Department of Biology, East Carolina University, Greenville, NC, USA

⁴Center for Health Disparities, Brody School of Medicine, East Carolina University, Greenville, NC, USA

⁵Department of Biostatistics and Data Science, Division of Public Health Sciences, Wake Forest University School of Medicine, Winston-Salem, NC, USA

⁶Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA

*Corresponding Author

Chad M. Aldridge, PT, DPT, MC-CR, NCS

Department of Neurology

University of Virginia

Charlottesville, VA

cma7n@uvahealth.org

Abstract

Background: In this genome wide association study (GWAS) we aimed to discover single nucleotide polymorphisms (SNPs) associated with motor recovery post-stroke. **Methods:** We used the Vitamin Intervention for Stroke Prevention (VISP) dataset of 2,100 genotyped patients with non-disabling stroke. Of these, 488 patients had motor impairment at enrollment. Genotyped data underwent strict quality control and imputation. The GWAS utilized logistic regression models with generalized estimating equations (GEE) to leverage the repeated NIH Stroke Scale (NIHSS) motor score measurements spanning 6 time points over 24 months. The primary outcome was a decrease in the motor drift score of ≥ 1 vs. < 1 at each timepoint. Our model estimated the odds ratio of motor improvement for each SNP after adjusting for age, sex, race, days from stroke to visit, initial motor score, VISP treatment arm, and principal components.

Results: Although no associations reached genome-wide significance ($p < 5 \times 10^{-8}$), our analysis detected 115 suggestive associations ($p < 5 \times 10^{-6}$). Notably, we found multiple SNP clusters near genes with plausible neuronal repair biology mechanisms. The CLDN23 gene had the most convincing association which affects blood-brain barrier integrity, neurodevelopment, and immune cell transmigration. **Conclusion:** We identified novel suggestive genetic associations with the first ever motor-specific post stroke recovery GWAS. The results seem to describe a distinct stroke recovery phenotype compared to prior genetic stroke outcome studies that use outcome measures, like the mRS. Replication and further mechanistic investigation are warranted. Additionally, this study demonstrated a proof-of-principle approach to optimize statistical efficiency with longitudinal datasets for genetic discovery.

1 Introduction

2 A reckoning is coming to the field of stroke recovery and genomics. The research, now merging
3 at the intersection of these fields, faces three major challenges. First, a majority of the studies on
4 stroke-related genes use a candidate gene approach [1], while there are only two genome-wide
5 association study published to date [2, 3]. Current understanding of stroke recovery genetics is
6 therefore limited to an extremely small portion of the genome, encompassing only 11 associated
7 genes [1]. However the complex and time-varying biology of stroke recovery is likely to involve a
8 much greater proportion of the genome. This suggests that study designs using genome-wide [4]
9 and epigenome-wide [5] associations are well suited to discover novel recovery-associated genes
10 and their variations. The second issue is that acute stroke treatment trials often collect blood
11 samples useful for subsequent genetic studies. However, they tend to lack detailed measures of
12 stroke recovery. Conversely, stroke recovery trials frequently collect these detailed and
13 domain-specific outcome measures, but lack biospecimens for subsequent genetic analyses. The
14 third challenge entails the issue that most studies on stroke recovery-related genes have defined
15 their recovery phenotypes using global outcome measures that combine multiple domains of
16 impairment (e.g. the modified Rankin Scale or total NIH Stroke Scale score) rather than using
17 domain-specific measures (e.g. the Upper Extremity Fugl-Meyer for the motor domain) [6].
18 It remains unclear whether the phenotype-genotype associations observed using multi-domain
19 measures differ from those observed using domain-specific measures of stroke recovery. For
20 example variants of the BDNF (brain derived neurotrophic factor) gene have shown to predict
21 poor stroke outcomes defined as the 90-day modified Rankin Scale (mRS) score ≤ 1 for ischemic
22 stroke, or Glasgow Outcome Scale score ≤ 3 for hemorrhagic stroke [7]. However, Cramer and
23 colleagues [8] recently showed that BDNF variants were not associated with a domain-specific
24 measure of arm motor function. This suggests that change in a multi-domain outcome measure
25 may represent a different phenotype-genotype relationship than change in a domain-specific
26 measure. The distinction is not trivial. As noted in the Stroke Recovery and Rehabilitation
27 Roundtable [9] guidelines, “brain repair maps best onto fine-grained movement quality measures
28 that are sensitive and specific.” In other words, using domain-specific measures of stroke
29 recovery is better suited for studies that aim to discover genetic mechanisms of brain plasticity.

30 Thus, genetic studies of stroke recovery using domain-specific measures are urgently needed.
31 In an effort to address this need, Braun and colleagues [10] argued that changes in NIHSS
32 subscores, which measure impairment in distinct neurological domains, can be considered as an
33 efficient and clinically feasible means to obtain domain-specific measures of stroke recovery.
34 They noted that the NIHSS motor impairment subscores are comparable to the Fugl-Meyer in
35 terms of being specific to arm and leg motor function. They also have good inter-rater reliability
36 (κ 0.77-0.78). The present study is the first effort to define a phenotype-genotype
37 association specific to post stroke motor-recovery using the change in NIHSS subscores.

38 **2 Methods and Materials**

39 **2.1 Discovery Cohort**

40 The Vitamin Intervention for Stroke Prevention trial (VISP) investigated the effect of vitamin
41 supplementation dose on the risk of recurrent stroke with a randomized double-blinded design.
42 The study enrolled patients who had a non-disabling ischemic stroke ($mRS < 3$) \geq 72 hours prior
43 to enrollment. Patients were randomized to a high dose or low dose vitamin supplementation arm
44 if they were at least 75% compliant of taking a low dose supplementation packet for one month
45 prior. All patients were reassessed every 3 months until a recurrent stroke event, but not longer
46 than 2 years [11]. The trial successfully enrolled a total of 3680 randomized patients. This study
47 was approved by the internal review boards (IRBs) of Wake Forest University School of Medicine,
48 University of North Carolina at Chapel Hill School of Medicine as well as individual recruiting sites
49 in accordance with the declaration of Helsinki. All patients provided written informed consent [12].
50 However, ten sites did not approve the genetic portion of the study resulting in 2,100 genotyped
51 patients. We included only those that had a motor drift weakness of an arm or leg at the initial
52 measurement of the NIH stroke scale at randomization. We excluded patients that had an
53 incident recurrent stroke during the trial. This resulted in 488 participants in this GWAS.

54 **2.2 Quality Control**

55 The Center for Inherited Disease Research at Johns Hopkins University performed genotyping on
56 the Illumina HumanOmni1-Quad-v1 array(Illumina, Inc.) The genotyped data underwent strict
57 quality control measures that filtered out SNPs as follows: 1) missing call rate > 2%, 2)
58 Mendelian errors in control trios, 3) deviation from Hardy-Weinburg equilibrium in controls, 4)
59 discordant calls in duplicate samples, 5) sex differences in allele frequency or heterozygosity, 6)
60 and minor allele frequency < 0.05 in line with previously published recommendations [13]. We
61 further increased the number of SNP with genetic imputation via the TOPMed Imputation server
62 [14, 15] which implements the Minimac Imputation procedure [16]. The TOPMed study [14] has a
63 large cohort of 97,256 individuals with a diverse set of backgrounds which was preferred because
64 of the sizeable proportion of non-European ancestry participants in the VISP genotyped cohort.
65 After filtering out imputed SNPs with poor imputation quality ($r^2 < 0.80$) and MAFs < 0.05, the
66 final count of SNPs came to 6,588,085.

67 **2.3 Phenotyping**

68 As suggested [10], we utilized the motor drift subscores of the NIHSS as a measurement of motor
69 weakness. The NIHSS subscores 5A/5B and 6A/6B defined the degree of limb weakness for the
70 upper and lower extremities also known as drift. The subscores rate limb weakness or drift on an
71 ordinal scale from 0 to 5: 0 is no drift, 1 drift is present, 2 observed some effort against gravity, 3
72 shows no effort against gravity, 4 there is no movement, and 5 the limb is amputated. Motor
73 improvement is defined as the decrease in the initial motor drift subscore of the weakest limb from
74 enrollment to each follow up period. If patients had equally affected upper and lower limbs, we
75 chose the upper limb. To maximize statistical power and model stability, we chose to dichotomize
76 motor improvement as a decrease in initial motor drift by ≥ 1 versus < 1 for each follow up period.

77 **2.4 Data Analysis Plan**

78 We implemented a logistic regression model with generalized estimating equations (GEE) with
79 the "gee" R package [17]. The GEE model allows the incorporation of repeated measurements of
80 the motor drift subscore over the 2 year trial duration[18], which provides notable statistical power

81 gains compared to the traditional case/control GWAS study design.

82 A priori we planned to adjust for age, sex, initial motor drift score, treatment arm, and population
83 stratification via principle components. We calculated the top ten principle components utilizing
84 the KING software [19] to account for population stratification in our cohort with genotyped SNPs
85 after pruning. To determine which principle components to include in GWAS model, we used a
86 backwards selection procedure optimizing the AIC with the "stepAIC" function from the MASS R
87 package [20]. This approach allows for more efficient population stratification adjustment.

88 In addition to the *a priori* covariates, time since stroke onset is an important covariate when
89 modeling stroke recovery because of changing rates of recovery based on well defined time
90 epochs(i.e. Acute, Early and Late Subacute, and Chronic)[21]. These epochs are tied to
91 biological processes of inflammation and scarring early on into recovery with a transition to
92 mainly endogenous plasticity in later stages. To account for this effect in the model, we added the
93 covariate of time from stroke onset to time of motor drift measurement in days for each follow up
94 period. Furthermore, we investigated the non-linear relationship of this covariate via binning the
95 time from stroke onset to follow up period into quartiles. Figure 1 shows the mean estimated
96 probability of motor drift improvement by each quartile. We decided to use a spline of time from
97 onset to measurement with 1 knot at 250 days to better model the non-linear relationship and
98 maintain clinical interpret-ability of the model for planned sensitivity analysis.

99 Lastly, we considered possible loss to follow up effects. We planned to investigate which baseline
100 characteristics predict missing motor drift scores. Any associated baseline characteristics would
101 be added to the final GEE model with an exchangeable correlation structure as covariates.

102 **2.5 Sensitivity Analysis**

103 We performed two sensitivity analyses. First, we evaluated the interaction of time of stroke onset
104 to follow-up period spline with each SNP that reached a p-value threshold of $p < 5 \times 10^{-6}$. We
105 suspected that the effect of the SNP may change depending on the stroke recovery phase.

106 Secondly, we observed a wide spread of time from stroke onset to VISP randomization (median
107 72 days; IQR 45.75 - 102 days). We generated an early versus late post-stroke enrollment
108 variable defined as $<$ the median (72 days) being early and \geq the median as late enrollment,

109 then estimate its interaction with each SNP as a separate sensitivity analysis.

110 **2.6 Look-Up Analysis**

111 We wished to investigate if the reported SNPs from the GISCOME GWAS study [2] on stroke
112 functional recovery replicate with our post stroke motor recovery associated SNPs. The
113 GISCOME study is the largest post stroke recovery GWAS by combined sample size ($n = 6,021$)
114 from 12 studies. Söderholm et al. defined good recovery as a mRS of ≤ 2 and a mRS of ≥ 3
115 signified poor recovery. We planned to compare our GWAS results with all SNPs with a p value
116 $< 5 \times 10^{-6}$ from the GISCOME study. The p-values of the look-up analysis will receive a multiple
117 comparison adjustment at a FDR of 10%. We chose the 10% rate because the look-up analysis
118 has SNPs in linkage disequilibrium. The FDR algorithm assumes that each hypothesis test is
119 independent from one another, which is violated when applied to SNPs within linkage
120 disequilibrium. This violation biases the adjusted p-values to the null which makes a FDR of 5%
121 exceedingly conservative.

122 **3 Results**

123 **3.1 Demographics**

124 The 488 patients provided 2,095 individual observations over the entire VISIP study 2 year period
125 from enrollment to months 1, 6, 12, 18 and 24. Patients had a median (IQR) 5 (4-5) number of
126 motor drift assessments with a minimum of 1 to a maximum of 5. Twenty six point six percent of
127 patients were lost to follow up by the 24 month visit. We found that sex and self-identified race
128 were associated with loss to follow up. Males made up 76% patients that were lost to follow
129 versus 59% ($p < 0.001$). Patients that identified as Black were more likely to be lost to follow up
130 (30% vs 15%; $p < 0.001$). In contrast, self-identified White patients were less likely to be lost to
131 follow up (62% versus 77%; $p < 0.001$). Table 1 shows the demographics of this cohort. As
132 expected, most of the patients had worse arm weakness (67%) than leg weakness likely due to
133 the VISIP inclusion criteria of non-disability strokes defined by a mRS ≤ 3 . It is well known that
134 the higher mRS scores are biased toward lower extremity weakness and inability to walk

135 compared to upper extremity weakness. Of note, the distribution of patient ancestry generally
136 reflects the national U.S. population.

137 **3.2 GWAS**

138 **3.2.1 Primary Results**

139 None of the SNPs reached genome-wide significance ($p < 5 \times 10^{-8}$). However, 115 SNPs
140 reached suggestive associations with motor improvement ($p < 5 \times 10^{-6}$). Figure 2 plots the
141 p-values of the odds ratio of motor improvement for each SNP. The calculated genomic control λ
142 of the GWAS is 1.01, which suggests no genomic inflation. Therefore, we did not adjust the
143 p-values.

144 The suggestive SNPs found themselves in chromosomes 1 (3), 6 (1), 8 (92), 9 (6), 12 (6), 14 (1),
145 16 (1), and 18 (5). The top two SNPs, rs12681936 and rs12680789, in chromosome 8 had the
146 smallest p-values (5.96×10^{-8}), which were just shy of genome-wide significance ($p < 5 \times 10^{-8}$).
147 See Supplement Table 1 for a full list of all suggestive SNP associations with annotations from
148 Ensembl.org's variant effect predictor software [22]. Figure 2 shows a strong signal on
149 chromosome 8. This locus is better visualized by the locus zoom plot in Figure 3 A. This locus is
150 within < 0.1 megabases of the CLD23 gene.

151 **3.2.2 Sensitivity Analysis**

152 Sensitivity analysis of the interaction between the spline of stroke onset to motor drift
153 measurement revealed that two SNPs had significant interactions at a FDR of 10%. They were
154 rs113693489 in chromosome 6 and rs2967308 in chromosome 16. Supplement Table 2 contains
155 all the interaction estimates and their q values. In general, SNP interactions with the first part of
156 the spline (Days from stroke onset to measurement < 250) had a mean (\pm sd) Odds Ratio of
157 0.967 (± 1.33). The interactions with the second part of the spline (≥ 250 days) had a mean (\pm
158 sd) Odds Ratio of 0.806 (± 1.33). Highlighting the chromosome 8 locus, Figure 4 shows the odds
159 ratio point estimate and their 95% confidence intervals.

160 The second sensitivity analysis focused on the interaction of the suggestive SNPs based on early
161 versus later enrollment into the VISIP trial from stroke onset. The SNPs' p-values underwent a

162 FDR adjustment of 10%. In contrast to the interaction analysis, all Early and Late enrollment
163 q-values were significant; See Supplemental Table 3. Early enrollment interactions had a mean
164 (\pm sd) Odds Ratio of 0.419 (\pm 1.68). Late enrollment interaction had a similar mean (\pm sd) Odds
165 Ratio of 0.397 (\pm 1.70). Figure 5 exhibits the estimates for each suggestive SNP interaction with
166 Early versus Late enrollment in the chromosome 8 locus. While our sensitivity analysis models
167 show that each SNP interaction is an independent predictor of motor improvement, the 95%
168 confidence intervals have large overlaps. The overlaps suggest that Odds Ratios for each SNP
169 interaction do not differ from Early versus Late enrollment in the VISIP trial from stroke onset.

170 **3.2.3 Look-Up Analysis**

171 Out of the 500 reported SNPs ($p < 5 \times 10^{-6}$) from the GISCOME study [2], only 414 were present
172 in our analysis results. After applying a FDR of 10%, none of the look up SNPs from the
173 GISCOME study reached significance.

174 **4 Discussion**

175 Our GWAS of post stroke motor recovery failed to show genome-wide significant associations.
176 However, we found 115 novel suggestive SNPs linked to the odds of motor recovery over a two
177 years in the first ever motor-specific post stroke recovery GWAS. These suggestive SNPs
178 mapped to genomic loci connected to genes that are previously unknown as either candidate
179 genes or ones from prior GWAS studies [2, 3]. Following from here, we discuss the genetic loci of
180 interest associated with motor recovery individually.

181 **4.1 Post Stroke Motor Recovery Genetic Loci**

182 The Chromosome 8 locus's apex, as seen in Figure 3 A, is < 0.1 megabases from the CLD23
183 gene. CLD23 (Claudin 23) is a protein encoding gene part of the Claudin gene family which are
184 integral membrane proteins and components to tight junction strands[23]. CLDN23 has related
185 pathways affecting the blood brain barrier and immune cell transmigration according to
186 [genecards.org's pathway unification database \(https://pathcards.genecards.org/\)](https://pathcards.genecards.org/).

187 Additionally, CLDN23 variants are associated with blood cholesterol, triglyceride, and lipid
188 measurements[24–26].

189 Unlike the chromosome 8 locus, the chromosome 9 locus finds itself within the PTPRD gene; part
190 of the protein tyrosine phosphate (PTP) family. The PTPRD gene has a protein to protein
191 interaction at the neuronal synapse located at the presynaptic terminal surface. It has related
192 pathways of cell growth, differentiation, mitotic cycle, and oncogenic transformation[27, 28].

193 Interestingly, PTPRD has an association with glioblastoma [29].

194 Figure 3 C shows the chromosome 12 locus within in the RIMS-Binding Protein (RIMBP2) gene.

195 As the name suggests, this gene produces a binding protein. The function of this protein is
196 predicted to involve neuromuscular synaptic transmission. It is also highly expressed in brain
197 tissue. Butola et al. in 2021 reported that the role of RIM-BP2 is to link voltaged-gated Ca^{2+}
198 channels and release sites of synaptic vesicles[30]. They explain that RIMBP2 disruption leads to
199 alterations in Cav2.1 channel topography at active zones. These active zones affect
200 neurotransmitter release. The top SNP (rs73156962) of this locus has a direct biological
201 interpretation ($p = 0.034$) in nucleus accumbens located in the basal ganglia, a highly dense
202 interconnected neuronal tissue [31].

203 The chromosome 18 locus sits almost equally between two genes, RTTN and SOCS6, each
204 within 0.1 megabases. See Figure 3 D. RTTN (Rotatin) encodes a large protein without a known
205 specific function. However, knockout mice models result in neural tube defects [32]. In humans,
206 RTTN pathological variants lead to microcephaly and polymicrogyria with seizures [33, 34]. Even
207 though RTTN is linked to neurological structure and disorder in humans, there remains a notable
208 lack of published literature on this gene and its biological mechanisms. However, SOCS6
209 (Supressor of Cytokine Signalling 6) is part of the supressor cytokine signalling protein family
210 which plays a key role in inflammation regulation and insulin signalling in human brain tissue,
211 especially brain tissue affected by a neuro-degenerative disease [35].

212 **4.2 Literature Comparison**

213 Söderholm et al. [2] performed the largest GWAS of stroke functional recovery to date. Their
214 analysis consisted of 12 studies which totaled 6,021 patients. They defined functional stroke

215 recovery as the obtainment of a mRS score ≤ 2 as "Good Recovery" versus ≥ 3 as "Poor
216 Recovery" in their case/control approach. The study found only 1 SNP (rs1842681) significant at
217 the genome-wide level ($p < 5 \times 10^{-8}$) located in the LOC105372028 gene. The LOC105372028
218 gene has no known biological function. They also found 33 suggestive SNPs among 12 different
219 loci. When they utilized the mRS scores as a ordinal response instead of a binary one, the
220 number of suggestive SNPs increased to 75 spread over 17 distinct loci without an increase in
221 genome-wide significant SNPs.

222 When comparing Söderholm et al.'s study with ours, there are two notable distinctions. First, the
223 set of associated genes of each study are unique. None of our associated genes replicated with
224 theirs. This fact is intriguing because the unique set of genes from each study may be due to the
225 functional recovery measures utilized. Our GWAS analysis used the motor drift scores from the
226 NIHSS as a specific motor behavior marker. Plus, most of the patients in our discovery cohort
227 had the greatest weakness in the upper extremity instead of the lower which suggests that motor
228 drift score changes may not correlate with changes in the mRS. Unlike the motor drift score, the
229 mRS encompasses multiple phenotypic domains like cognition, motor strength, balance, and
230 mortality. The mRS measure has the probability of associating with genes that have general or
231 systemic biological effects as well as include genes expressed in other tissues that interact with
232 brain tissue like cardiovascular and lymphatic tissues.

233 The second distinction is the clear difference in the number of patients in each study, ours
234 $n = 488$ and Söderholm et al. $n = 6,201$. We capitalized on the repeated motor drift scores
235 measurements. The logistic regression model with GEE greatly enhanced the statistical
236 efficiency to find SNPs of interest associated with post-stroke motor recovery. In fact this study's
237 had about one-tenth the minimum recommended sample size for GWAS studies [36, 37]. Thus,
238 our analysis is a proof-of-principle that longitudinal observational studies can be a strong design
239 for future stroke recovery genomic studies.

240 To note, the genetic loci of Söderholm et al. and ours, did not include well published candidate
241 stroke recovery genes of APOE, BDNF or COX-2 [1, 38–41]. This has particular interest since
242 one would imagine that at least one of these genes would present themselves in either our results
243 or those of Söderholm et al. Even more so in line with Söderholm et al. because of the use of the
244 mRS as a recovery measure like previous candidate gene studies. One possible explanation is

245 that GWAS studies remain too under-powered to detect the effect size of known candidate genes.
246 The effect sizes of the candidate genes may be smaller than anticipated. To address the issue of
247 being under-powered, the stroke recovery community needs more genetic data linked to specific
248 stroke recovery phenotypes of interest or at least capitalize on observational stroke outcome
249 studies with longitudinal designs overlaid on relevant recovery milestones.

250 **4.3 Limitations**

251 Unfortunately, our study does not have a replication cohort, despite searching internationally for
252 other cohorts with NIHSS subscores and genetic data. For example, the National Institute of
253 Neurological Disorders and Stroke (NINDS) archived clinical research database has 22 publicly
254 available stroke study datasets, but none of them have NIHSS subscores and genetic data.
255 However, we performed a look-up analysis based on Söderholm et al.'s findings [2] as a
256 reasonable surrogate replication cohort. Another limitation of study is related to the VISP
257 enrollment criteria. Patients enrolled into the VISP trial must have had a stroke due to
258 atheroembolic mechanisms. Potential patients were excluded if their stroke was the result of a
259 cardioembolic source. It is possible that our discovery cohort of patients had more small vessel
260 strokes compared to large artery and other stroke types. The biological mechanisms deployed
261 and their effect on stroke recovery may differ among these subtypes, especially since large artery
262 and cardioembolic stroke tend to have larger stroke lesion volumes than small vessel strokes.
263 Small vessel stroke may relate more to chronic inflammatory or hypertension exposures which
264 may explain CLDN23 as the most promising finding. Unfortunately, the VISP trial did not collect
265 stroke subtype data like the TOAST criteria [42]. We are unable to investigate how the genetic
266 associations may differ among stroke subtypes.

267 **5 Conclusion**

268 We demonstrated the first ever use of repeated measurements and a domain-specific phenotype
269 in a stroke recovery GWAS. This resulted in the discovery of new genes associations. As a
270 proof-of-principle, this GWAS repurposed the NIHSS in a rich stroke clinical trial data set in line
271 with the recommendations from Braun et al. [10] and the Stroke Recovery and Rehabilitation

272 Roundtable [9]. This study's approach may have a great impact on future genetic stroke study
273 design. Longitudinal design allows one to investigate if the SNP effect is associated with changes
274 over time, which we believe is critical in stroke recovery genetics research.

275 **6 Author Contributions**

276 CA, RB, KK, FH, BW contributed to the conception and design of the study. MS and BW were
277 instrumental in the acquisition of phenotype and genotype data. CA and BW performed
278 phenotype harmonization. CA performed the statistical analysis. FH consulted on the statistical
279 analysis. CA wrote the first draft of the manuscript. All authors contributed to manuscript revision,
280 read, and approved the submitted version.

281 **7 Acknowledgments**

282 We wish to thank all of the participants in the Vitamin Intervention for Stroke Prevention trial that
283 made this study possible. The authors acknowledge Research Computing at The University of
284 Virginia for providing computational resources and technical support that have contributed to the
285 results reported within this publication. URL: <https://rc.virginia.edu>

286 **8 Funding**

287 The GWAS component of the VISP study was supported by the United States National Human
288 Genome Research Institute (NHGRI), Grant U01 HG005160 (PI Michèle Sale & Bradford
289 Worrall), as part of the Genomics and Randomized Trials Network (GARNET). Genotyping
290 services were provided by the Johns Hopkins University Center for Inherited Disease Research
291 (CIDR), which is fully funded through a federal contract from the NIH to the Johns Hopkins
292 University. Assistance with data cleaning was provided by the GARNET Coordinating Center
293 (U01 HG005157; PI Bruce S Weir). Study recruitment and collection of datasets for the VISP
294 clinical trial were supported by an investigator-initiated research grant (R01 NS34447; PI James
295 Toole) from the United States Public Health Service, NINDS, Bethesda, Maryland. Control data

296 for comparison with European ancestry VISP stroke cases were obtained through the database
297 of genotypes and phenotypes (dbGAP) High Density SNP Association Analysis of Melanoma:
298 Case-Control and Outcomes Investigation (phs000187.v1.p1; R01CA100264, 3P50CA093459,
299 5P50CA097007, 5R01ES011740, 5R01CA133996, HHSN268200782096C; PIs Christopher
300 Amos, Qingyi Wei, Jeffrey E. Lee). For VISP stroke cases of African ancestry, a subset of the
301 Healthy Aging in Neighborhoods of Diversity across the Life Span study (HANDLS) were used as
302 stroke free controls. HANDLS is funded by the National Institute of Aging (1Z01AG000513; PI
303 Michele K. Evans).

304 **9 Disclosures**

305 The authors declare that the research was conducted in the absence of any commercial or
306 financial relationships that could be construed as a potential conflict of interest. Dr. Bradford
307 Worrall is the Deputy Editor for the journal *Neurology*.

308 **10 Data Availability Statement**

309 The genetic associations results generated by this study can be found in the Cerebrovascular
310 Disease Knowledge Portal (<https://cd.hugeamp.org/XXXX>).

311 **References**

- 312 1. Lindgren, A. & Maguire, J. Stroke recovery genetics. *Stroke* **47**, 2427–2434 (2016).
- 313 2. Söderholm, M. *et al.* and the Genetics of Ischaemic Stroke Functional Outcome (GISCOME)
314 Network. Genome-wide association meta-analysis of functional outcome after ischemic
315 stroke. *Neurology* **92**, e1271–e1283 (2019).
- 316 3. Mola-Caminal, M. *et al.* PATJ low frequency variants are associated with worse ischemic
317 stroke functional outcome: a genome-wide meta-analysis. *Circulation research* **124**,
318 114–120 (2019).

- 319 4. Tam, V. et al. Benefits and limitations of genome-wide association studies. *Nature Reviews*
320 *Genetics* **20**, 467–484 (2019).
- 321 5. Campagna, M. P. et al. Epigenome-wide association studies: current knowledge, strategies
322 and recommendations. *Clinical epigenetics* **13**, 1–24 (2021).
- 323 6. Fugl-Meyer, A. R., Jääskö, L., Leyman, I., Olsson, S. & Steglind, S. A method for evaluation
324 of physical performance. *Scand J Rehabil Med* **7**, 13–31 (1975).
- 325 7. Siironen, J. et al. The Met allele of the BDNF Val66Met polymorphism predicts poor outcome
326 among survivors of aneurysmal subarachnoid hemorrhage. *Stroke* **38**, 2858–2860 (2007).
- 327 8. Cramer, S. C. et al. Genetic Factors, Brain Atrophy, and Response to Rehabilitation
328 Therapy After Stroke. *Neurorehabilitation and neural repair* **36**, 131–139 (2022).
- 329 9. Kwakkel, G. et al. Standardized measurement of quality of upper limb movement after
330 stroke: consensus-based core recommendations from the second stroke recovery and
331 rehabilitation roundtable. *Neurorehabilitation and neural repair* **33**, 951–958 (2019).
- 332 10. Braun, R. G. et al. What the Modified Rankin Isn't Ranking: Domain-Specific Outcomes for
333 Stroke Clinical Trials. *Neurology* (2021).
- 334 11. Toole, J. F. et al. Lowering homocysteine in patients with ischemic stroke to prevent
335 recurrent stroke, myocardial infarction, and death: the Vitamin Intervention for Stroke
336 Prevention (VISP) randomized controlled trial. *JAMA* **291**, 565–75. ISSN: 1538-3598.
337 <http://www.ncbi.nlm.nih.gov/pubmed/14762035> (5 Feb. 2004).
- 338 12. Keene, K. L. et al. Genetic associations with plasma B12, B6, and folate levels in an
339 ischemic stroke population from the Vitamin Intervention for Stroke Prevention (VISP) trial.
340 *Frontiers in public health* **2**, 112 (2014).
- 341 13. Consortium, I. H. 3. et al. Integrating common and rare genetic variation in diverse human
342 populations. *Nature* **467**, 52 (2010).
- 343 14. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program.
344 2019. *Genomics* (2019).
- 345 15. Das, S. et al. Next-generation genotype imputation service and methods. *Nature genetics*
346 **48**, 1284–1287 (2016).

- 347 16. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation.
348 *Bioinformatics* **31**, 782–784 (2015).
- 349 17. Vincent J Carey, T. L., src/dgedi.f, B. R. F. &
350 src/dgefa.f are for LINPACK authored by Cleve Moler. Note that maintainers are not
351 available to give advice on using a package they did not author. *gee: Generalized*
352 *Estimation Equation Solver* R package version 4.13-20 (2019).
353 <https://CRAN.R-project.org/package=gee>.
- 354 18. Zeger, S. L. & Liang, K.-Y. Longitudinal data analysis for discrete and continuous outcomes.
355 *Biometrics*, 121–130 (1986).
- 356 19. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies.
357 *Bioinformatics* **26**, 2867–2873 (2010).
- 358 20. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S* Fourth. ISBN
359 0-387-95457-0. <https://www.stats.ox.ac.uk/pub/MASS4/> (Springer, New York, 2002).
- 360 21. Bernhardt, J. et al. Agreed definitions and a shared vision for new standards in stroke
361 recovery research: the stroke recovery and rehabilitation roundtable taskforce. *International*
362 *Journal of Stroke* **12**, 444–450 (2017).
- 363 22. McLaren, W. et al. The ensembl variant effect predictor. *Genome biology* **17**, 1–14 (2016).
- 364 23. Katoh, M. & Katoh, M. CLDN23 gene, frequently down-regulated in intestinal-type gastric
365 cancer, is a novel member of CLAUDIN gene family. *International journal of molecular*
366 *medicine* **11**, 683–689 (2003).
- 367 24. Sinnott-Armstrong, N. et al. Genetics of 35 blood and urine biomarkers in the UK Biobank.
368 *Nat Genet* **53**, 185–194 (Feb. 2021).
- 369 25. Richardson, T. G. et al. Characterising metabolomic signatures of lipid-modifying therapies
370 through drug target mendelian randomisation. *PLoS Biol* **20**, e3001547 (Feb. 2022).
- 371 26. Adewuyi, E. O., Mehta, D. & Nyholt, D. Genetic overlap analysis of endometriosis and
372 asthma identifies shared loci implicating sex hormones and thyroid signalling pathways.
373 *Human Reproduction* **37**, 366–383 (2022).

- 374 27. Pulido, R., Krueger, N. X., Serra-Pagès, C., Saito, H. & Streuli, M. Molecular
375 Characterization of the Human Transmembrane Protein-tyrosine Phosphatase δ :
376 EVIDENCE FOR TISSUE-SPECIFIC EXPRESSION OF ALTERNATIVE HUMAN
377 TRANSMEMBRANE PROTEIN-TYROSINE PHOSPHATASE δ ISOFORMS (\square). *Journal of*
378 *Biological Chemistry* **270**, 6722–6728 (1995).
- 379 28. Mizuno, K. *et al*. MPTP delta, a putative murine homolog of HPTP delta, is expressed in
380 specialized regions of the brain and in the B-cell lineage. *Molecular and cellular biology* **13**,
381 5513–5523 (1993).
- 382 29. Wu, G. *et al*. The genomic landscape of diffuse intrinsic pontine glioma and pediatric
383 non-brainstem high-grade glioma. *Nature genetics* **46**, 444 (2014).
- 384 30. Butola, T. *et al*. RIM-Binding Protein 2 organizes Ca²⁺ channel topography and regulates
385 release probability and vesicle replenishment at a fast central synapse. *Journal of*
386 *Neuroscience* **41**, 7742–7767 (2021).
- 387 31. Consortium, G. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene
388 regulation in humans. *Science* **348**, 648–660 (2015).
- 389 32. Kia, S. K. *et al*. RTTN mutations link primary cilia function to organization of the human
390 cerebral cortex. *The American Journal of Human Genetics* **91**, 533–540 (2012).
- 391 33. Cavallin, M. *et al*. Recurrent RTTN mutation leading to severe microcephaly, polymicrogyria
392 and growth restriction. *European journal of medical genetics* **61**, 755–758 (2018).
- 393 34. Stouffs, K. *et al*. Biallelic mutations in RTTN are associated with microcephaly, short stature
394 and a wide range of brain malformations. *European journal of medical genetics* **61**, 733–737
395 (2018).
- 396 35. Walker, D., Whetzel, A. & Lue, L.-F. Expression of suppressor of cytokine signaling genes in
397 human elderly and Alzheimer’s disease brains and human microglia. *Neuroscience* **302**,
398 121–137 (2015).
- 399 36. Ziyatdinov, A. *et al*. Estimating the effective sample size in association studies of
400 quantitative traits. *G3* **11**, jkab057 (2021).

- 401 37. Moore, C. M., Jacobson, S. A. & Fingerlin, T. E. Power and sample size calculations for
402 genetic association studies in the presence of genetic model misspecification. *Human*
403 *heredity* **84**, 256–271 (2019).
- 404 38. Kim, D. Y., Quinlan, E. B., Gramer, R. & Cramer, S. C. BDNF Val66Met polymorphism is
405 related to motor system function after stroke. *Physical therapy* **96**, 533–539 (2016).
- 406 39. Kim, J.-M. *et al.* Associations of BDNF genotype and promoter methylation with acute and
407 long-term stroke outcomes in an East Asian cohort. *PloS one* **7**, e51280 (2012).
- 408 40. Chan, A., Yan, J., Csurhes, P., Greer, J. & McCombe, P. Circulating brain derived
409 neurotrophic factor (BDNF) and frequency of BDNF positive T cells in peripheral blood in
410 human ischemic stroke: effect on outcome. *Journal of neuroimmunology* **286**, 42–47 (2015).
- 411 41. Gómez-Pinilla, F., Ying, Z., Roy, R. R., Molteni, R. & Edgerton, V. R. Voluntary exercise
412 induces a BDNF-mediated mechanism that promotes neuroplasticity. *Journal of*
413 *neurophysiology* **88**, 2187–2195 (2002).
- 414 42. Jr, H. P. A. *et al.* Classification of subtype of acute ischemic stroke. Definitions for use in a
415 multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. *stroke* **24**,
416 35–41 (1 1993).

417 **11 Tables**

VISP Cohort (n=488)	
Treatment Arm	
High Dose	225 (47%)
Low Dose	251 (53%)
Age in Years	
Mean (SD)	66 (± 11)
Sex	
Male	310 (64%)
Weakest Limb	
Arm	333 (67%)
Stroke Onset to Enrollment	
Days	72.5 (± 31.2)
Body Mass Index (BMI)	
N-Miss	7
Mean (SD)	28.54 (6.47)
Hypertension	
N-Miss	1
No	114 (23%)
Yes	383 (77%)
Ever Smoker	
N-Miss	1
No	173 (36%)
Yes	314 (64%)
Diabetes Mellitus Type II	
No	331 (68%)
Yes	157 (32%)
Ancestry	
European	347 (73%)
African	92 (19%)
Other	37 (8%)

Table 1: Vitamin Intervention for Stroke Prevention (VISP) trial demographics. Information is presented as counts (percentages) or means (standard deviations).

418 12 Figure Legends

419 **Figure 1:** Shows the non-linear relationship of the mean probability of motor improvement for
420 each follow up time point since study enrollment. The green and orange lines segments highlight
421 the notable change in slopes from 1 to 6 month visits to 6 to 24 month visits. The difference in
422 slope between the two time periods is similar to stroke rehab trials of the upper extremity.

423 **Figure 2:** This Manhattan plot shows each SNP and its $-\log_{10}(p \text{ value})$ associated with post
424 stroke motor improvement. None of the SNPs reached genome-wide significant (above the red
425 line). However, 115 SNPs had suggestive associations (above the blue line) with 2 right under
426 the red line. The most convincing genetic loci is the large spike in chromosome 8; near the
427 Claudin 23 gene. This gene affects blood brain barrier and immune cell transmigration.

428 **Figure 3:** Panel plot of Locus Zoom figures (A-D) corresponding to genetic loci of interest. The
429 colors refers to correlation of each SNP to the top SNP in each panel with red having an
430 $r^2 \geq 0.80$. **A** Genetic locus near the CLDN23 gene on chromosome 8. **B** Genetic locus within the
431 PTPRD gene on chromosome 9. **C** Genetic locus within the RIMBP2 gene on chromosome 12. **D**
432 Genetic locus between the RTTN and SOCS6 genes on chromosome 18.

433 **Figure 4:** Shows the estimates and 95% confidence intervals of the interaction of the study
434 timepoints 1 to 6 months versus 6 to 24 months with each suggestive SNP found in Chromosome
435 8. Interestingly, the interaction estimates of many SNPs seem to straddle one with SNP
436 estimates at 1 to 6 months having a greater odds of motor recovery, while SNP estimates at 6 to
437 24 months having less.

438 **Figure 5:** Shows odds ratios and 95% confidence intervals of motor drift score improvement from
439 the interaction Early versus Late post-stroke enrollment by SNP in Chromosome 8. The odds
440 ratios estimates for Early versus Late do not have a discernible pattern or consistency.

441 **13 Figures**

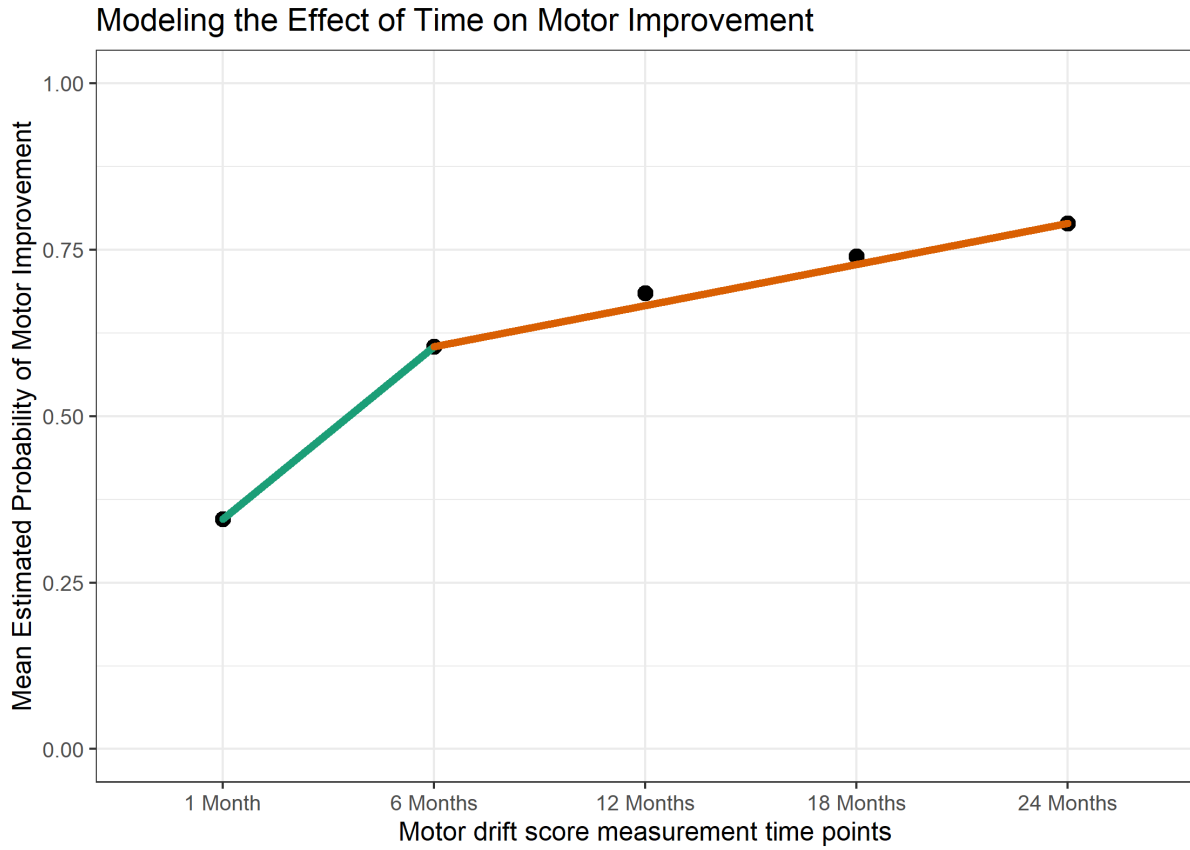


Figure 1: Shows the non-linear relationship of the mean probability of motor improvement for each follow up time point since study enrollment. The green and orange line segments highlight the notable change in slopes from 1 to 6 month visits to 6 to 24 month visits. The difference in slope between the two time periods is similar to stroke rehab trials of the upper extremity.

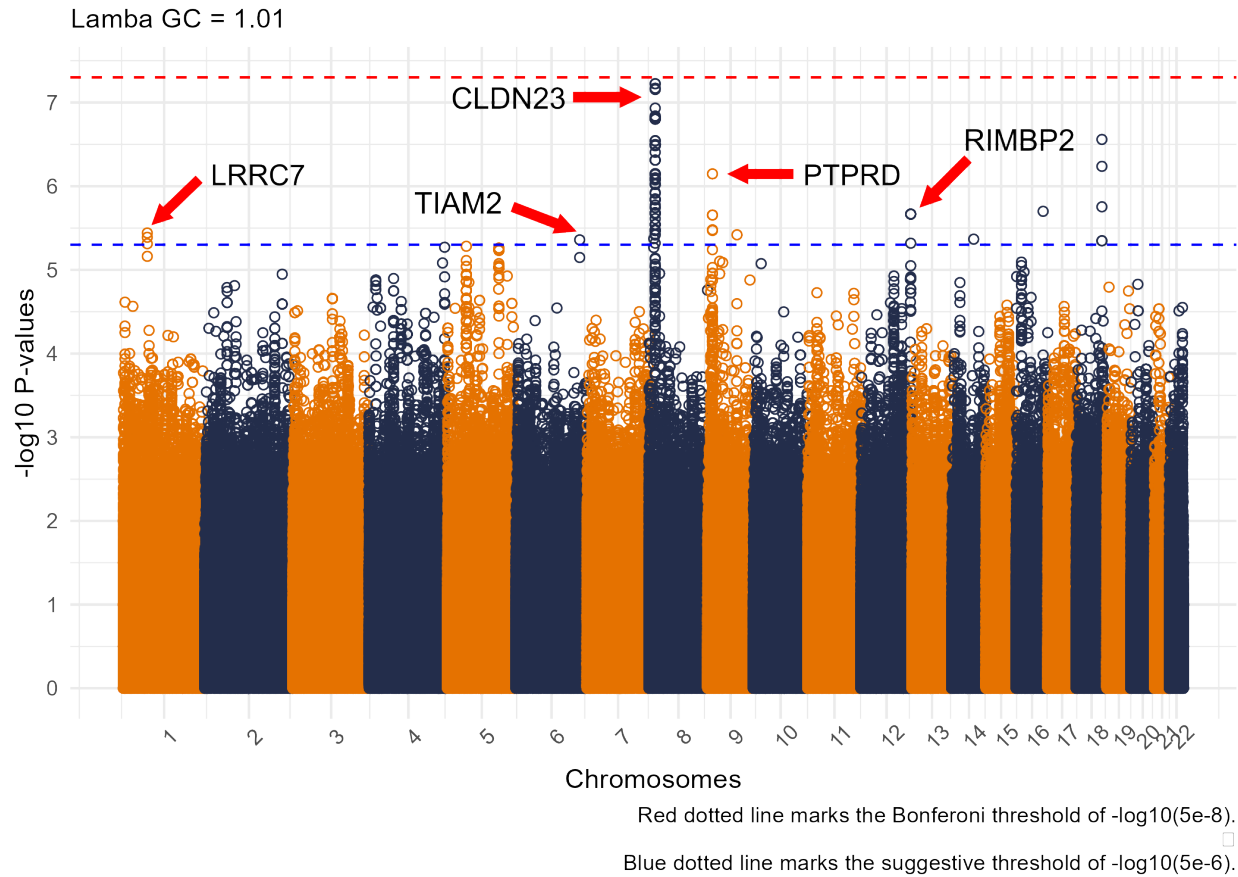


Figure 2: This Manhattan plot shows each SNP and its $-\log_{10}(p \text{ value})$ associated with post stroke motor improvement. None of the SNPs reached genome-wide significant (above the red line). However, 115 SNPs had suggestive associations (above the blue line) with 2 right under the red line. The most convincing genetic loci is the large spike in chromosome 8; near the Claudin 23 gene. This gene affects blood brain barrier and immune cell transmigration.

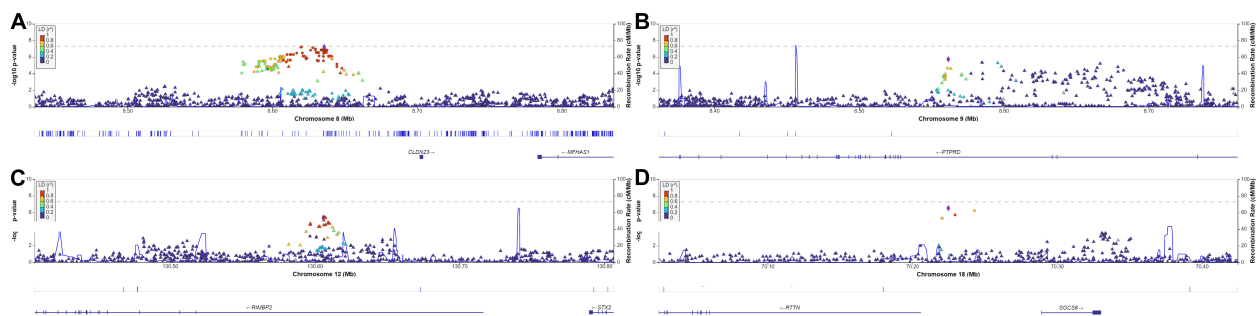


Figure 3: Panel plot of Locus Zoom figures (A-D) corresponding to genetic loci of interest. The colors refers to correlation of each SNP to the top SNP in each panel with red having an $r^2 \geq 0.80$. **A** Genetic locus near the CLDN23 gene on chromosome 8. **B** Genetic locus within the PTPRD gene on chromosome 9. **C** Genetic locus within the RIMBP2 gene on chromosome 12. **D** Genetic locus between the RTTN and SOCS6 genes on chromosome 18.

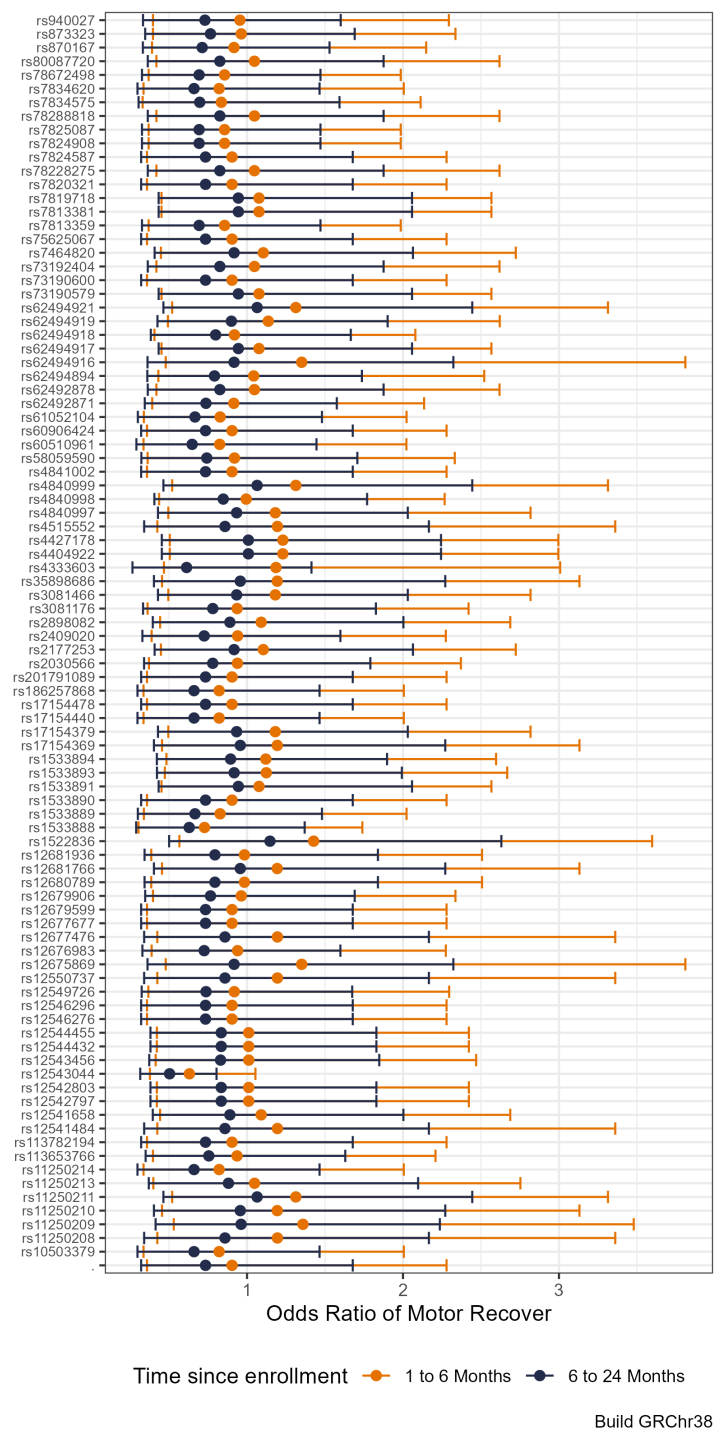


Figure 4: Shows the estimates and 95% confidence intervals of the interaction of the study time-points 1 to 6 months versus 6 to 24 months with each suggestive SNP found in Chromosome 8. Interestingly, the interaction estimates of many SNPs seem to straddle one with SNP estimates at 1 to 6 months having a greater odds of motor recovery, while SNP estimates at 6 to 24 months having less.

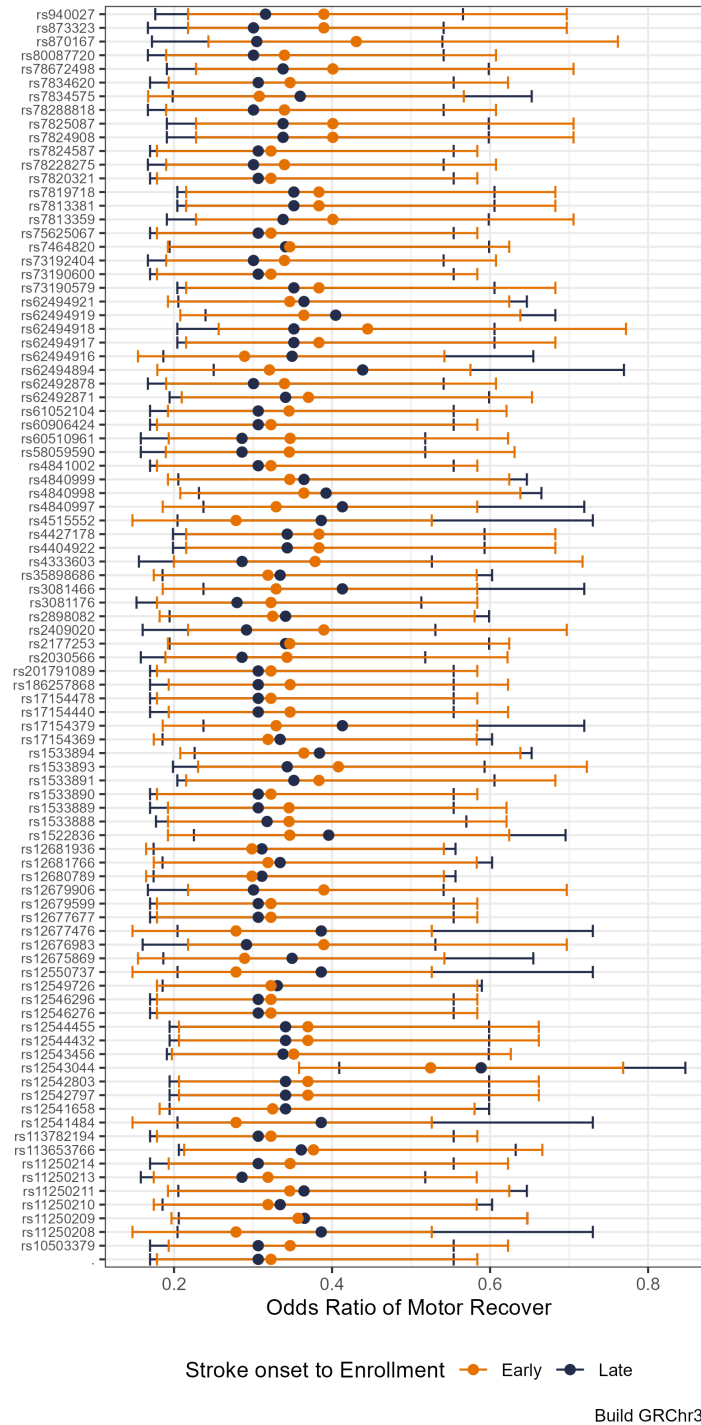


Figure 5: Shows odds ratios and 95% confidence intervals of motor drift score improvement from the interaction Early versus Late post-stroke enrollment by SNP in Chromosome 8. The odds ratios estimates for Early versus Late do not have a discernible pattern or consistency.