Supplementary Document : More Results

Ja-an Lin [a], Hongtu Zhu[a,d], Rebecca Knickmeyer [b],

Martin Styner [c], John Gilmore [b], and Joseph G. Ibrahim[a]

Department of [a]Biostatistics, [b]Psychiatry, and [c]Computer Science,

and Biomedical Research Imaging Center[d],

University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Address for Correspondence:

Dr. Hongtu Zhu,

Department of Biostatistics, University of North Carolina at Chapel Hill,

McGavran Greenberg Hall, CB#7420,

Chapel Hill, NC 27599, U.S.A.

Email: hzhu@bios.unc.edu,

Phone: 1-919-966-7272.

This is a supplementary document for the paper *Projection Regression Models for Multivariate Imaging Phenotype*. In this document, we include two more clinical data analyses to demonstrate our proposed PRM method. Also, the detailed simulation results and the auxiliary data analysis results for the main paper are given here.

Additional Clinical Data Analysis

Besides the neonatal data analysis in the main paper, we also applied the PRM method on two additional datasets. The purpose of the first analysis is to explore the relationship between the cerebral fiber tract development in neonates and some potential impact factors. The primary aim of the other is to search the putative SNPs which have effects on Alzheimer's disease progression in terms of brain volume degeneration.

Diffusion Tensor Imaging Data

The first supplementary dataset is to investigate the potential impact factors on brain structure development in terms of cerebral fiber tracts. A total of 143 healthy full-term neonates (87M and 56F) with mean postnatal age of 25$\pm$17.9 days (range from 9 to 144 days) were recruited at the University of North Carolina at Chapel Hill (UNC-CH). Each subject was assessed with medical resonance (MR) images and certain demographic information at an one time visit. Efforts were made to ensure the subjects slept comfortably inside the MR scanner, and thus none of the subjects were sedated during imaging sessions. All images were acquired on a 3T Allegra head only MR system (Siemens Medical Inc., Erlangen, Germany) with a maximal gradient strength of 40 mT/m and a maximal slew rate of 400 mT/(m·msec). A single shot EPI DTI sequence (TR/TE=5400/73 msec) with eddy current compensation was used to obtain DTI images. Diffusion gradients with a b-value of 1000 s/mm^2 were applied in six non-collinear directions, (1,0,1), (-1,0,1), (0,1,1), (0,1,-1),

(1,1,0), and (-1,1,0). The reference scan (b=0) was also acquired for the construction of diffusion tensor matrices. Contiguous slices with slice thickness of 2mm were scanned to cover the whole brain. The voxel resolution was isotropic 2mm, and the in-plane field of view was 256mm in both directions. A total of five scans were acquired and averaged to improve the signal-to-noise ratio of the images.

We chose 3 fiber tracts of interest including the splenium, the genu and the left uncinate. For each tract, we computed the fractional anistropy (FA) values, and the three eigenvalues of the diffusion tensors at each grid point. The FA values denote the inhomogeneous extent of local barriers to water diffusion, and the three eigenvalues of the diffusion tensor may, respectively, reflect the magnitude of water diffusivity along and perpendicular to the long axis of white matter fibers [Song et al., 2003]. The four diffusion properties was then applied to PRM as responses, where the covariates of interest include age after birth in days, gender and gestational age at scanning. To acquire higher accuracy, we choose the number of wild bootstrap to be $10,000$. In the hypothesis testing, each covariate of interest was tested after adjusting for other covariates and the p-values were corrected by the false discovery rate (FDR) at the 0.05 significant level. The result shows that postnatal age has an influential effect on the 3 brain regions while the effects are not detected in gender and gestational age at birth. Table I in this document provides the numbers of grid points which show significance after FDR correction and the p-values along each fiber tract are illustrated in Figure 1.

For comparison purposes, we also applied PCA and CWR methods to the same data. In the PCA method, at each grid point we regressed the first principle component of the four diffusion properties on the same covariates of interest and the same null hypotheses were tested at 0.05 significant level. Both the unadjusted and FDR-adjusted testing results are shown in Table II.

In CWR method, a similar approach in simulation is implemented. At each grid point, each diffusion property was regressed on the same covariates by an univariate linear model. Then the

Table I: **Numbers of significant point Analyzed by PRM**

Brain Structural Region	Grid Points Measured	Gender	Age after Birth	G-Age
Genu	45	0	0	43
Splenium	70	0	0	51
Left Uncinate	55	0	0	31

The numbers of significant point analyzed by PRM for each covariate of interest - gender, age after birth (Age after Birth) and gestational age at scanning (G-Age) - after FDR correction using a 0.05 significant level in each brain structural region are provided.

Table II: **Numbers of Significant Point Analyzed by PCA**

Brain Structural Region	Grid Points Measured	Gender	Age after Birth	G-Age
Genu	45	0	7	10
Splenium	70	2	8	13
Left Uncinate	55	5	7	6
Genu	45	0	0	0
Splenium	70	0	0	3
Left Uncinate	55	0	0	0

The numbers of significant point analyzed by PCA using the *first* principle component for each covariate of interest before (upper section) and after FDR (lower section) correction using a 0.05 significant level in each brain structural region are provided.

Table III: **Numbers of Significant Point Analyzed by CWR**

Brain Structural Region	Grid Points Measured	Gender	Age after Birth	G-Age
Genu	45	0	16	32
Splenium	70	2	13	48
Left Uncinate	55	4	12	41

The numbers of significant point analyzed by component wise regression (CWR) for each covariate of interest after FDR correction using 0.05 significant level in each brain structural region are provided.

FDR correction at 0.05 significant level was applied to adjust the p-values after hypotheses testing. The Table III shows the number of significant grid points based on this approach.

Region of Interest Data

The second data is from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study. It is an ongoing longitudinal study with the primary purpose of exploring the genetic and neuroimaing information associated with late-onset Alzheimer disease (LOAD). The study recruited patients older than 65 diagnosed with mild cognitive impairment (MCI) ($n \approx 400$) or Alzheimer's disease (AD) ($n \approx 200$). Around 200 elders with healthy memory were recruited as controls as well. Each subject is followed at least 3 years. During the study period, the subjects are assessed with MRI images and diagnosis status evaluation at certain time points.

Our primary aim is to explore the potential SNPs effect on disease progression observing in brain volume degeneration. The brain volumes are characterize by 93 selected regions of interest extracted from the MRI images and the SNPs effect on disease progression are summarized by the interaction terms between the disease status and the SNPs. We selected 12 SNPs based on the database of

Table IV: **P-values of selected SNP obtained from PRM for ADNI data**

SNP	P-value	SNP	P-value	SNP	P-value
rs5930	0.944	rs213045	0.404	**rs1801133**	**0.006**
rs541458	0.678	rs939348	0.05	rs1049296	0.563
rs1143634	0.783	rs2075650	0.019	rs2306604	0.954
s4878104	0.08	**rs11136000**	**0.003**	rs17269348	0.958

AlzGene [Bertram et al., 2007] and then applied the PRM method to test if the interaction term equals zero. To increase the statistical power, we only included one SNP at a time. Besides the interaction term, we also incorporated the covariates gender, baseline age, baseline diagnosis status, intracranial volume (ICV) as well as the main effect of the SNP. To prevent the population effect, the subjects are restricted to Caucasians. The final sample size is 711 with 193 healthy controls, 343 MCIs and 175 ADs who have no missing data in ROIs, SNPs and other covariates. The selected SNPs and their corresponding p-values adjusted after FDR correction at 0.05 significant level are listed in Table IV.

We found that the SNP rs11136000 on chromosome 8, which is associated with the clusterin CLU gene and also known as the APOJ gene, shows strong association with Alzheimer's disease status in terms of brain volume change. Additionally, SNP rs1801133 on the MTHFR gene of Chromosome 1 is also detected having significant impact on the disease progression. These results are consistent with the findings in AlzGene. Hu et al. [2011] performed a longitudinal meta-analysis with 1034 ADs and 1186 control (including 300 ADs and 190 Controls from ADNI) using the Clinical Dementia Rating (CDR) Scale as the main response that the CDR is widely used clinically to assess the dementia level. They compared the differences between the two disease groups by a simple chi-square tests using PLINK [Purcell et al., 2007]. However, restricting the samples to ADNI subjects,

Table V: **P-values from the PCA method to analyze ADNI data**

SNP	P-value	SNP	P-value	SNP	P-value
rs5930	0.99	rs213045	0.99	rs1801133	0.93
rs541458	0.77	rs939348	0.99	rs1049296	0.69
rs1143634	0.99	rs2075650	0.88	rs2306604	0.99
s4878104	0.99	rs11136000	0.92	rs17269348	1

the CLU gene did not show significance (p-value = 0.787) on the CDR Scale. None of the current papers in AlzGene reported MTHFR using ADNI study. Thus, we anticipate this finding provides more pathological information for the future studies about Alzheimer's disease.

Similarly, to demonstrate comparison, we also applied the PCA and the CWR methods and tested the same null hypotheses. In the PCA approach, the first 3 principle components, which explains around 88% of variation from 93 ROIs, were regressed on the same set of covariates in a multivariate linear model. None of the interaction terms between disease status and the SNPs were significant using the Hotelling's T^2 test at 0.05 significant level. The detail of the testing results are displayed in Table V. Using the CWR method, each ROI was fitted with the same group of covariates by an univariate linear model. The interaction terms were tested by Student's t-test and the p-values were then adjusted by FDR at 0.05 significant level. None of the 12 SNPs were detected being significant using this approach as well.

Simulation Result in Detail

We list the detailed simulation results in Tables VI - XIII for reader's reference.

Table VI: **Type I Errors - Part 1**

q	50	100	150	200	50	100	150	200
p	N=150				N=200			
0.05	0.053	0.06	0.027	0.047	0.093	0.06	0.067	0.06
0.1	0.04	0.04	0.02	0.033	0.047	0.06	0.033	0.073
0.2	0.053	0.06	0.053	0.033	0.053	0.06	0.02	0.04
Principle Component p								
0.05	0.06	0.04	0.06	0.027	0.053	0.053	0.047	0.053
0.1	0.087	0.087	0.093	0.067	0.093	0.107	0.087	0.087
0.2	0.132	0.1	0.127	0.113	0.133	0.167	0.127	0.127
Multiple Comparison p								
0.05	0.033	0.04	0.047	0.047	0.073	0.08	0.067	0.087
0.1	0.087	0.067	0.107	0.093	0.12	0.113	0.14	0.113
0.2	0.132	0.16	0.172	0.147	0.153	0.16	0.18	0.153

This table gives the results of the *type I error* based on the sample sizes $N = 150$ and *200* for PRM multiple comparison and principle component methods, with *minor allele frequencies* 0.05, 0.1 and 0.2 to simulate the distribution of the genetic groups and number of responses q.

Table VII: **Type I Errors - Part 2**

q	50	100	150	200	50	100	150	200
p	N=150				N=200			
0.3	0.087	0.02	0.033	0.04	0.027	0.06	0.033	0.093
0.4	0.047	0.04	0.06	0.08	0.067	0.053	0.067	0.06
0.5	0.053	0.053	0.067	0.053	0.067	0.047	0.073	0.04
Principle Component p								
0.3	0.139	0.133	0.147	0.133	0.16	0.173	0.167	0.14
0.4	0.152	0.167	0.153	0.167	0.167	0.2	0.173	0.173
0.5	0.185	0.173	0.173	0.18	0.18	0.207	0.187	0.207
Multiple Comparison p								
0.3	0.219	0.193	0.205	0.173	0.265	0.207	0.258	0.213
0.4	0.265	0.207	0.258	0.213	0.2	0.247	0.247	0.233
0.5	0.298	0.213	0.311	0.253	0.227	0.26	0.313	0.253

The table gives the results of the *type I error* based on the sample sizes $N = 150$ and *200* for PRM multiple comparison and principle component methods, with *minor allele frequencies* 0.3, 0.4 and 0.5 to simulate the distribution of the genetic group and number of responses q.

Table VIII: **Type I Errors - Part 3**

q	50	100	150	200	50	100	150	200
p	N=250				N=300			
0.05	0.027	0.04	0.027	0.047	0.033	0.053	0.048	0.052
0.1	0.027	0.067	0.027	0.04	0.027	0.067	0.048	0.048
0.2	0.067	0.067	0.047	0.033	0.053	0.067	0.032	0.068
Principle Component p								
0.05	0.053	0.06	0.033	0.03	0.053	0.06	0.06	0.033
0.1	0.12	0.093	0.047	0.073	0.1	0.107	0.08	0.08
0.2	0.153	0.113	0.093	0.087	0.113	0.153	0.1	0.127
Multiple Comparison p								
0.05	0.06	0.04	0.073	0.053	0.053	0.08	0.06	0.04
0.1	0.107	0.087	0.127	0.093	0.073	0.153	0.1	0.12
0.2	0.153	0.16	0.18	0.153	0.147	0.12	0.16	0.127

The table gives the results of the *type I error* based on sample sizes $N = 250$ and *300* for PRM, multiple comparison and principle component methods, with *minor allele frequencies* 0.05, 0.1 and 0.2 to simulate the distribution of the genetic group and number of responses q.

Table IX: **Type I Errors - Part 4**

q	50	100	150	200	50	100	150	200
p	N=250				N=300			
0.3	0.073	0.047	0.053	0.087	0.073	0.107	0.06	0.04
0.4	0.06	0.027	0.06	0.04	0.093	0.053	0.068	0.044
0.5	0.06	0.033	0.053	0.047	0.107	0.053	0.068	0.032
Principle Component p								
0.3	0.2	0.14	0.113	0.1	0.12	0.167	0.107	0.147
0.4	0.213	0.173	0.14	0.14	0.16	0.213	0.16	0.167
0.5	0.24	0.207	0.193	0.153	0.173	0.22	0.193	0.187
Multiple Comparison p								
0.3	0.213	0.16	0.179	0.18	0.16	0.22	0.233	0.187
0.4	0.245	0.187	0.205	0.213	0.187	0.293	0.253	0.233
0.5	0.264	0.207	0.225	0.247	0.193	0.3	0.28	0.28

The table gives the results of the *type I error* based on sample sizes $N = 250$ and *300* for PRM, multiple comparison and principle component methods, with *minor allele frequencies* 0.3, 0.4 and 0.5 to simulate the distribution of the genetic group and number of responses q.

Table X: **Power - Part 1**

q	50	100	150	200	50	100	150	200
p	N=150				N=200			
0.05	0.44	0.413	0.293	0.367	0.78	0.67	0.573	0.46
0.1	0.66	0.52	0.467	0.433	0.853	0.807	0.713	0.587
0.2	0.713	0.627	0.53	0.513	0.9	0.887	0.78	0.713
Principle Component p								
0.05	0.087	0.04	0.067	0.027	0.08	0.047	0.043	0.067
0.1	0.113	0.087	0.1	0.093	0.147	0.107	0.107	0.107
0.2	0.173	0.127	0.133	0.147	0.207	0.153	0.153	0.14
Multiple Comparison p								
0.05	0.133	0.093	0.06	0.073	0.2	0.2	0.12	0.133
0.1	0.3	0.273	0.247	0.219	0.44	0.47	0.384	0.38
0.2	0.645	0.553	0.57	0.503	0.788	0.769	0.656	0.709

The table gives the results of the *power* based on sample sizes $N = 150$ and *200* for PRM, multiple comparison and principle component methods, with *minor allele frequencies* 0.05, 0.1 and 0.2 to simulate the distribution of the genetic group and number of responses q.

Table XI: **Power - Part 2**

q	50	100	150	200	50	100	150	200
p	N=150				N=200			
0.3	0.753	0.707	0.593	0.56	0.94	0.907	0.813	0.773
0.4	0.807	0.72	0.627	0.593	0.967	0.92	0.86	0.813
0.5	0.827	0.727	0.653	0.607	0.967	0.927	0.88	0.8
Principle Component p								
0.3	0.227	0.167	0.18	0.167	0.26	0.187	0.193	0.167
0.4	0.273	0.2	0.193	0.193	0.3	0.227	0.207	0.213
0.5	0.313	0.227	0.213	0.26	0.347	0.257	0.247	0.247
Multiple Comparison p								
0.3	0.834	0.768	0.795	0.695	0.894	0.94	0.848	0.9
0.4	0.94	0.887	0.868	0.788	0.967	0.967	0.907	0.954
0.5	0.96	0.907	0.914	0.868	0.987	0.974	0.968	0.967

The table gives the results of the *power* of sample sizes $N = 150$ and *200* for PRM, multiple comparison and principle component methods, with *minor allele frequencies* 0.3, 0.4 and 0.5 to simulate the distribution of the genetic group and number of responses q.

Table XII: **Power - Part 3**

q	50	100	150	200	50	100	150	200
p	N=250				N=300			
0.05	0.893	0.86	0.813	0.72	0.967	0.893	0.86	0.847
0.1	0.96	0.947	0.887	0.807	0.987	0.94	0.94	0.913
0.2	0.973	0.993	0.953	0.893	0.993	0.973	0.987	0.98
Principle Component p								
0.05	0.067	0.06	0.033	0.033	0.073	0.073	0.06	0.04
0.1	0.18	0.1	0.073	0.087	0.14	0.107	0.093	0.1
0.2	0.22	0.14	0.133	0.107	0.193	0.2	0.147	0.153
Multiple Comparison p								
0.05	0.268	0.153	0.193	0.14	0.32	0.293	0.225	0.18
0.1	0.533	0.467	0.53	0.353	0.662	0.647	0.523	0.513
0.2	0.914	0.834	0.887	0.775	0.947	0.927	0.94	0.894

The table gives the results of the *power* of sample sizes $N = 250$ and *300* for PRM, multiple comparison and principle component methods, with *minor allele frequencies* 0.05, 0.1 and 0.2 to simulate the distribution of the genetic group and number of responses q.

Table XIII: **Power - Part 4**

q	50	100	150	200	50	100	150	200
p	N=250				N=300			
0.3	0.98	0.98	0.967	0.9	1	0.993	0.993	0.967
0.4	0.987	0.987	0.96	0.927	1	0.993	0.993	0.967
0.5	1	0.987	0.98	0.9	1	0.993	0.993	0.987
Principle Component								
p								
0.3	0.287	0.207	0.18	0.133	0.247	0.253	0.18	0.2
0.4	0.331	0.26	0.227	0.187	0.32	0.307	0.252	0.24
0.5	0.364	0.3	0.28	0.212	0.371	0.327	0.298	0.267
Multiple Comparison								
p								
0.3	0.987	0.96	0.947	0.921	1	0.98	0.987	0.973
0.4	0.993	0.986	0.993	0.967	1	1	1	1
0.5	1	1	1	0.993	1	1	1	1

The table gives the results of the *power* of sample sizes $N = 250$ and *300* for PRM, multiple comparison and principle component methods, with *minor allele frequencies* 0.3, 0.4 and 0.5 to simulate the distribution of the genetic group and number of responses q.

Table XIV: **P-values from PCA method to analyze neonatal data**

SNP	P-value	SNP	P-value	SNP	P-value
rs4680	0.99	rs821616	0.79	rs6675281	0.95
rs35753505	0.99	rs6994992	0.99	rs9340799	0.79
rs2234693	0.92	rs6265	0.99	rs2270335	0.98

Neonatal Data Analysis Result in Detail

We display the detailed results using the PCA method to analyze the neonatal data discussed in the main paper in Table XIV.

Reference

Bertram L, McQueen M, Mullin K, Blacker D, Tanzi RE. 2007. Systematic meta-analyses of alzheimer disease genetic association studies: the alzgene database. Nature Genetics 39:17-23.

Hu X, Pickering E, Liu Y, Hall S, Fournier H, Katz E, Dechairo B, John S, Eerdewegh PV, Soares H, ADNI. 2011. Meta-analysis for genome-wide association study identifies multiple variants at the bin1 locus associated with late-onset alzheimer's disease. PLoS One 24;6: e16616.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PLW, Daly MJ, Sham PC. 2007. Plink: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81:559-575.

S. K. Song and S. W. Sun and W. K. Ju and S. J. Lin and A. H. Cross and A. H. Neufeld. 2003. Diffusion tensor imaging detects and differentiates axon and myelin degeneration in mouse optic nerve after retinal ischemia. Neuroimage 20:1714-1722.

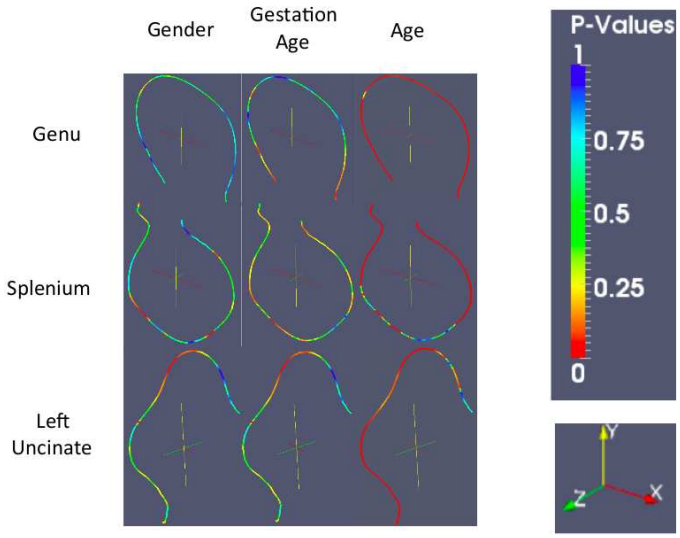

Fig. 1: The p-values along each fiber tract. The fiber tract from the top row to the lowest row are *Genu*, *Splenium* and *Left Uncinate*, respectively. The effects from left to right are gender, age after birth and gestational age at scanning. The color closer to red means the grid point at the fiber is more significant to the corresponding effect.