Biometrics 000, 000–000 DOI: 000
000 0000

Supporting Information for "Identifying Disease-Associated Biomarker Network Features Through Conditional Graphical Model" by Shanghong Xie, Xiang Li, Peter McColgan, Rachael I. Scahill, Donglin Zeng, and Yuanjia Wang

Shanghong Xie[1], Xiang Li[2], Peter McColgan[3], Rachael I. Scahill[3], Donglin Zeng[4], and Yuanjia Wang[1,5]

[1]Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY, U.S.A.

[2]Statistics and Decision Sciences, Janssen Research & Development, LLC, Raritan, NJ, U.S.A.

[3]Huntington's Disease Centre, Department of Neurodegenerative Disease, UCL Institute of Neurology, London, WC1N 3BG, UK.

[4]Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina, U.S.A.

[5]Departments of Psychiatry, Columbia University Medical Center, New York, NY, U.S.A.

This paper has been submitted for consideration for publication in *Biometrics*

1. Web Appendix A: High-Risk Network vs Low-Risk Network in Motivating Study

Web Figure S1 shows the networks of high-risk and low-risk groups in our motivating study estimated by graphical lasso. The nodes are cortical thickness of 18 regions of interest (ROIs) listed in Web Table S6.

[Figure 1 about here.]

2. Web Appendix B: Detailed Parameter Updates in the Algorithm

ζ_{js} is estimated by

$$\hat{\zeta}_{js} = \frac{S\left(\frac{1}{n}\sum_{i=1}^{n} X_{is}\left\{M_{ij} - \hat{\sigma}^2_{\varepsilon_j}\sum_{i=1}^{n}(\sum_{k\neq j}\hat{\omega}_{jk}(\boldsymbol{X}_i)M_{ik} + \sum_{r\neq s}\hat{\zeta}_{jr}X_{ir})\right\}, \lambda_1\right)}{\left(\frac{1}{n}\sum_{i=1}^{n} X^2_{is}\right)\hat{\sigma}^2_{\varepsilon_j}}, s = 1, ..., q,$$

and $\hat{\alpha}_{jks}$ is estimated by

$$\hat{\alpha}_{jks} = \frac{S\left(\frac{1}{n}\sum_{i=1}^{n} X_{is}\left(2M_{ik}M_{ij} - \hat{\sigma}^2_{\varepsilon_j}A - \hat{\sigma}^2_{\varepsilon_k}B\right), \lambda_1\right)}{(\frac{1}{n}\sum_{i=1}^{n} X^2_{is}M^2_{ik})\hat{\sigma}^2_{\varepsilon_j} + (\frac{1}{n}\sum_{i=1}^{n} X^2_{is}M^2_{ij})\hat{\sigma}^2_{\varepsilon_k}}, s = 1, ..., q$$

where $S(a, b) = \text{sign}(a)(|a| - b)_+$ is the soft thresholding operator,

$A = M_{ik}\left\{\hat{\kappa}_j(\boldsymbol{X}_i) + \sum_{t\neq j,k}\hat{\omega}_{jt}(\boldsymbol{X}_i)M_{it} + (\sum_{r\neq s}\hat{\alpha}_{jkr}X_{ir})M_{ik}\right\}$, and

$B = M_{ij}\left\{\hat{\kappa}_j(\boldsymbol{X}_i) + \sum_{t\neq j,k}\hat{\omega}_{kt}(\boldsymbol{X}_i)M_{it} + (\sum_{r\neq s}\hat{\alpha}_{jkr}X_{ir})M_{ij}\right\}$. The initial values of $\hat{\sigma}^2_{\varepsilon_j}$ are obtained by taking the first partial derivative of $l_n(\boldsymbol{\zeta}, \boldsymbol{\alpha}, \boldsymbol{\sigma}^2)$ with respect to $\sigma^2_{\varepsilon_j}$. The closed-form solution to update $\sigma^2_{\varepsilon_j}$ can be obtained from solving

$$\frac{\partial l_n(\boldsymbol{\zeta}, \boldsymbol{\alpha}, \boldsymbol{\sigma}^2)}{\partial \sigma^2_{\varepsilon_j}} = -\frac{n}{2\sigma^2_{\varepsilon_j}} + \frac{1}{2(\sigma^2_{\varepsilon_j})^2}\sum_{i=1}^{n} M^2_{ij} - \frac{1}{2}\sum_{i=1}^{n}(\kappa_j(\boldsymbol{X}_i) + \sum_{k\neq j}\omega_{jk}(\boldsymbol{X}_i)M_{ik})^2 = 0.$$

3. Web Appendix C: Detailed Simulation Settings

In Settings 1 and 2, we simulated covariates $\boldsymbol{X}_i$ independently from a standard normal distribution. In Setting 1, we set common $\sigma^2_{\varepsilon_j} = \sigma^2_{\varepsilon} = 0.2$ for $j = 1, ..., p$. We assumed that four pairs of biomarkers are connected while the others are conditionally independent. The matrix $(\boldsymbol{\alpha}_{jk})_{j>k}$ is a $q \times \frac{p(p-1)}{2}$ matrix. Each column represents one pair of

biomarkers. We set column 1 of the matrix to be $\{-0.5, -1, -1.5, 0, ..., 0\}^T$, column 5 to be $\{-1, -0.5, -1.5, 0, ..., 0\}^T$, column 8 to be $\{1.5, -0.5, -1, 0, ..., 0\}^T$, column 10 to be $\{-0.5, -1.5, 1, 0, ..., 0\}^T$, and the remaining columns all zeros. Column 1, 5, 8, 10 represent $\boldsymbol{\alpha}_{jk}$ for $(j,k) \in \{(1,2),(2,3),(3,4),(4,5)\}$ when $p = 5$, represent $\boldsymbol{\alpha}_{jk}$ for $(j,k) \in \{(1,2),(1,6),(1,9),(2,3)\}$ when $p = 10$, and represent $\boldsymbol{\alpha}_{jk}$ for $(j,k) \in \{(1,2),(1,6),(1,9),(1,11)\}$ when $p = 20$. The matrix $(\boldsymbol{\zeta}_j)$ is a $q \times p$ matrix. We set $\boldsymbol{\zeta}_1 = \{0.5, 1, 1.5, 0, ..., 0\}^T$, $\boldsymbol{\zeta}_2 = \{1, 0.5, 1.5, 0, ..., 0\}^T$, $\boldsymbol{\zeta}_3 = \{-1.5, 1, 0.5, 0, ..., 0\}^T$ and $\boldsymbol{\zeta}_j = \mathbf{0}$ for $j = 4, ..., p$. We generated outcome from $Y_i = X_{i1} + 2X_{i2} + M_{i1} + 3M_{i2} + W_{i,s_1r_1} + 2W_{i,s_5r_5} + \epsilon_i$, where $\boldsymbol{X}_i$, $\boldsymbol{M}_i$, and $\boldsymbol{W}_i$ are standardized and $\epsilon_i \sim N(0,1)$. W_{i,s_1r_1} is the 1st mutual information where $(s_1, r_1) = (1,2)$ for $p = 5, 10, 20$. W_{i,s_5r_5} is the 5th mutual information where $(s_5, r_5) = (2,3)$ for $p = 5$, and $(s_5, r_5) = (1,6)$ for $p = 10, 20$.

In Setting 2, we set one third of $\sigma^2_{\varepsilon_j}$ to be 0.15, one third to be 0.2, and one third to be 0.25. For example, $\sigma^2_{\varepsilon_j} = 0.15$ for $j \in \{1, 4, 7, 10\}$, $\sigma^2_{\varepsilon_j} = 0.2$ for $j \in \{2, 5, 8\}$ and $\sigma^2_{\varepsilon_j} = 0.25$ for $j \in \{3, 6, 9\}$ when $p = 10$. In matrix $(\boldsymbol{\alpha}_{jk})_{j>k}$, we set column 1 to be $\{-0.5, -1, -1.5, 0, ..., 0\}^T$, column 5 to be $\{1, -0.5, -0.5, 0, ..., 0\}^T$, column 8 to be $\{-1, 1, -1, 0, ..., 0\}^T$, column 10 to be $\{-1, -2, 0.5, 0, ..., 0\}^T$, and the remaining columns all zeros. The settings for $(\boldsymbol{\zeta}_j)$ and outcome are as the same as in Setting 1.

In Setting 3, we included a binary covariate. The second covariate X_{i2} follows a Bernoulli distribution, $Bernoulli(0.5)$ and the other covariates follow standard normal distribution independently. We set heterogeneous $\sigma^2_{\varepsilon_j}$ as the same as in Setting 2. The settings for $(\boldsymbol{\alpha}_{jk})_{j>k}$, $(\boldsymbol{\zeta}_j)$, and outcome are as the same as in Setting 1.

In Setting 4, we considered $(p,q) = (5,3)$, $(p,q) = (10,5)$, and $(p,q) = (20,10)$. The settings for $(\boldsymbol{\alpha}_{jk})_{j>k}$, $(\boldsymbol{\zeta}_j)$, $\sigma^2_{\varepsilon_j}$ and outcome are as the same as Setting 1.

In Setting 5, we considered $n = p = 100$ with $q = 5$. The settings for $(\boldsymbol{\alpha}_{jk})_{j>k}$, $(\boldsymbol{\zeta}_j)$, $\sigma^2_{\varepsilon_j}$ and outcome are as the same as Setting 1.

In Setting 6, $p = 18$, $q = 17$ and we included a constant one in $\boldsymbol{X}_i$ with $n = 500$ and $1,000$. We set eighteen non-zero $\boldsymbol{\alpha}_{jk}$'s: $\boldsymbol{\alpha}_{1,7} = \{-1, 0, -1.5, 0, \ldots, 0\}^T$, $\boldsymbol{\alpha}_{1,15} = \{0, 0, 0, -1, 1, 0, \ldots, 0\}^T$, $\boldsymbol{\alpha}_{3,14} = \{1.5, 0, 0, -1, 0, \ldots, 0\}^T$, $\boldsymbol{\alpha}_{3,18} = \{0, 1.5, -2, 0, \ldots, 0\}^T$, $\boldsymbol{\alpha}_{4,5} = \{0, 0, 0, 0, 0, -1, 0, 1.5, -1, 0, \ldots, 0\}^T$, $\boldsymbol{\alpha}_{4,14} = \{0, 0, 0, 1, 0, -1, 0, \ldots, 0\}^T$, $\boldsymbol{\alpha}_{5,13} = \{1.5, -1, 0, \ldots, 0\}^T$, $\boldsymbol{\alpha}_{5,17} = \{1, 1, -1, 0, \ldots, 0\}^T$, $\boldsymbol{\alpha}_{6,12} = \{0, 0, 0, 0, 0, 0, 0, -1.5, 0, 1.5, 0, \ldots, 0\}^T$, $\boldsymbol{\alpha}_{6,15} = \{0, 0, 0, 0, 0, -1.5, 1.5, 0, \ldots, 0\}^T$, $\boldsymbol{\alpha}_{6,16} = \{1, 0, 0, -1, 0, \ldots, 0\}^T$, $\boldsymbol{\alpha}_{6,18} = \{0, 0, 0, 0, 0, 1.5, 0, 0, 2, 0, \ldots, 0\}^T$, $\boldsymbol{\alpha}_{7,11} = \{1, 1, 0, \ldots, 0\}^T$, $\boldsymbol{\alpha}_{8,17} = \{1, 0, 0, 0, 0, 0, -1, 0, \ldots, 0\}^T$, $\boldsymbol{\alpha}_{9,15} = \{0, 0, 1.5, 0, -2, 0, \ldots, 0\}^T$, $\boldsymbol{\alpha}_{10,12} = \{1.5, 0, -1, 0, \ldots, 0\}^T$, $\boldsymbol{\alpha}_{10,15} = \{0, 0, 0, 0, 0, 0, 1, -1.5, 0, \ldots, 0\}^T$, $\boldsymbol{\alpha}_{11,13} = \{0, 0, 0, 0, 0, 0, -1, 0, 0, 0, 1, 0, \ldots, 0\}^T$.

We set fourteen non-zero $\boldsymbol{\zeta}_j$'s: $\boldsymbol{\zeta}_1 = \{0, 0, 0, 0, 0, 1, 0, \ldots, 0\}^T$, $\boldsymbol{\zeta}_3 = \{0, 0, 0, -1, 0, \ldots, 0\}^T$, $\boldsymbol{\zeta}_4 = \{0, -1, 0, 0, 0, 0, 0, 0, 0, 0, -1, 0, 0, 0, 0, 0, 0, -1\}^T$, $\boldsymbol{\zeta}_6 = \{0, -1, 0, \ldots, 0\}^T$, $\boldsymbol{\zeta}_7 = \{0, \ldots, 0, 1, 0\}^T$, $\boldsymbol{\zeta}_8 = \{0, \ldots, 0, -1.5, 1, 0, 0, 0\}^T$, $\boldsymbol{\zeta}_9 = \{0, \ldots, 0, 1, 0, 0, 0, 0, 0, 0, 0\}^T$, $\boldsymbol{\zeta}_{11} = \{0, 0, 0, 0, 0, 0, 0, 1, 0, \ldots, 0\}^T$, $\boldsymbol{\zeta}_{13} = \{0, \ldots, 0, 1.5, -1\}^T$, $\boldsymbol{\zeta}_{14} = \{0, \ldots, 0, -1\}^T$, $\boldsymbol{\zeta}_{15} = \{0, \ldots, 0, -1, 0, 0, 0, 0\}^T$, $\boldsymbol{\zeta}_{16} = \{0, \ldots, 0, -1, 0, 0, 0, 0, 0, 1, 0\}^T$, $\boldsymbol{\zeta}_{17} = \{0, \ldots, 0, 1, 0, 0\}^T$, $\boldsymbol{\zeta}_{18} = \{0, \ldots, 0, 1, 0, 0, 0, 0\}^T$. σ_j^2's are heterogenous with $\sigma_j^2 = 0.2$ or $\sigma_j^2 = 0.125$. In the second stage outcome model, ten connections, four covariates and two biomarkers are associated with outcome: $Y_i = 2X_{i3} + X_{i,15} + X_{i,16} + X_{i,17} + M_{i,17} + M_{i,18} + 1.5W_{i,1,15} - 1.5W_{i,3,18} + W_{i,4,14} + W_{i,5,13} + 1.5W_{i,5,17} - W_{i,6,18} - W_{i,8,17} - 1.5W_{i,10,12} + W_{i,10,15} - W_{i,11,13}$.

4. Web Appendix D: Additional Simulation Results

Web Figure S2 (Settings 1-4), Web Figure S3 (Setting 5) and Web Figure S4 (Setting 6) visualize the number of times (at least once among 100 simulations) that an edge is identified in the network structure of $\boldsymbol{M}_i$ in the first stage analysis. An edge between M_{ij} and M_{ik} is defined as exist if $\|\boldsymbol{\alpha}_{jk}\|_2 \neq 0$. Our method correctly identifies all TP edges in all settings (black lines; line thickness proportional to the number of times identified; true edges are

the black lines.). When p and q increase, some FP edges are selected with small frequencies (about 2% or 3% times across simulations). In more complicated settings (Settings 2 and 3), more FP edges are selected especially when $p = q = 20$. The performance is improved when the sample size increases to $1,000$. The performance of $p > q$ (Setting 4) is similar to that of Setting 1. More FP edges are selected when $p = 20$ and $n = 500$ in Setting 4 due to the presence of a large number of parameters (up to $2,120$). However, when the sample size increases to $1,000$, most FP edges are eliminated. When $n = p = 100$ (Setting 5, Web Figure S3), our method still correctly identified all TP edges with high frequency and selects some FP edges with small frequency. When the network is denser (Setting 6, Web Figure S4), all TP edges were identified correctly in each simulation and FP edges were selected only 2% times.

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

Larger SSE, fewer TPs, and more FPs appear in Setting 3 (Web Table S1). Our method can still identify almost all TPs and the SSE of $\boldsymbol{\gamma}$ (connection effects) is decreased by 78% when $n = 1,000$ compared to that when $n = 500$. MCC remains large. When $p = q = 20$ with sample size $n = 1,000$, MCC of $(\boldsymbol{\alpha}_{jk})_{j>k}$ can retain close to 0.9 and MCC of $\boldsymbol{\gamma}$ is 0.803. The results suggest that our method performs better when the subject-specific effects are stronger. For large p, increasing sample size would improve the estimation and selection performance. In Setting 4 where $p > q$, Web Table S2 shows a similar performance to that of $p = q$ in terms of estimation and selection accuracy.

[Table 1 about here.]

[Table 2 about here.]

When $n = p = 100$ (Setting 5, Web Table S3), larger SSE and more FPs appear due to a large number of parameters and a small sample size. Our method can still identify almost all TPs in $(\boldsymbol{\alpha}_{jk})_{j>k}$ while most of TNs are not selected, and MCC of $(\boldsymbol{\alpha}_{jk})_{j>k}$ is 0.807. We also identify almost all the TP variables in $\boldsymbol{W}_i$, $\boldsymbol{X}_i$ and $\boldsymbol{M}_i$, and MCC of $\boldsymbol{\gamma}$ retains above 0.85 when the sample size is small.

[Table 3 about here.]

In Setting 6 (Web Table S4) with a lower sparsity rate, our method can still identify almost all TPs with reasonable FPs in both the first and second stages. MCCs of $(\boldsymbol{\alpha}_{jk})_{j>k}$ and $(\boldsymbol{\zeta}_j)$ are above 0.9 and MCC of $\boldsymbol{\gamma}$ is about 0.85.

[Table 4 about here.]

We also compared our method to FGL for Setting 1 with p = q = 20. FGL requires that the subgroups are known. We classified the sample into two groups based on covariate X_1 to implement FGL. Our method outperformed FGL in terms of true positive edges, false positive edges, true negative edges, and false negative edges in the first stage network model (Web Table S5). FGL identified less true edges and selected many more null edges.

[Table 5 about here.]

5. Web Appendix E: Additional Details for Analysis of PREDICT-HD

5.1 *E1: Imaging Measures and Covariates in PREDICT-HD Analysis*

Web Table S6 lists the 6 subcortical ROIs and 18 cortical thickness ROIs used in the analysis.

[Table 6 about here.]

The covariates $\boldsymbol{X}_i$ include 15 baseline clinical variables: CAP score which is a product of CAG and age measuring disease burden (Zhang et al., 2011), four motor scores (bradykinesia, rigidity, dystonia, chorea) from the Unified Huntington's Disease Rating Scale (UHDRS),

total functional capacity (TFC), symbol digit modality test raw score (SDMT), three Stroop scores (color naming, word reading, and interference), four verbal test scores, the University of Pennsylvania smell identification test percentage of correct (smell identification test percent).

5.2 *E2: Modular Classification*

The modularity optimization algorithm was performed on the connection matrix whose elements were the absolute values of the estimated partial correlation. We repeated this algorithm 1,000 times and a consensus clustering procedure was used to obtain a consensus partition (Lancichinetti and Fortunato, 2012). These were implemented in the brain connectivity toolbox (BCT, version 2017-01-15, Rubinov and Sporns, 2010). We set the resolution parameter to represent the classic modularity. Eighteen cortical thickness ROIs were classified into four modules (Figure 4, left panel: module 1 (blue nodes), left and right rostral middle frontal, left and right pars triangularis, and right caudal anterior cingulate; module 2 (orange nodes), right superior frontal, left precuneus, left superior parietal, left and right lateral occipital; module 3 (green nodes), left and right supramarginal, left and right bankssts (i.e., cortical areas around superior temporal sulcus); module 4 (red nodes), left inferior temporal, left fusiform, left and right lingual). From the first-stage analysis, our conditional Gaussian graphical model reveals that the cortical thickness network is dense but organizes in a modular fashion in premanifest HD, so that brain regional cortical thinning acts in a dependent manner.

The cortical "rich club" regions identified in our grey matter structural network are classified to the same module as in the white matter connectivity data (van den Heuvel and Sporns, 2011). Module 2 is almost consistent with cortical "rich club" regions established in healthy human brains, which include the right and left superior frontal, superior parietal, and precuneus regions, and the "rich club" regions are more connected to each other than with other regions (van den Heuvel and Sporns, 2011). Selective loss of "rich club" connectivity is

an organizational principle of white matter connectivity loss in HD (McColgan et al., 2015). In other modules, it can be seen that nodes in the same module fall into similar anatomical regions. These results show that our network modularity is consistent with existing literature when using the absolute value of partial correlation (and also mutual information) as a network connection measure.

5.3 *E3: Identified Top Covariates Associated with Cortical Connections In the First Stage*

Among all covariates, the top three covariates with the largest total effects aggregated across all cortical connections (based on the L_2-norm) are bradykinesia score, smell identification test percent, and Stroop word reading score. The total effects of covariates are summarized in Web Table S7.

[Table 7 about here.]

5.4 *E4: Estimated Cortical Connection Effects from Second Stage*

Web Table S8 summarized the estimated effect sizes of cortical connections identified in the second stage analysis.

[Table 8 about here.]

5.5 *E5: Identified Pathways Through Cortical Connections*

Web Figure S5 shows the identified paths from a baseline covariate to the clinical motor symptom through brain grey matter connections. We focus on four important baseline covariates: CAP (CAG-age product), bradykinesia (measure of motor symptom severity), Stroop word reading (measure of cognition), and smell identification test percent (measure of cognitive state). The paths between a covariate and brain network connections visualize the analyses results in the first stage, which show that different brain connections are selectively related to different covariates. Bradykinesia and smell identification test percent are associated with the largest number of brain connections (10 connections) among the four

covariates. The subcortical connection between hippocampus and thalamus is associated with CAP, bradykinesia, and smell identification test percent but not Stroop word test. The associations between subcortical connection and CAP and baseline bradykinesia are consistent with the clinical consequences of the loss of indirect and direct basal ganglia pathways in HD (hyperkinetic chorea and hypokinetic bradykinesia, respectively). In addition, Stroop word reading test is more associated with intra-hemispheric cortical connections (6 links) than inter-hemispheric cortical connection (only 1 link), while CAP almost equally affects inter-hemispheric and intra-hemispheric cortical connections. As a measure of patient's cognition, the association between baseline Stroop word reading score and cortical connections instead of subcortical connections is consistent with the existing literature (Rosas et al., 2008).

[Figure 5 about here.]

5.6 *E6: Net Reclassification Improvement of Imaging Model*

A measure to quantify the improvement in prediction performance of a new model with additional predictors as compared to an existing model is the net reclassification improvement (NRI) for predicting conversion to HD diagnosis. The continuous NRI statistic (Pencina et al., 2011) is defined as

$$NRI = NRI_e + NRI_{\bar{e}},$$

where NRI_e quantifies the amount of correct reclassification among converters and $NRI_{\bar{e}}$ quantifies that among non-converters. Specifically, $NRI_e = P(\widehat{Y}_{i1} > \widehat{Y}_{i0}|D_i = 1) - P(\widehat{Y}_{i1} < \widehat{Y}_{i0}|D_i = 1)$ and $NRI_{\bar{e}} = P(\widehat{Y}_{i1} < \widehat{Y}_{i0}|D_i = 0) - P(\widehat{Y}_{i1} > \widehat{Y}_{i0}|D_i = 0)$, where D_i is an indicator of HD diagnosis, $\widehat{Y}_{i1}$ and $\widehat{Y}_{i0}$ are the estimated rate of change in motor score based on the new model and the existing model in the test set, respectively. A positive NRI indicates that the new model improves prediction performance compared to the existing model. The average NRI based on the testing data is 0.36 (95% CI: $[0.03, 0.70]$, $p < 0.05$). In addition, the average

NRI_e was 0.24 and the average $NRI_{\bar{e}}$ was 0.12. The imaging biomarker model improves the reclassification rate for converters more than for non-converters. These results demonstrate the gain of predictive power of imaging biomarkers and cortical network connections in addition to standard covariates.

6. Web Appendix F: Additional Details in Validation on TRACK-HD

In Section 4.3, we conducted a validation study of the cortical connections identified in PREDICT-HD using TRACK-HD subjects. The covariates used in re-calibrating the strength of subject-specific cortical connections included baseline CAP, SDMT, TFC, UHDRS diagnostic confidence level, and Stroop word reading score.

7. Web Appendix G: Biological Implications and Insights

We compared the obtained grey matter networks with the white matter connectivity networks obtained on the TRACK-HD (McColgan et al., 2017). Two strongest connections identified in the first stage model are consistent with McColgan et al. (2017), showing that inter-hemispheric white matter connections are particularly vulnerable in premanifest HD. In particular, McColgan et al. (2015) found that degrees (number of brain connections) of the left superior frontal and left rostral middle frontal regions were correlated with TMS. The identified grey matter modulars are also similar to white matter modulars. In our study, among all cortical network connections identified in the first-stage, only a small percentage (17%) are associated with the clinical manifestation of HD as measured by the rate of change in TMS. Long-range and inter-hemispheric links are more likely to associate with motor impairment of HD, similar to the white matter connectivity study in McColgan et al. (2017). Only one connection in the subcortical regional volumetric network was identified as informative of motor symptom. Stroop word reading was identified as one of the top three most important covariates for cortical connections, which is consistent with Stout

et al. (2012). In addition, performance of the baseline Stroop word test is related to cortical connections instead of subcortical connections, similar to the white matter connection study (McColgan et al., 2017).

The strong similarities seen between our grey matter connection study and the white matter connectivity study (McColgan et al., 2017) suggest a shared underlying mechanism of HD. These findings are biologically important since they address a question that is still largely unknown in the HD research community – is white matter loss a direct result of neuronal loss (i.e., caused by the dying back of axons) as opposed to loss of myelin or dysmyelination. The shared pattern between grey matter and white matter connectivity suggests the former.

We note the presence of both positive and negative cortical network connections in the first stage results. These findings are in accordance with several previous studies of cortical thickness co-variation patterns in normal subjects, aging adults, and patients with Alzheimer's disease (Chen et al., 2011; He et al., 2007, 2008). The distribution of the direction of network connections of our results are similar to those of a previous age-related cortical thickness network study (Chen et al., 2011): the number of connections with positive and negative effects on clinical outcome are the same (positive/negative: 11/11), which is consistent with the age-related study in which the number of increased and decreased correlation comparing normal aging network to young network were almost the same (decrease/increase: 17/16). In addition, the negative-effect connection between the right superior frontal and left precuneus regions and the positive-effect connection between the left superior parietal and left precuneus regions were also identified in Chen et al. (2011). Some negative-effect connections are seen in frontal areas, which are also in accordance with Chen et al. (2011).

8. Web Appendix H: Code

Code of the proposed methodology is also available on Github:
https://github.com/shanghongxie/Covariate-adjusted-network.

References

Chen, Z. J., He, Y., Rosa-Neto, P., Gong, G., and Evans, A. C. (2011). Age-related alterations in the modular organization of structural cortical network by using cortical thickness from mri. *Neuroimage* **56,** 235–245.

He, Y., Chen, Z., and Evans, A. (2008). Structural insights into aberrant topological patterns of large-scale cortical networks in alzheimer's disease. *Journal of Neuroscience* **28,** 4756–4766.

He, Y., Chen, Z. J., and Evans, A. C. (2007). Small-world anatomical networks in the human brain revealed by cortical thickness from mri. *Cerebral Cortex* **17,** 2407–2419.

Lancichinetti, A. and Fortunato, S. (2012). Consensus clustering in complex networks. *Scientific Reports* **2,** 336.

McColgan, P., Seunarine, K. K., Gregory, S., Razi, A., Papoutsi, M., Long, J. D., et al. (2017). Topological length of white matter connections predicts their rate of atrophy in premanifest Huntington's disease. *JCI Insight* **2,**.

McColgan, P., Seunarine, K. K., Razi, A., Cole, J. H., Gregory, S., Durr, A., et al. (2015). Selective vulnerability of rich club brain regions is an organizational principle of structural connectivity loss in Huntington's disease. *Brain* **138,** 3327–3344.

Pencina, M. J., D'Agostino, R. B., and Steyerberg, E. W. (2011). Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in Medicine* **30,** 11–21.

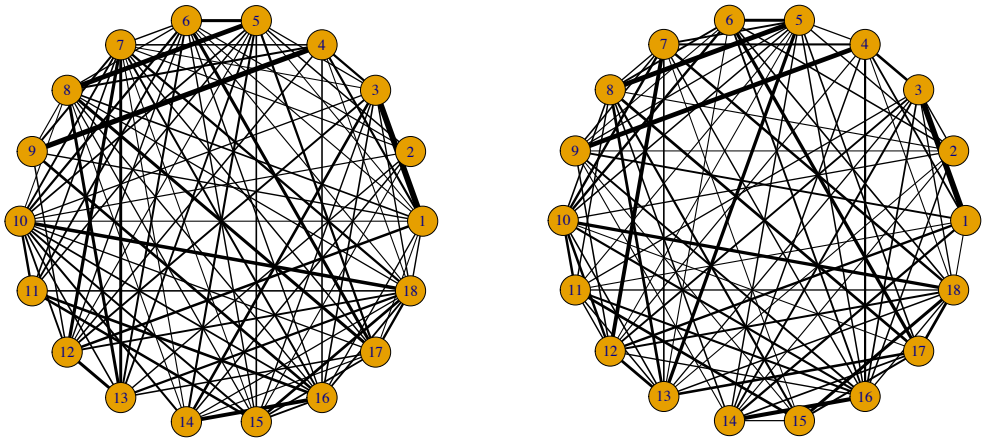
Rosas, H. D., Salat, D. H., Lee, S. Y., Zaleta, A. K., Pappu, V., Fischl, B., et al. (2008). Cerebral cortex and the clinical expression of huntington's disease: complexity and heterogeneity. *Brain* **131,** 1057–1068.

Rubinov, M. and Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* **52,** 1059–1069.

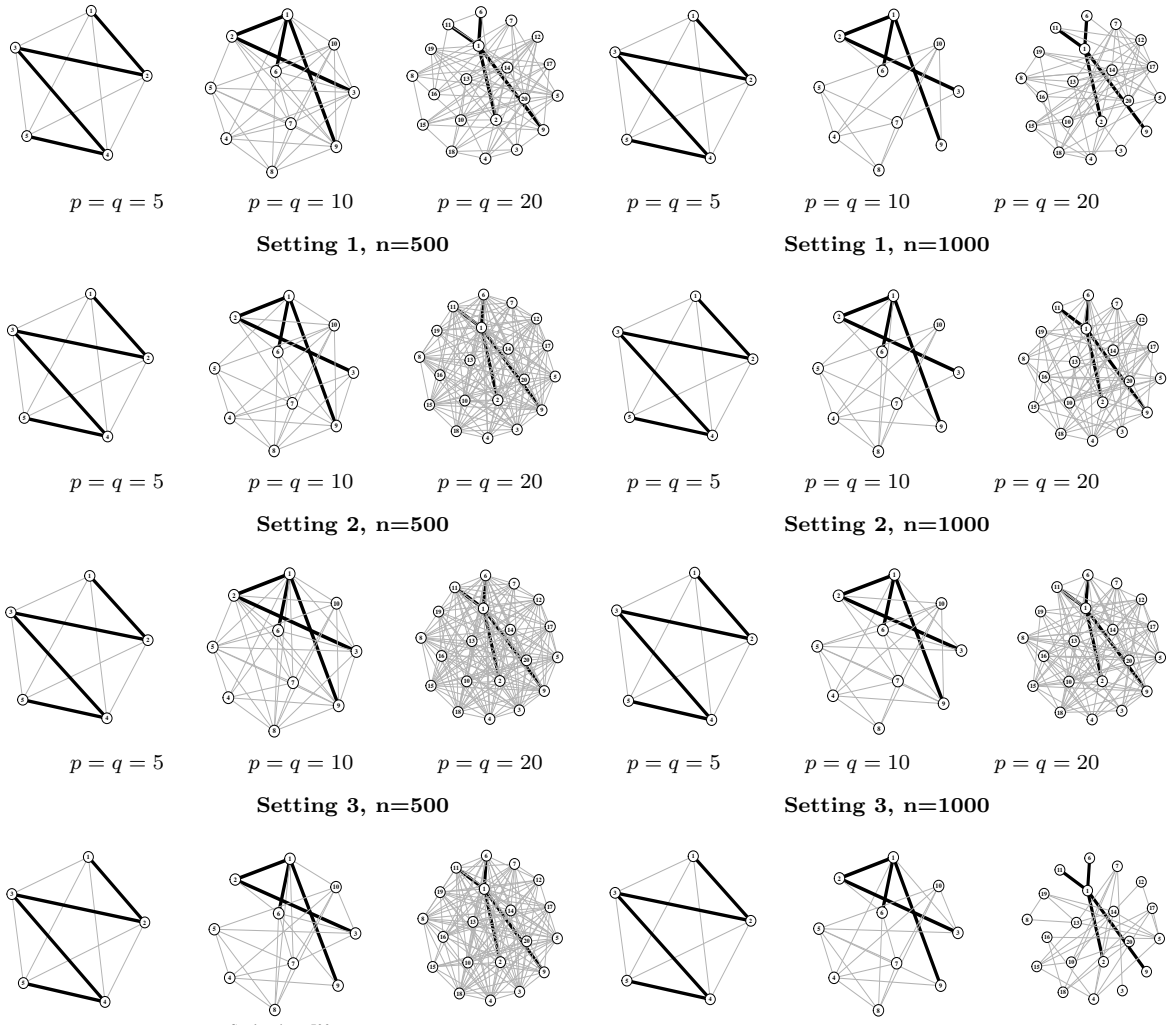
Stout, J. C., Jones, R., Labuschagne, I., O'regan, A. M., Say, M. J., Dumas, E. M., et al. (2012). Evaluation of longitudinal 12 and 24 month cognitive outcomes in premanifest and early huntington's disease. *Journal of Neurology and Neurosurgery Psychiatry* **83,** 687–694.

van den Heuvel, M. P. and Sporns, O. (2011). Rich-club organization of the human connectome. *Journal of Neuroscience* **31,** 15775–15786.

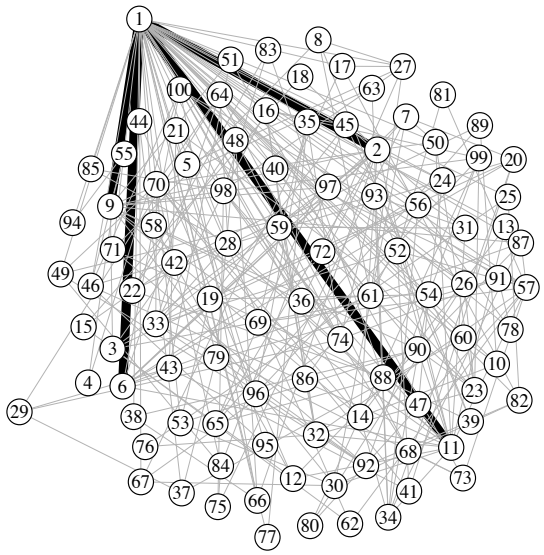
Zhang, Y., Long, J. D., Mills, J. A., Warner, J. H., Lu, W., and Paulsen, J. S. (2011). Indexing disease progression at study entry with individuals at-risk for Huntington disease. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **156,** 751–763.



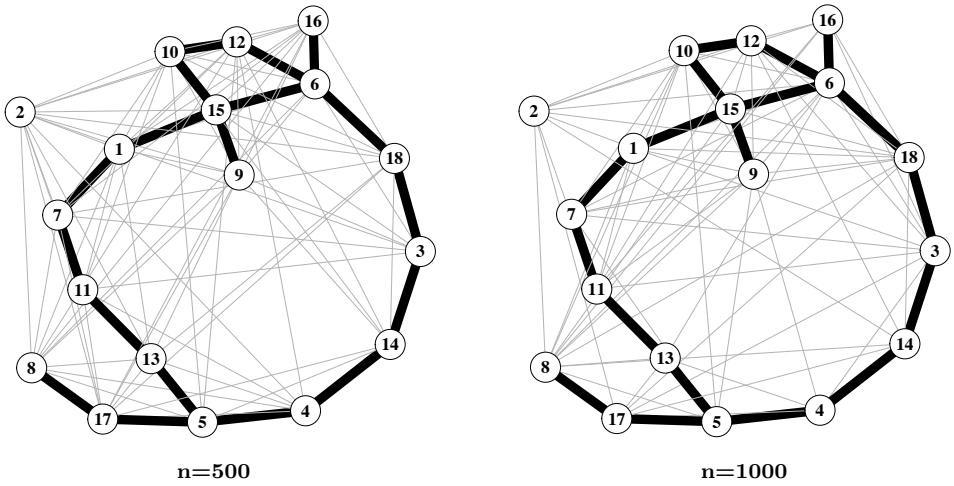
Web Figure S1: High-risk network (left panel) and low-risk network (right panel) estimated by graphical lasso in the motivating study. Node: cortical ROI in Web Table S6. Edge width: proportional to partial correlation.



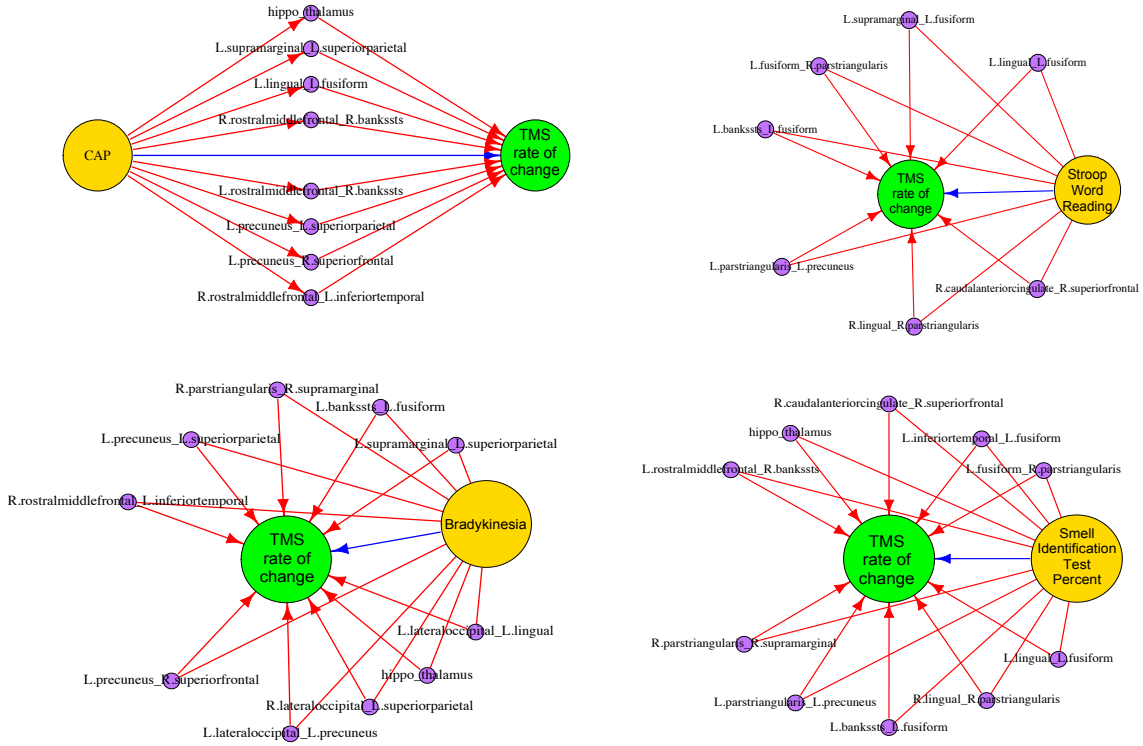
Web Figure S2: Frequency of edges selected in the first stage across 100 simulations in Settings 1-4. Edge width is proportional to the number of times an edge was identified. Black: true positive edges; Grey: false positive edges. Actual edges are the black lines.



Web Figure S3: Frequency of edges selected in the first stage across 100 simulations in Setting 5. Edge width is proportional to the number of times an edge was identified. Black: true positive edges; Grey: false positive edges. Actual edges are the black lines.



Web Figure S4: Frequency of edges selected in the first stage across 100 simulations in Setting 6. Edge width is proportional to the number of times an edge was identified. Black: true positive edges; Grey: false positive edges. Actual edges are the black lines.



Web Figure S5: Identified paths through brain connections. Blue line: covariate effect directly on outcome; Red line: covariate effect through network connections; Orange node: covariate; Purple node: brain connection (intermediate variable); Green node: outcome. Node size is proportional to the degree of the node (for node with directed connection, the degree is the sum of outdegree and indegree).

Web Table S1: **Estimation and Selection Performance of Simulations in Setting 3.**

		$n = 500$						$n = 1000$					
		SSE[1]	**TP**[2]	**FP**[3]	**TN**[4]	**FN**[5]	**MCC**[6]	**SSE**[1]	**TP**[2]	**FP**[3]	**TN**[4]	**FN**[5]	**MCC**[6]
							$p = 5$,	$q = 5$					
1st stage	$\boldsymbol{\zeta}$	0.465	8.95	1.49	14.51	0.05	0.888	0.212	8.99	1.26	14.74	0.01	0.909
	$\boldsymbol{\alpha}$	0.368	12.00	2.57	35.43	0.00	0.900	0.139	12.00	1.65	36.35	0.00	0.930
2nd stage	$\boldsymbol{\gamma}$	0.126	2.00	0.61	7.39	0.00	0.867	0.036	2.00	0.43	7.57	0.00	0.903
	$\boldsymbol{\beta}$	0.026	2.00	0.50	2.50	0.00	0.849	0.008	2.00	0.46	2.54	0.00	0.864
	$\boldsymbol{\eta}$	0.016	2.00	0.49	2.51	0.00	0.855	0.006	2.00	0.47	2.53	0.00	0.852
							$p = 10$,	$q = 10$					
1st stage	$\boldsymbol{\zeta}$	1.091	8.51	2.15	88.85	0.49	0.879	0.489	8.99	1.45	89.55	0.01	0.934
	$\boldsymbol{\alpha}$	0.864	11.73	4.17	433.83	0.27	0.899	0.382	11.99	2.59	435.41	0.01	0.929
2nd stage	$\boldsymbol{\gamma}$	0.583	1.96	1.01	41.99	0.04	0.816	0.119	2.00	0.95	42.05	0.00	0.833
	$\boldsymbol{\beta}$	0.090	2.00	1.55	6.45	0.00	0.735	0.025	2.00	1.09	6.91	0.00	0.802
	$\boldsymbol{\eta}$	0.053	2.00	1.48	6.52	0.00	0.746	0.014	2.00	1.14	6.86	0.00	0.786
							$p = 20$,	$q = 20$					
1st stage	$\boldsymbol{\zeta}$	2.168	8.08	4.36	386.64	0.92	0.815	1.084	8.88	4.76	386.24	0.12	0.855
	$\boldsymbol{\alpha}$	1.833	11.50	7.30	3780.70	0.50	0.854	0.828	11.99	4.77	3783.23	0.01	0.888
2nd stage	$\boldsymbol{\gamma}$	0.889	1.92	1.45	186.55	0.08	0.746	0.200	2.00	1.21	186.79	0.00	0.803
	$\boldsymbol{\beta}$	0.250	2.00	2.92	15.08	0.00	0.645	0.063	2.00	2.79	15.21	0.00	0.674
	$\boldsymbol{\eta}$	0.045	2.00	3.13	14.87	0.00	0.643	0.012	2.00	2.23	15.77	0.00	0.724

[1]SSE: Average sum of squared error across 100 simulations; [2]TP: Average number of true positive parameters across 100 simulations; [3]FP: Average number of false positive parameters across 100 simulations; [4] TN: Average number of true negative parameters across 100 simulations; [5] FN: Average number of false negative parameters across 100 simulations; [6] MCC: Matthews correlation coefficient

Web Table S2: **Estimation and Selection Performance of Simulations in Setting 4.**

		$n = 500$						$n = 1000$					
		SSE[1]	**TP**[2]	**FP**[3]	**TN**[4]	**FN**[5]	**MCC**[6]	**SSE**[1]	**TP**[2]	**FP**[3]	**TN**[4]	**FN**[5]	**MCC**[6]
							$p = 5,\ q = 3$						
1st stage	$\boldsymbol{\zeta}$	0.269	8.99	0.57	5.43	0.01	0.931	0.111	9.00	0.44	5.56	0.00	0.942
	$\boldsymbol{\alpha}$	0.148	12.00	2.36	15.64	0.00	0.882	0.052	12.00	1.95	16.05	0.00	0.894
2nd stage	$\boldsymbol{\gamma}$	0.043	2.00	0.26	7.74	0.00	0.939	0.016	2.00	0.29	7.71	0.00	0.933
	$\boldsymbol{\beta}$	0.006	2.00	0.14	0.86	0.00	1.000	0.003	2.00	0.15	0.85	0.00	1.000
	$\boldsymbol{\eta}$	0.020	2.00	0.44	2.56	0.00	0.878	0.009	2.00	0.47	2.53	0.00	0.859
							$p = 10,\ q = 5$						
1st stage	$\boldsymbol{\zeta}$	0.528	8.95	1.78	39.22	0.05	0.914	0.270	9.00	1.30	39.70	0.00	0.936
	$\boldsymbol{\alpha}$	0.368	12.00	2.03	210.97	0.00	0.934	0.218	12.00	2.16	210.84	0.00	0.937
2nd stage	$\boldsymbol{\gamma}$	0.083	2.00	0.49	42.51	0.00	0.917	0.025	2.00	0.30	42.70	0.00	0.948
	$\boldsymbol{\beta}$	0.008	2.00	0.44	2.56	0.00	0.868	0.003	2.00	0.22	2.78	0.00	0.928
	$\boldsymbol{\eta}$	0.041	2.00	1.34	6.66	0.00	0.763	0.010	2.00	1.01	6.99	0.00	0.813
							$p = 20,\ q = 10$						
1st stage	$\boldsymbol{\zeta}$	1.654	8.76	3.70	187.30	0.24	0.863	0.440	8.99	2.51	188.49	0.01	0.913
	$\boldsymbol{\alpha}$	1.133	12.00	5.05	1882.95	0.00	0.905	0.245	12.00	1.35	1886.65	0.00	0.964
2nd stage	$\boldsymbol{\gamma}$	0.061	2.00	0.4	187.60	0.00	0.930	0.022	2.00	0.25	187.75	0.00	0.955
	$\boldsymbol{\beta}$	0.009	2.00	0.96	7.04	0.00	0.816	0.005	2.00	0.95	7.05	0.00	0.821
	$\boldsymbol{\eta}$	0.024	2.00	2.26	15.74	0.00	0.720	0.013	2.00	1.98	16.02	0.00	0.742

[1]SSE: Average sum of squared error across 100 simulations; [2]TP: Average number of true positive parameters across 100 simulations; [3]FP: Average number of false positive parameters across 100 simulations; [4] TN: Average number of true negative parameters across 100 simulations; [5] FN: Average number of false negative parameters across 100 simulations; [6] MCC: Matthews correlation coefficient

Web Table S3: **Estimation and Selection Performance of Simulations in Setting 5 with $n = p = 100$.**

		$n = p = 100, \quad q = 5$					
		SSE[1]	**TP**[2]	**FP**[3]	**TN**[4]	**FN**[5]	**MCC**[6]
1st stage	$\boldsymbol{\zeta}$	8.009	3.90	2.38	488.62	5.10	0.543
	$\boldsymbol{\alpha}$	5.196	11.18	7.50	24730.5	0.82	0.807
2nd stage	$\boldsymbol{\gamma}$	1.114	1.73	0.32	4945.68	0.27	0.866
	$\boldsymbol{\beta}$	0.130	2.00	0.25	2.75	0.00	0.921
	$\boldsymbol{\eta}$	0.631	1.96	3.87	94.13	0.04	0.697

[1]SSE: Average sum of squared error across 100 simulations; [2]TP: Average number of true positive parameters across 100 simulations; [3]FP: Average number of false positive parameters across 100 simulations; [4] TN: Average number of true negative parameters across 100 simulations; [5] FN: Average number of false negative parameters across 100 simulations; [6] MCC: Matthews correlation coefficien

Web Table S4: **Estimation and Selection Performance of Simulations in Setting 6.**

		$n = 500$						$n = 1000$					
		SSE[1]	**TP**[2]	**FP**[3]	**TN**[4]	**FN**[5]	**MCC**[6]	**SSE**[1]	**TP**[2]	**FP**[3]	**TN**[4]	**FN**[5]	**MCC**[6]
							$p = 18,$	$q = 17$					
1st stage	$\boldsymbol{\zeta}$	2.402	19.00	3.54	301.46	0.00	0.937	1.123	19.00	1.90	303.10	0.00	0.961
	$\boldsymbol{\alpha}$	6.491	37.92	9.46	2706.54	0.08	0.910	2.762	38.00	10.74	2705.26	0.00	0.908
2nd stage	$\boldsymbol{\gamma}$	2.792	9.63	3.24	139.76	0.37	0.842	1.552	9.90	3.15	139.85	0.10	0.862
	$\boldsymbol{\beta}$	2.441	4.00	3.55	9.45	0.00	0.649	1.837	4.00	3.36	9.64	0.00	0.667
	$\boldsymbol{\eta}$	0.044	2.00	2.89	13.11	0.00	0.665	0.016	2.00	3.04	12.96	0.00	0.660

[1]SSE: Average sum of squared error across 100 simulations; [2]TP: Average number of true positive parameters across 100 simulations; [3]FP: Average number of false positive parameters across 100 simulations; [4] TN: Average number of true negative parameters across 100 simulations; [5] FN: Average number of false negative parameters across 100 simulations; [6] MCC: Matthews correlation coefficient

Web Table S5: **Comparison with FGL in Simulation Setting 1 with $p = q = 20$.**

	Proposed Method		**FGL**	
	$N = 500$	$N = 1000$	$N = 500$	$N = 1000$
TP	4	4	3.34	3.64
FP	0.99	0.93	64.86	93.41
TN	185.01	185.07	121.14	92.59
FN	0	0	0.66	0.36

Web Table S6: Subcortical and cortical ROIs included in PREDICT-HD study analyses

6 Subcortical ROIs	
caudate	putamen
accumbens	globus pallidus
hippocampus	thalamus

18 Cortical ROIs	
1. Lateral occipital.L	2. Caudal anterior cingulate.R
3. Lateral occipital.R	4. Lingual.R
5. Rostral middle frontal.L	6. Pars triangularis.L
7. Precuneus.L	8. Rostral middle frontal.R
9. Lingual.L	10. Supramarginal.L
11. Bankssts.L	12. Superior parietal.L
13. Superior frontal.R	14. Inferior temporal.L
15. Bankssts.R	16. Fusiform.R
17. Pars triangularis.R	18. Supramarginal.R

Web Table S7: Estimated total effect (L_2 norm) of covariates on network connections in the first stage gray matter subcortical and cortical network model (PREDICT-HD study)

Covariate	Subcortical[1]	Cortical[1]
intercept	4.432	6.710
CAP	0.262	1.743
Bradykinesia	0.323	1.914
Rigidity	0.399	1.648
Dystonia	0.234	0.806
Chorea	0.295	1.471
TFC	0.206	1.216
SDMT	0.316	0.926
Stroop color naming	0.457	1.409
Stroop word reading	0.483	1.858
Stroop interference	0.122	1.614
Dual verbal working memory total	0.883	1.749
Hopkins verbal learning test delayed recall	0.400	1.772
Verbal fluency total correct	0.327	1.844
Hopkins verbal learning test total	0.417	1.289
Smell identification test percent	0.505	1.907
Gray matter volume total	0.288	1.816
White matter volume total	0.426	1.766

[1]: L_2 norm of all the estimated connection coefficients for each covariate.

Web Table S8: 22 identified cortical connections in PREDICT-HD study.

Connection	Estimated effect size [1]
Lateral occipital.L/Precuneus.L	0.108
Lateral occipital.L/Lingual.L	-0.027
Caudal anterior cingulate.R/Superior frontal.R	-0.082
Lateral occipital.R/Superior parietal.L	0.055
Lingual.R/Pars triangularis.R	0.111
Rostral middle frontal.L/Bankssts.R	0.059
Pars triangularis.L/Precuneus.L	0.015
Precuneus.L/Superior parietal.L	0.002
Precuneus.L/Superior frontal.R	-0.054
Precuneus.L/Supramarginal.R	-0.034
Rostral middle frontal.R/Inferior temporal.L	0.135
Rostral middle frontal.R/Bankssts.R	-0.015
Rostral middle frontal.R/Pars triangularis.R	-0.032
Lingual.L/Fusiform.L	-0.074
Supramarginal.L/Superior parietal.L	0.039
Supramarginal.L/Fusiform.L	-0.122
Bankssts.L/Fusiform.L	-0.149
Inferior temporal.L/Fusiform.L	0.033
Bankssts.R/Pars triangularis.R	-0.080
Bankssts.R/Supramarginal.R	0.245
Fusiform.R/Pars triangularis.R	0.174
Pars triangularis.R/Supramarginal.R	-0.002

[1]Estimated effect size: Estimated coefficient in the refitted linear regression model (second stage outcome model).