Sparse Representation based Biomarker Identification for Schizophrenia

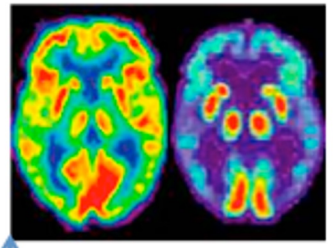
DATA INTEGRATION OF IMAGING AND GWAS DATA

Yin Yao
National Institutes of Mental Health, NIH,
Bethesda, 20892, USA

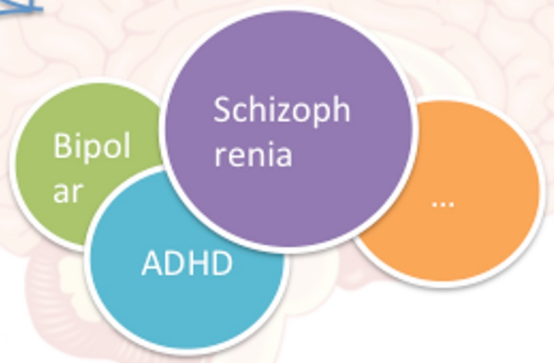
Multiple layers of networks (correlations) in neuropsychiatric disorders

NGS

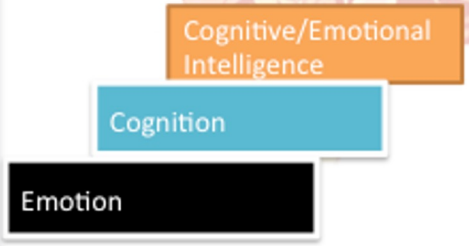
Neuroimaging: structure/ function



Psychiatric disorders



The developmental brain

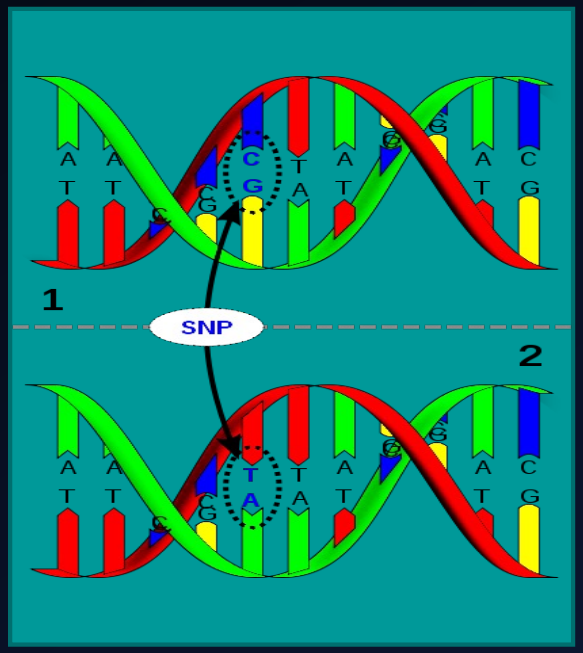


Background

- **Schizophrenia (SCZ)**
  - One of the most chronically disabling psychiatric illnesses.
  - Global median lifetime morbid risk of 7.2/1000 persons.
  - Both genomic (e.g. GWAS) and brain imaging data (e.g. fMRI) were used to explore the pathogenesis of SCZ.
- **Data Integration**
  - Take advantage of complementary information.
  - Seek higher power to identify potential biomarkers that might be missed by using a single type of analysis.

Genomic Data

- A genome-wide association study (**GWAS**) has been a way to explore potential effects in human diseases.
- GWAS typically focus on association(s) between single nucleotide polymorphisms (**SNPs**) and human diseases.
- GWAS has been viewed as a powerful method for identifying susceptible genes for many common diseases.



Limitations of GWAS

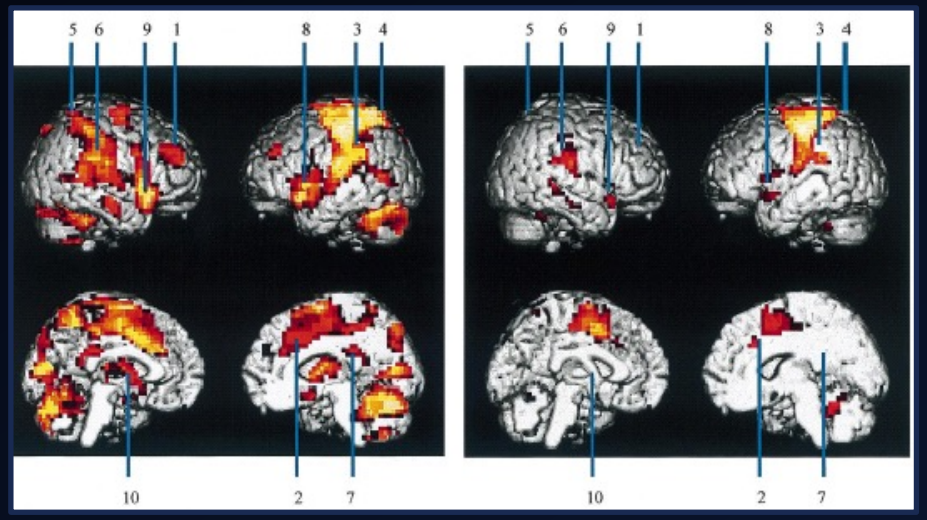
- **Lack of information**
  - single SNP has very small effect on a disease.
- **High rates of false-positive**
  - Many “associated” variants are not causal, need large sample size.

- Investigators typically search the entire genome for associations.

fMRI Imaging Data: Neurological Study on Schizophrenia

- Functional Magnetic Resonance Imaging (**fMRI**)
  - Task-fMRI (**tfMRI**)
  - Resting state fMRI (**rsfMRI**)

- Identify both structural and functional abnormalities in brain regions



Locating functional differences in the frontal lobes, hippocampus and temporal lobes in brain of Schizophrenia patients

Proposed Regression Model for integrative data analysis

The representation of combined data set

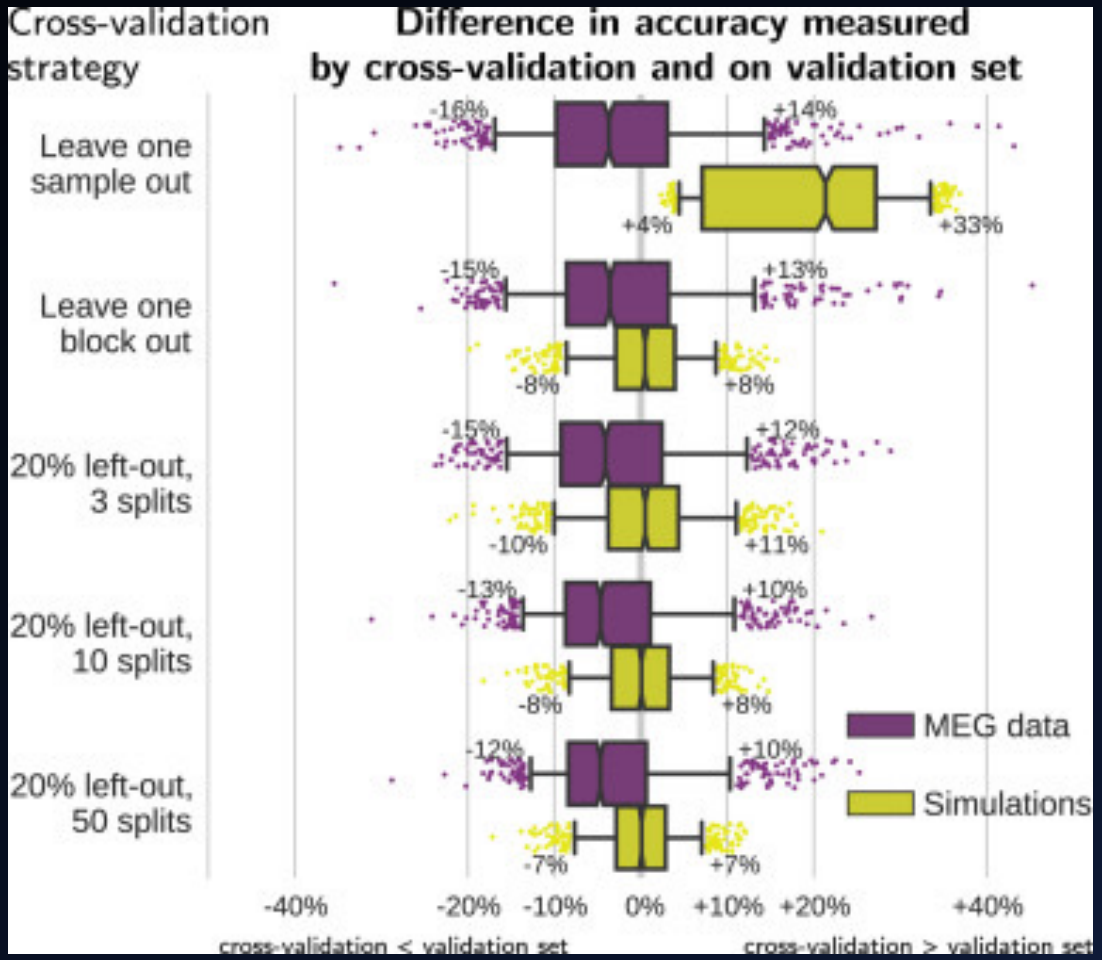
$$\mathbf{y} = [\boldsymbol{\alpha_1 A_1}, \boldsymbol{\alpha_2 A_2}] \begin{bmatrix} \boldsymbol{x_1} \\ \boldsymbol{x_2} \end{bmatrix} + \boldsymbol{\varepsilon} = \boldsymbol{Ax} + \boldsymbol{\varepsilon} \quad \text{(Eq. 2)}$$

- where, $y \in R^{n\times 1}$ is the observation vector;
- $\mathrm{A}_1 \in R^{n\times p_1}$ and $\mathrm{A}_2 \in R^{n\times p_2}$ are measurements of two different data types $\mathrm{X} = [\alpha_1 \mathrm{A}_1, \alpha_2 \mathrm{A}_2] \in R^{n\times p}$
  - where $\alpha_1 + \alpha_2 = 1$, and $\alpha_1, \alpha_2 > 0$ are the weight factors for the two types of data.
- $\varepsilon \in R^{n\times 1}$ is the measurement error caused by noises.

Weighting factors and classification

- $\mathbf{y} = [\boldsymbol{\alpha_1 A_1}, \boldsymbol{\alpha_2 A_2}] \begin{bmatrix} \boldsymbol{x_1} \\ \boldsymbol{x_2} \end{bmatrix} + \boldsymbol{\varepsilon} = \boldsymbol{Ax} + \boldsymbol{\varepsilon}$ (Eq. 2)
- To determine optimal weighting factors α1 and α2, cross validation can be used
  - So that the weighting factors that generates the best classification ratio (CR). Cross validation is used to generate the best classification ratio (CR).
- In each run of the 10-fold cross-validation, 90 percent subjects from both cases and controls were randomly chosen for variable/biomarker selection, while the rest were used for testing. For each method, we carried out 100 runs and the average of the classification ratios was used as the final identification accuracy.

Cross-Validation



- Varoquaux et al., Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines, NeuroImage, Volume 145, 2017, Pages 166-179

Weighting factors and classification

- $\mathbf{y} = [\boldsymbol{\alpha_1 A_1}, \boldsymbol{\alpha_2 A_2}] \begin{bmatrix} \boldsymbol{x_1} \\ \boldsymbol{x_2} \end{bmatrix} + \boldsymbol{\varepsilon} = \boldsymbol{Ax} + \boldsymbol{\varepsilon}$ (Eq. 2)
- We also used the cross-validation to determine the optimal weighting factors in Eq. (2).
  - For different pair of weighting factors, different variable groups will be selected, resulting in different classification ratios.
- Using **cross validation**, we can select the best weighting factors that lead to the highest classification ratio.

Solve the Model: Sparse Regression

Multivariate regression

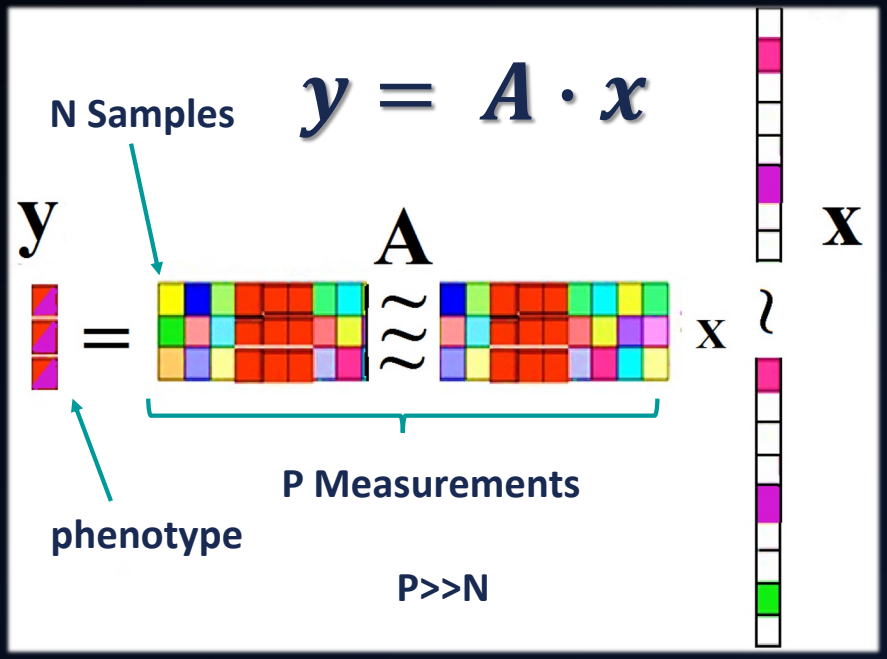


Fig. Schematic diagram of a sparse regression model

SRVS Algorithm

1. Initial $\delta^{(0)} = 0$;
2. For the Step l, randomly choose k columns from $X = \{x_1, \ldots, x_p\} \in R^{n \times p}$ to construct a $n \times k$ sub-matrix denoted as $X_l \in R^{n \times k}$; and mark the selected columns' indexes as $I_l \in R^{1 \times k}$;

$\boldsymbol{P >> N}$

SRVS ALGORITHM

Sparse Representation Variable Selection

SRVS ALGORITHM

$P >> N$

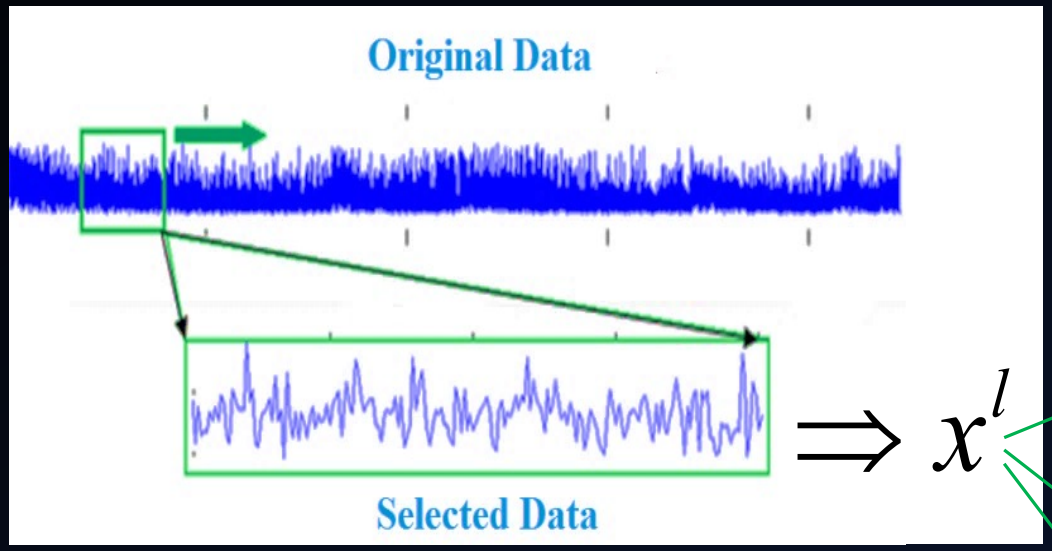
Sparse Representation Variable Selection Algorithm

3. Solve the following L_p minimization problem to find the optimal sparse solution $\delta_l \in R^{k\times 1}$:

$$min\|\delta_l\|_p \quad \text{s.t.} \quad \|y - X_l\delta_l\|_2 \leq \varepsilon \tag{3}$$

4. Update $\delta^{(l)} \in R^{p\times 1}$ with δ_l: $\delta^{(l)}(I_l) = \delta^{(l-1)}(I_l)+\delta_l$; where $\delta^{(l)}(I_l)$ and $\delta^{(l-1)}(I_l)$ denote the I_l th entries in $\delta^{(l)}$ and $\delta^{(l-1)}$ respectively;
5. If $\left\|\delta^{(l)}/l - \delta^{(l-1)}/(l-1)\right\|_2 > \alpha$, where α is a predefined constant, update $l = l+1$, and go to Step 2. Otherwise, set $\delta = \delta^{(l)}/l$. The non-zero entries in δ correspond to the column vectors selected.

Finding X using SRVS



$\Rightarrow x^l$

$$X = \frac{1}{L}\sum x^l$$

Sparse Representation Variable Selection

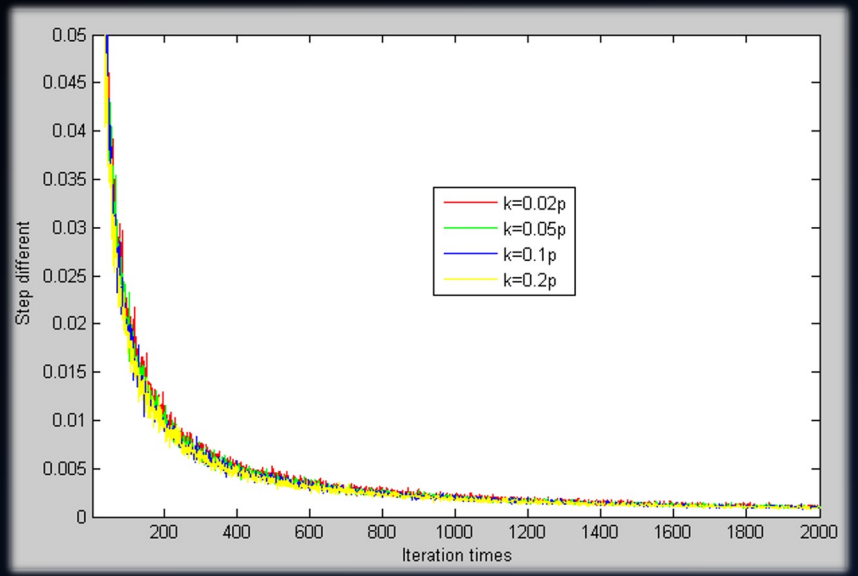
Properties of SRVS

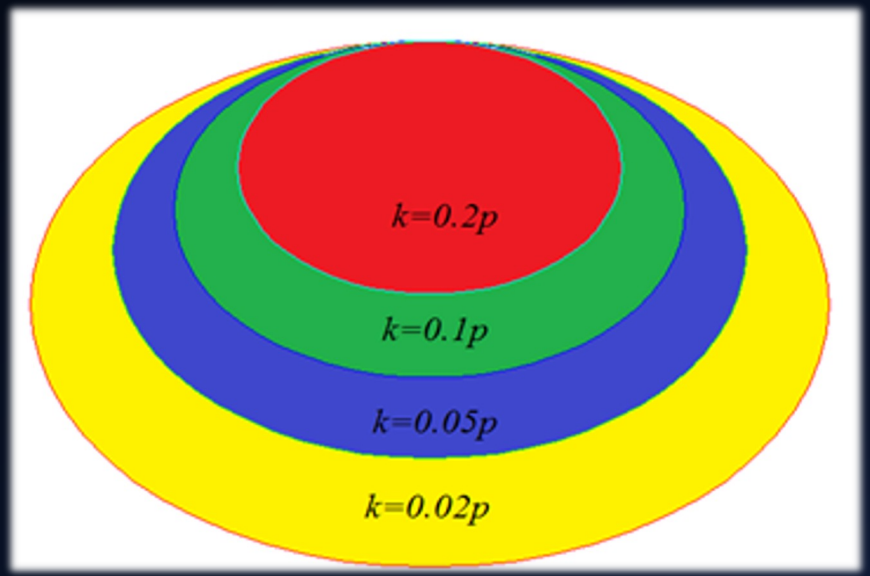
SRVS Algorithm

1. Initial $\delta^{(0)} = 0$;
2. For the Step l, randomly choose k columns from $X = \{x_1, \ldots, x_p\} \in R^{n \times p}$ to construct a $n \times k$ sub-matrix denoted as $X_l \in R^{n \times k}$; and mark the selected columns' indexes as $I_l \in R^{1 \times k}$;

CONVERGENCE $\|y - Ax\|_2 \leq \varepsilon$

MULTI-SCALE PROPERTY





Sparse Representation Variable Selection

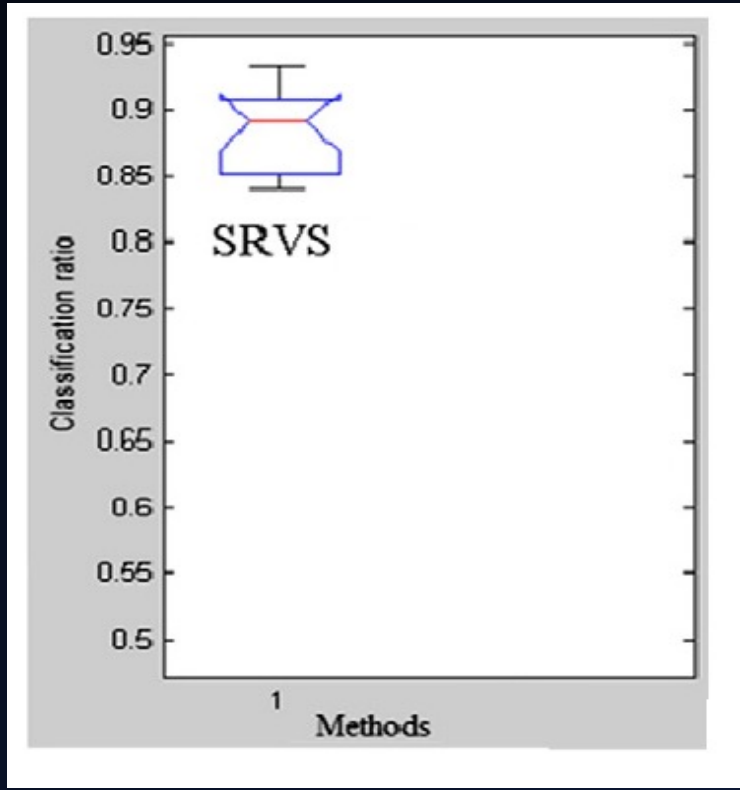
Application 1: GWAS data and tfMRI data integration

- **Simulated GWAS Data**
  - 92 cases, 116 controls, with 759,075 SNPs
  - Use SNPs as features
- **tfMRI data**
  - The sample size is the same as the GWAS data
  - Stimulus-on vs. stimulus-off images were collected from both cases and controls
  - A total of 153,594 fMRI voxels features available

(Cao et al., 2014, NeuroImage)

Application 1: Results

- Classification ratio (**CR**) from Cross validation (%)
  - SNPs alone: 83.1
  - tfMRI voxels alone: 63.1
  - Integration features by SRVS: 89.7



(Cao et al., 2014, NeuroImage)

Application 2: GWAS data and rfMRI data integration

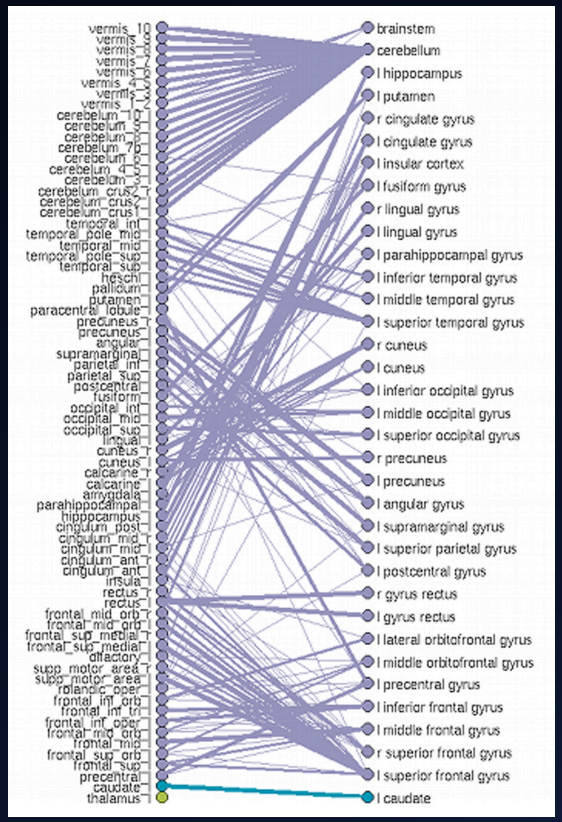
- **Simulated GWAS Data**
  - 100 cases vs. 100 controls, with 10,000 SNP data
- **rfMRI data**
  - 100 cases vs. 100 controls
  - Feature: Connectivity between/within 116 AAL brain regions
  - $\frac{116 \cdot 117}{2} = 6786$ features

Application 2: Feature extraction using rfMRI data

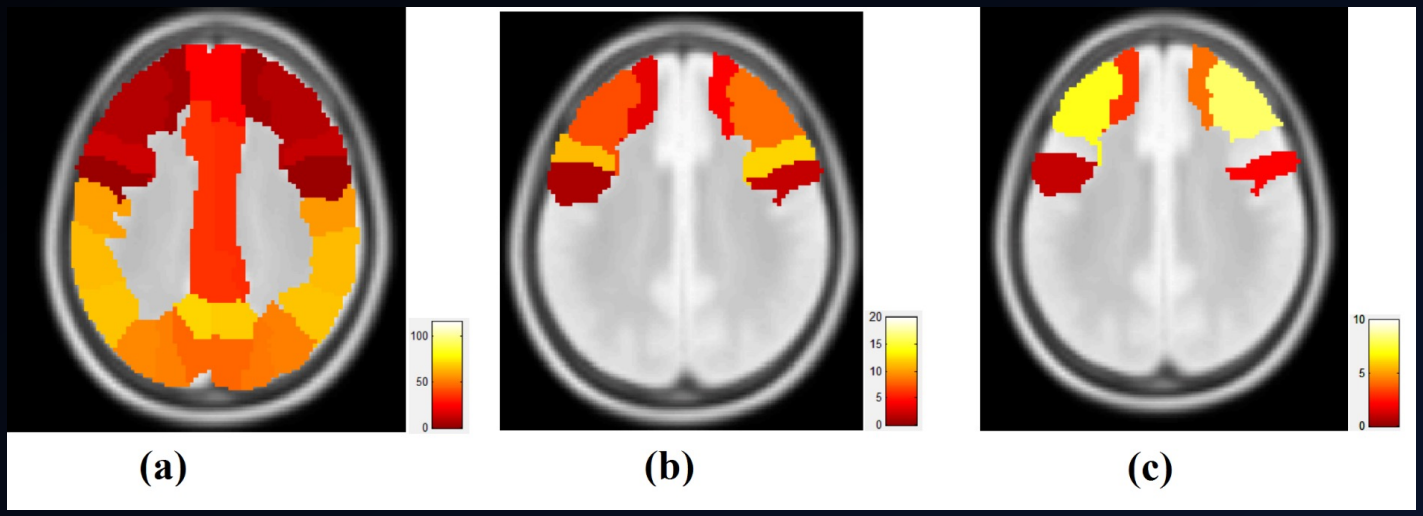
- **Features from rfMRI**: Connectivity (**CN**) between/within 116 AAL brain regions
- CN =sum(corr(V_i, V_j))/N
  - where V_i and V_j represent the intensity of ith and jth voxel;
  - corr(): Pearson correlation coefficients;
  - N is the number of voxel pairs
- Finally,

$$\frac{116(116+1)}{2}$$ (6786 features)

unique CNs are extracted

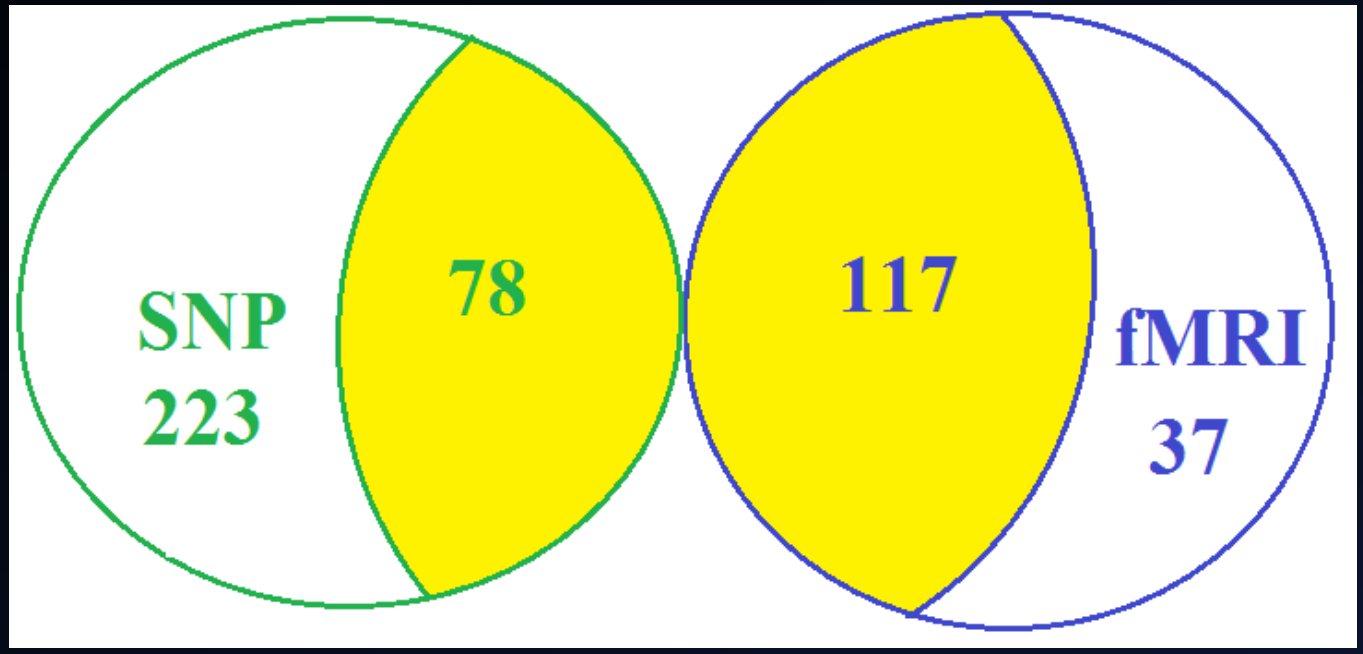


Application 2: Results 1



(a) 116 AAL brain regions; (b) using rfMRI data alone selected brain regions; (c) brain regions selected from integrated biomarker selection with SRVS

Application 2: Results 2

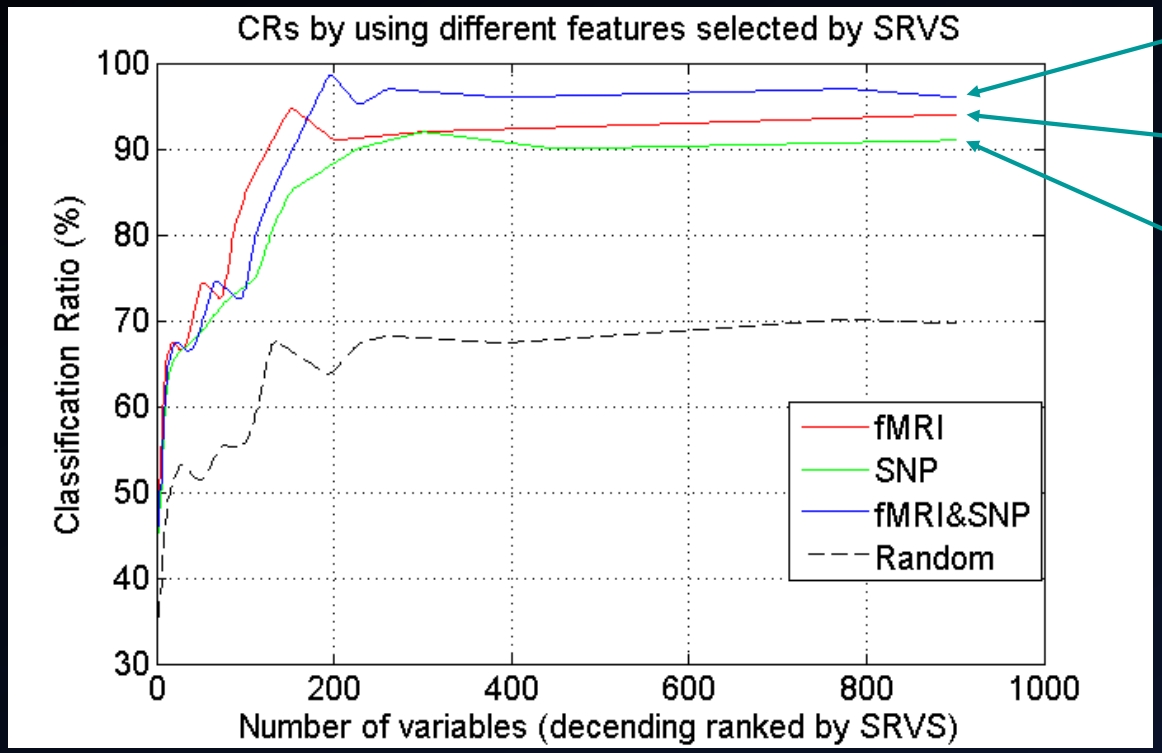


Using SNP feature alone:
301 SNP selected; CR=91.7%

Using rfMRI feature alone:
154 fMRI selected; CR=94.8%

Using both SNP & rfMRI features:
195 fMRI selected; CR=98.9%

Application 2: Results 3

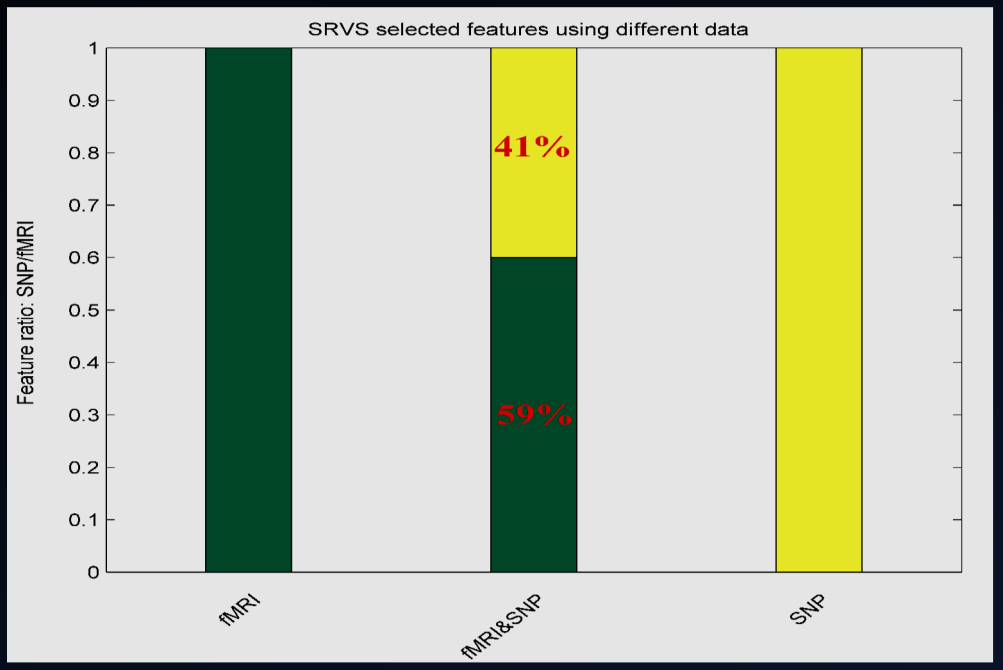


Using both SNP & rfMRI features:
195 rfMRI selected; CR=98.9%

Using rfMRI feature alone:
154 rfMRI selected; CR=94.8%

Using SNP feature alone:
301 SNP selected; CR=91.7%

Application 2: Results 4



Using SNP feature alone:
301 SNP selected; CR=91.7%

Using rfMRI feature alone:
154 fMRI selected; CR=94.8%

Using both SNP & rfMRI features:
195 fMRI selected; CR=98.9%

Summary

- Integrating both **fMRI** data and **SNPs** seems to point to better accuracy for SCZ diagnosis.
- Both **rfMRI** and **tfMRI** can be integrated with SNPs.
- The Sparse-Representation-Variable-Selection method is effective in selecting biomarkers when the number of variables is large and the sample size is small.

To Conclude:

- We addressed the data integration problem by developing a generalized sparse model (GSM) using weighting factors (α_1 and α_2) to integrate multi-modality data for “empirical predictor” selection

- More applied projects are on the way

Acknowledgement

- Hong Bao Cao conducted the data analysis
- Yin Yao and Yu-Ping Wang conceived the concept

(Cao et al., 2014, NeuroImage)

Thanks!

QUESTIONS?