Similarity Data based TBI Patients Stratification and Feature Selection

[1]Emily Li, [1]Qi Cheng, [1]Alan D. Kaplan, [2]Geoffrey Manley

[1]Lawrence Livermore National Laboratory
[2]TRACK-TBI Consortium

May 15, 2019

LLNL-PRES-774228

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

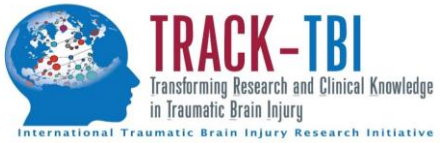
Lawrence Livermore National Laboratory

Introduction

- Traumatic brain injury (TBI) is a complex disorder. Though TBI is a major cause of death and disability, it is traditionally stratified based on clinical signs and symptoms, and few targeted treatments exist.
- The Transforming Research and Clinical Knowledge in Traumatic Brain Injury (TRACK-TBI) Pilot multicenter study enrolled 586 acute TBI patients and collected diverse common data elements (TBI-CDEs) across the study population, including imaging, genetics, and clinical outcomes.
  - The most highly-detailed ever collected data in neuroscience, with enormous opportunity for data-rich phenotyping and TBI precision medicine
- The goals of the project include developing better ways to diagnose and provide early prognosis of brain injury and to characterize various subtypes of brain injury that might have different underlying pathophysiology, prognosis, and optimal treatment.
- We conduct multidimensional exploratory analytics and use similarity data based approaches to reveal data-driven patterns in patient outcomes, and identify the key features that characterize these patterns.
  - Provide for diagnostic decision support and precision medicine in neurotrauma

Components of the TRACK-TBI Dataset and Data-Analysis Challenges

- **Data captured across 4 broad domains:**



- **Clinical assessments and demographic information,**
- **Blood biomarkers (genetics and proteomics)**
- **Neuroimaging (Head CT and MRI)**
- **Outcome measures at 3, 6, 12 months**

Missing values

Heterogeneity in data
- Mixed categorical, ordinal, continuous

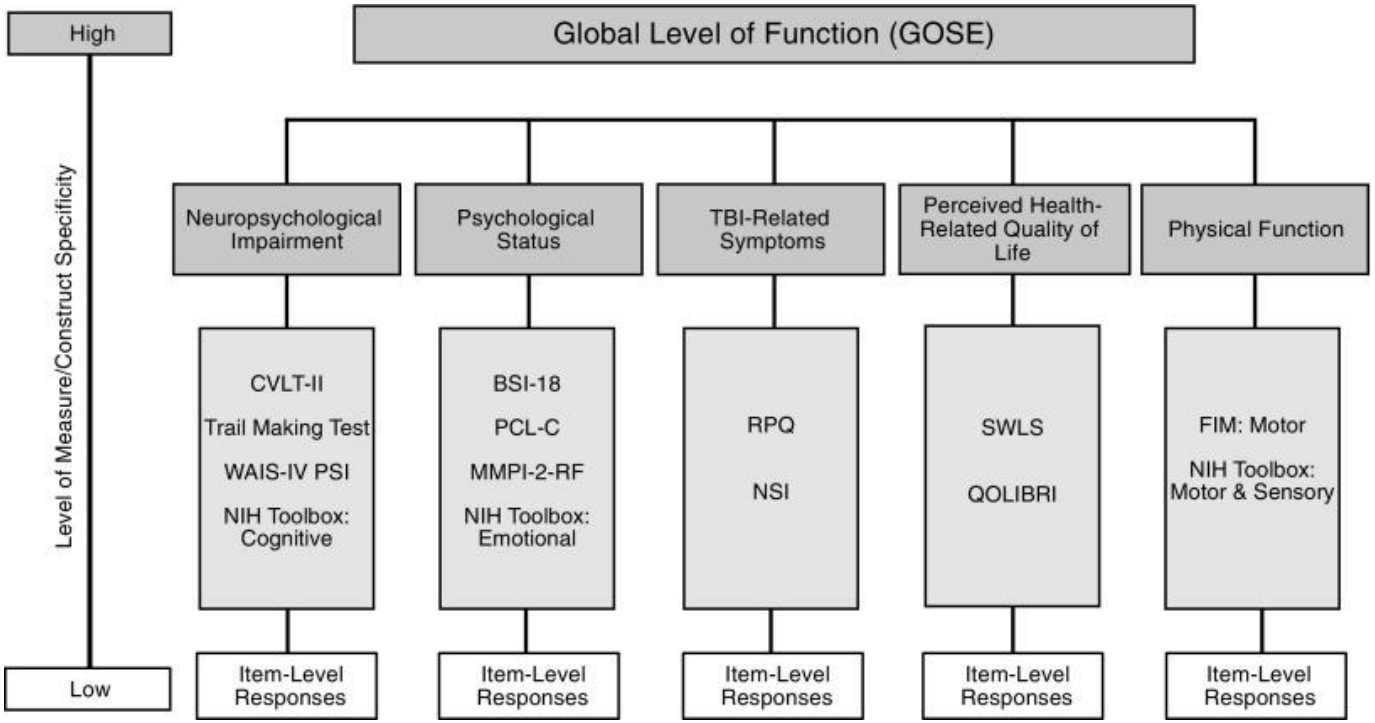
Multimodality
- Scalar and various imaging types

Multidimensional targets (outcomes)

Data Structure

- Primarily working with the 6-month outcome data
- 6-month outcome data breakdown by functional groups

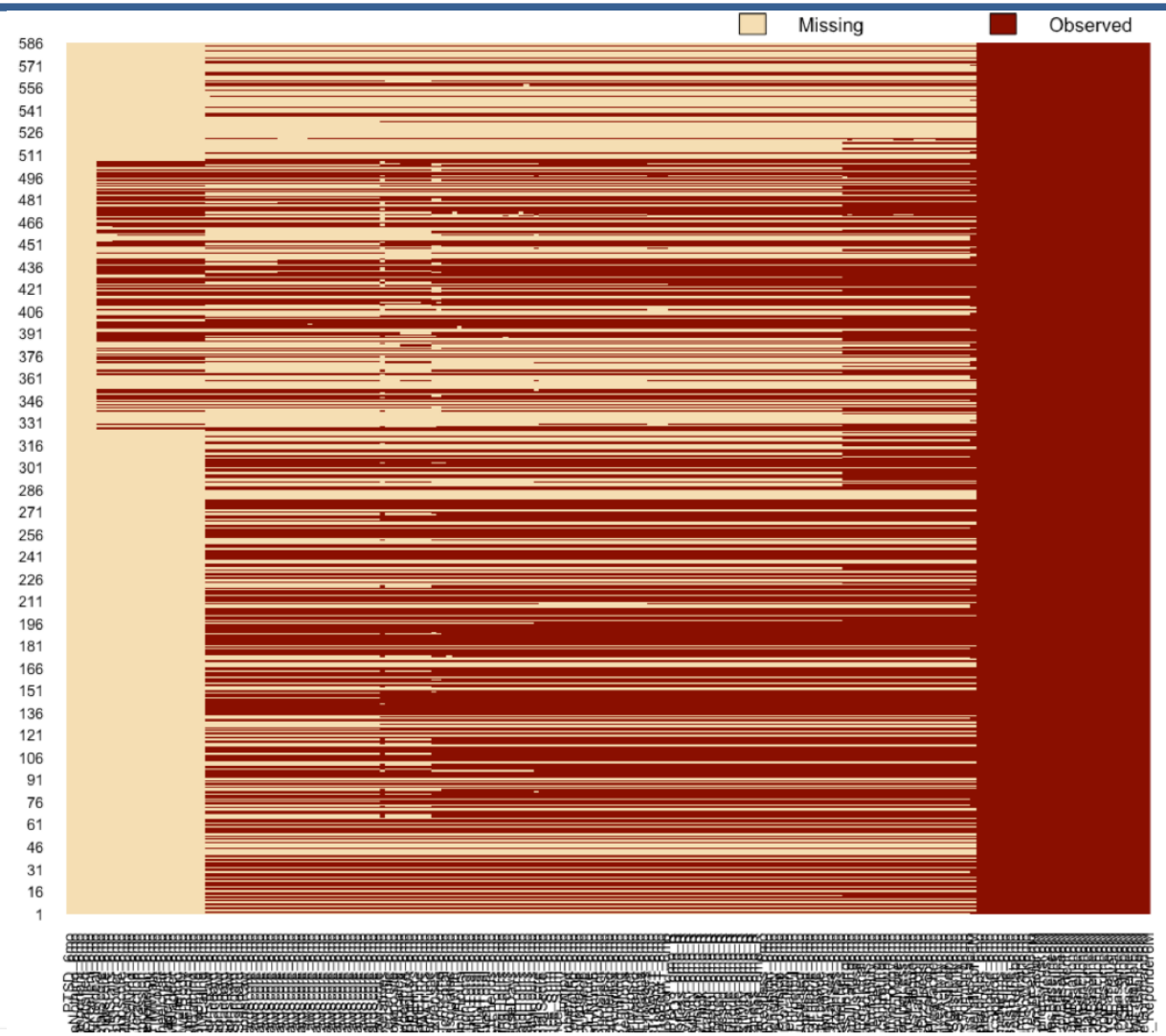
Data Structure



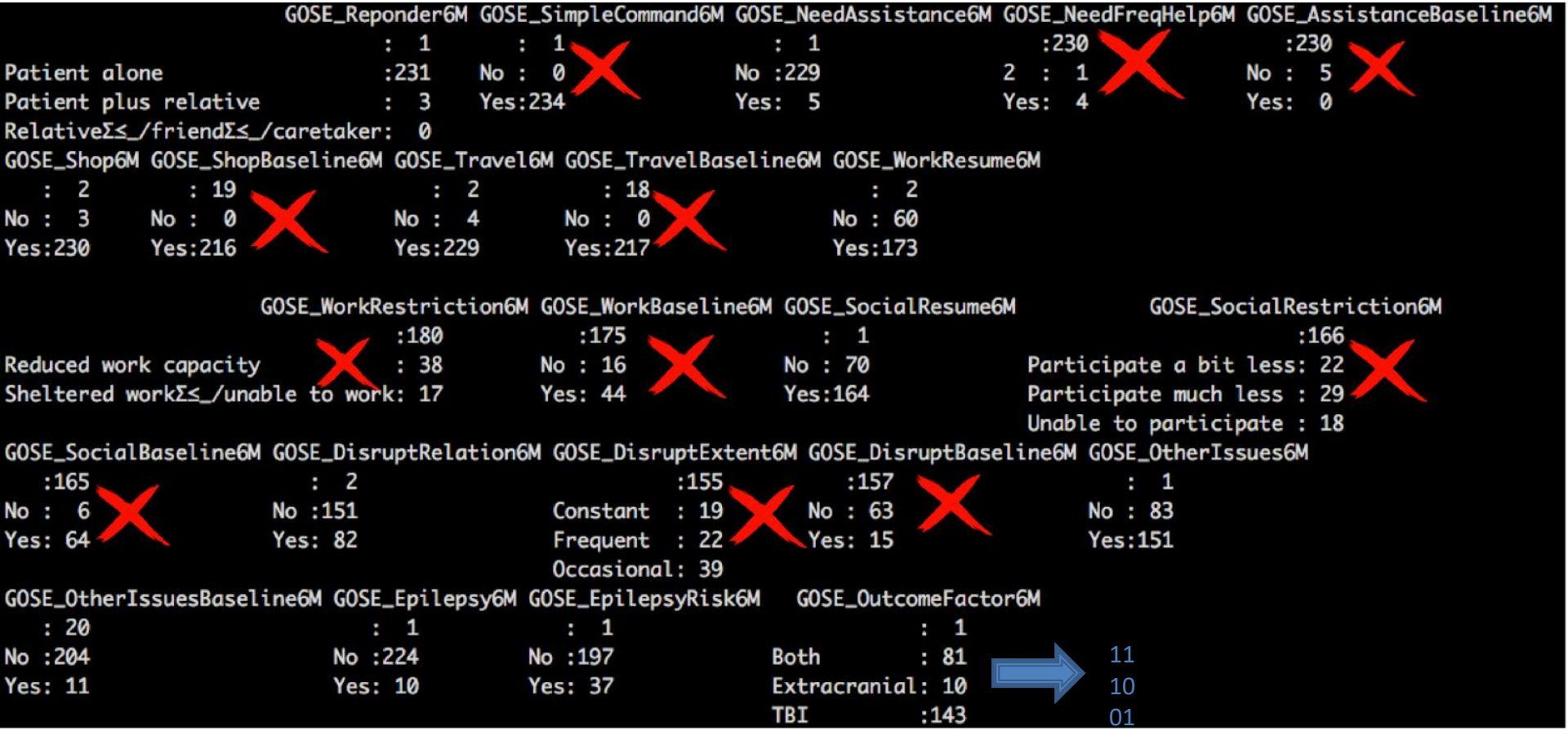
Nelson LD, Ranson J, Ferguson AR, Giacino J, Okonkwo DO, Valadka A, Manley G, McCrea M, (2017). Validating Multidimensional Outcome Assessment Using the TBI Common Data Elements: An Analysis of the TRACK-TBI Pilot Sample." J Neurotrauma. doi: 10.1089/neu.2017.5139.

Tbi.6m all r x c

- Data cleaning
  - Delete rows and columns with most/all values missing
  - Delete variables with constant values
  - Convert all categorical variables to numerical variables
    - Convert factors level into ordinal numerical variables if level relationship can be understood
    - Convert factors level to binary codes (with increased dimensions) if level relationship cannot be understood



Original GOSE Categorical Variables



Matrix Completion
- to impute missing values

- Reconstruct a low rank matrix from a subset of its entries
  - The low rank assumption is applicable here due to the redundancy among variables
- Let M be the original incomplete matrix and X be the estimated low rank complete matrix
- Define $(M^E)_{ij} = \begin{cases} M_{ij} & if\ M_{ij}\ is\ known \\ 0 & otherwise \end{cases}$ and $(X^E)_{ij} = \begin{cases} X_{ij} & if\ M_{ij}\ is\ known \\ 0 & otherwise \end{cases}$
- Then, min $\|M^E - X^E\|_F^2$ s.t. rank(X)≤r
- Finally,

$$\widehat{M}_{ij} = \begin{cases} M_{ij} & if\ M_{ij}\ is\ known \\ X_{ij} & otherwise \end{cases}$$

Method of Non-negative Matrix Factorization (NMF) for Similarity Data – for data clustering

- Similarity matrix S_{NxN}

$$S_{ij} = e^{-\frac{\|x_i - x_j\|^2}{\sigma^2}}$$

- Probability matrix $P = S/1_N^t S 1_N$
- Factorization $P = WH^t$, where

$$p_{ij} = P(x_i, x_j) = \sum_{v=1}^{K} w_{iv} h_{jv}$$

$$w_{iv} = P(x_i, c_v)$$

$$h_{jv} = P(x_j | c_v)$$

- The objective is to

$$\min_{W,H} C(P||WH^t) := -\sum_{i,j} [p_{ij} \log \sum_{v} (w_{iv} h_{jv})]$$

subject to $0 \le w_{iv}, h_{jv} \le 1, \forall i, j, v, \sum_{i,v} w_{iv} = 1, \sum_j h_{jv} = 1.$

- The label of x_i is

$$l_{x_i} = \arg\max_v w_{iv}$$

Model Selection
- to determine K

- The cross entropy keeps on decreasing with increased K
- Stability measure

$$S = \mathbb{E}\Big[\min_{\pi \in \Theta_K} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\pi\big(h_{\boldsymbol{X},\alpha_{\boldsymbol{X}}(\boldsymbol{X})}(\boldsymbol{X}_i')\big) \neq \alpha_{\boldsymbol{X}'}(\boldsymbol{X}_i')\}\Big]$$

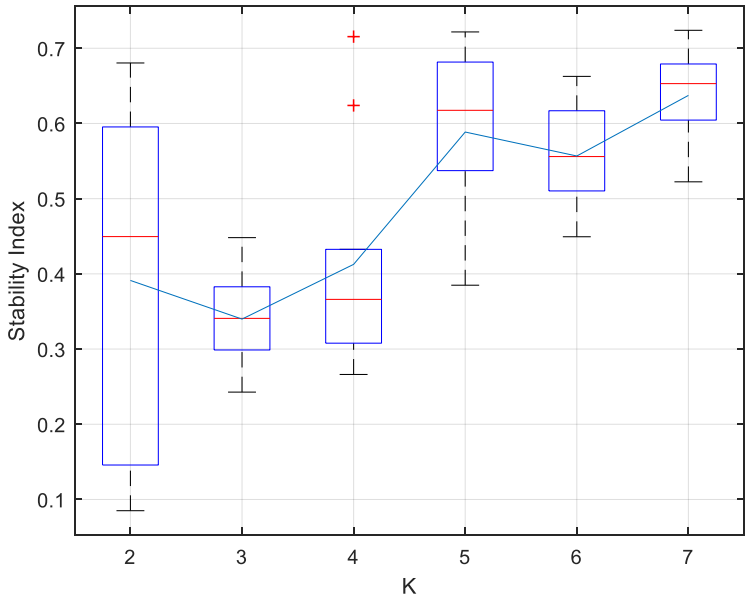
where $\alpha_{\boldsymbol{X}'}(\boldsymbol{X}_i')$ is the set of cluster labels for data $\boldsymbol{X}_i'$ using model order K and clustering method α, and $h_{\boldsymbol{X},\alpha_{\boldsymbol{X}}(\boldsymbol{X})}(\boldsymbol{X}_i')$ is the set of predicted labels of data $\boldsymbol{X}_i'$ using $\boldsymbol{X}, \alpha_{\boldsymbol{X}}(\boldsymbol{X})$ as the training data. Here, $\boldsymbol{X}$ and $\boldsymbol{X}_i'$ are from the same distribution. Θ_k is the set of all permutations of $\{1, \ldots, K\}$. On average, the smaller S, the more stable the model is.

TRACK-TBI Pilot Outcome Variables - 6 months: Summary Scores

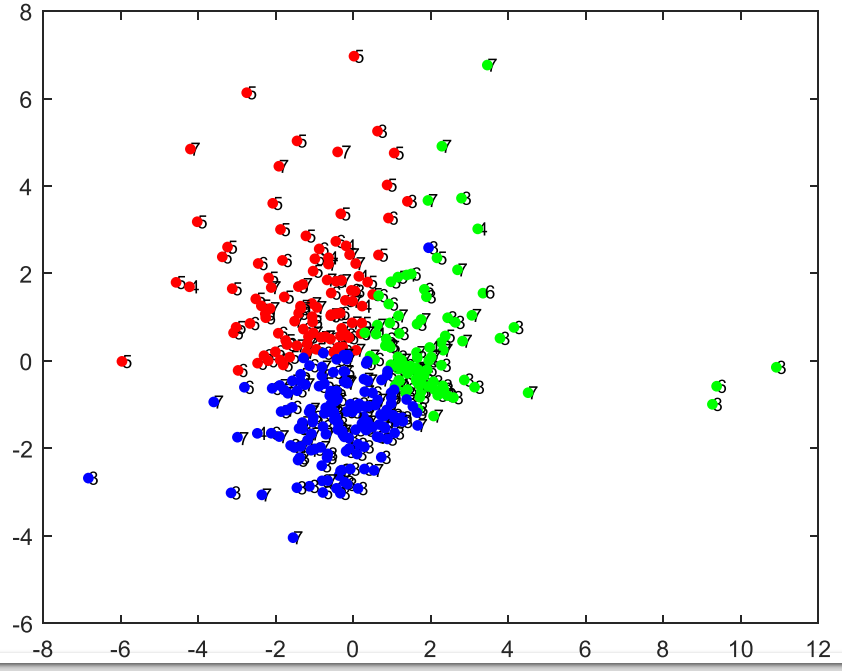
CDE Outcome Domain	Name of Measure	Variable Name	Coding
Global Outcome	Glasgow Outcome Scale Extended (GOSE)	GOSE_OverallScore	(1=death, 2=vegetative state, 3/4=lower/upper severe, 5/6=lower/ upper moderate, 7/8=lower/upper good recovery)
Post-concussive/TBI-related symptoms	Rivermead Post Concussion Symptoms Questionnaire (RPQ)	RPQ_Total	0 - 64 (higher = more injury-related symptoms)
Psychiatric and Psychological Status	Brief Symptom Inventory 18-item (BSI-18)*	BSI18DeprT	T scores with population M = 50, SD = 10; higher=more symptoms
		BSI18AnxT	T scores with population M = 50, SD = 10; higher=more symptoms
		BSI18SomT	T scores with population M = 50, SD = 10; higher=more symptoms
	PTSD Checklist--Civilian (PCL-C)	PCLTotalScore	17-84 (higher = more posttraumatic stress disorder symptoms)
Generic Quality of Life	Satisfaction With Life Scale (SWLS)	SWLSTotalScore	1-35 (higher=more satisfied with life)
Neuropsychological Impairment	Trail Making Test	TMTPartATime	# seconds (higher = slower psychomotor speed)
		TMTPartBTime	# seconds (higher = slower speed/mental flexibility/set-shifting)
	Wechsler Adult Intelligence Scale (WAIS-IV) Processing Speed Index (PSI)	WAIS_PSI_Composite	Standard score (population M = 100, SD = 15); higher = better/faster processing speed
	California Verbal Learning Test - Second Edition (CVLT-II)	CVLTTrial1To5RawScore	Higher = better memory
		CVLTTrialBRawScore	Higher = better memory
		CVLTShortDelayFreeRecall	Higher = better memory
		CVLTLongDelayFreeRecall	Higher = better memory
		CVLTTotalRecognitionDiscriminability	Higher = better memory
		CVLTTotalIntrusionsRaw	Higher = more intrusion errors (worse performance)
		CVLTTotalRepetitionsRaw	Higher = more repetitive responses (worse performance)

Analysis of Summary Scores: model selection

- Stability assessment boxplot showing the distribution of 10 repeats of stability value on y-axis for each model across different model on x-axis.

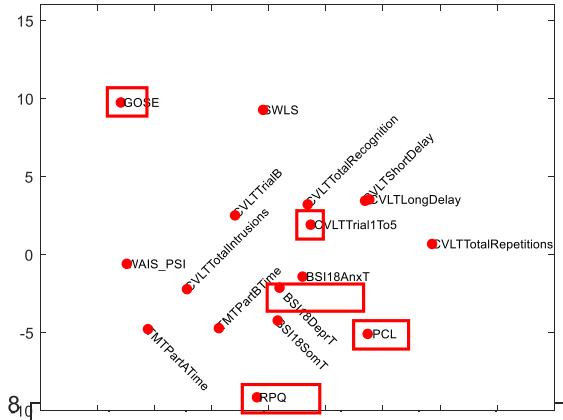


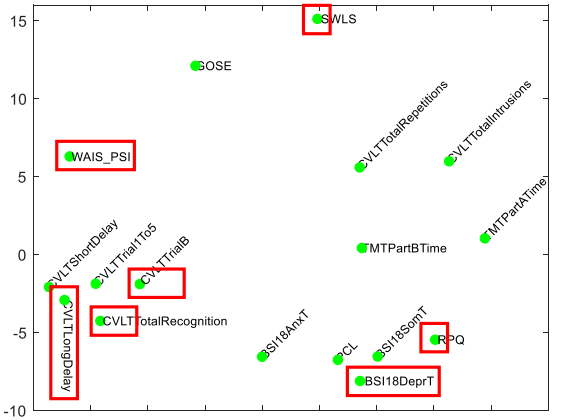
Analysis of Summary Scores: clustering for K=3

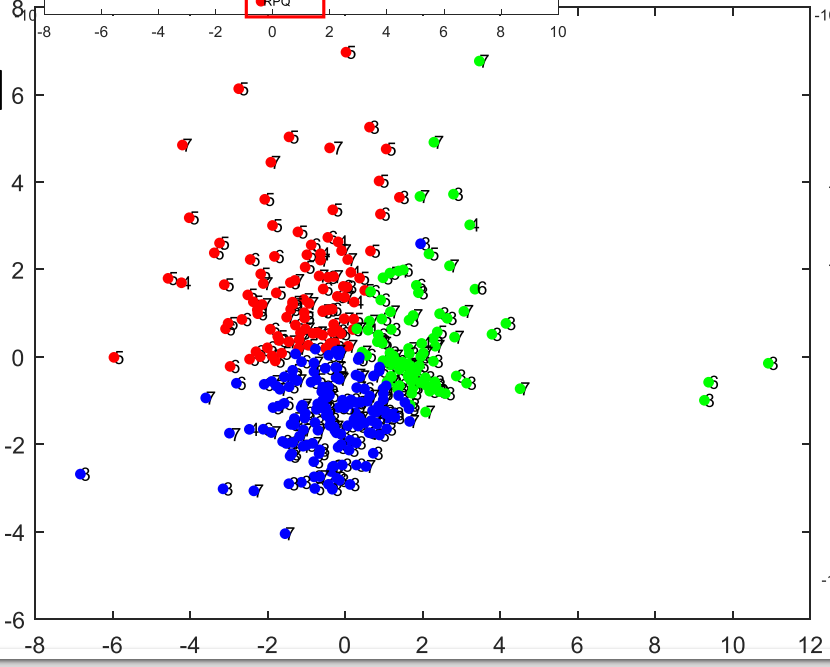


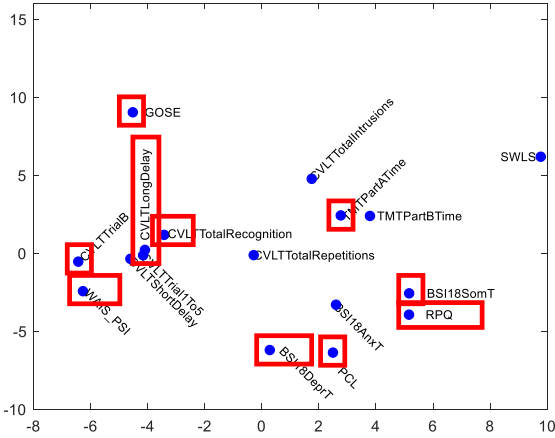
Analysis of Summary Scores: clustering for K=3

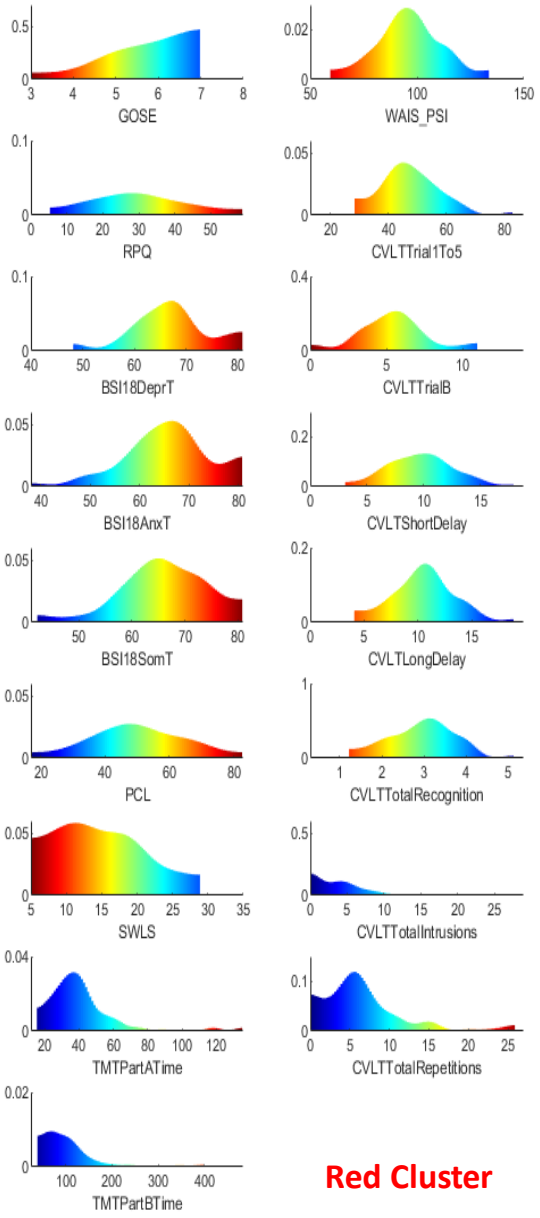
- To understand and characterize each cluster of patients, a binomial logistic regression model is built based on the assigned cluster labels and the step-wise feature selection procedure is conducted to select the contributing features in distinguishing these clusters.

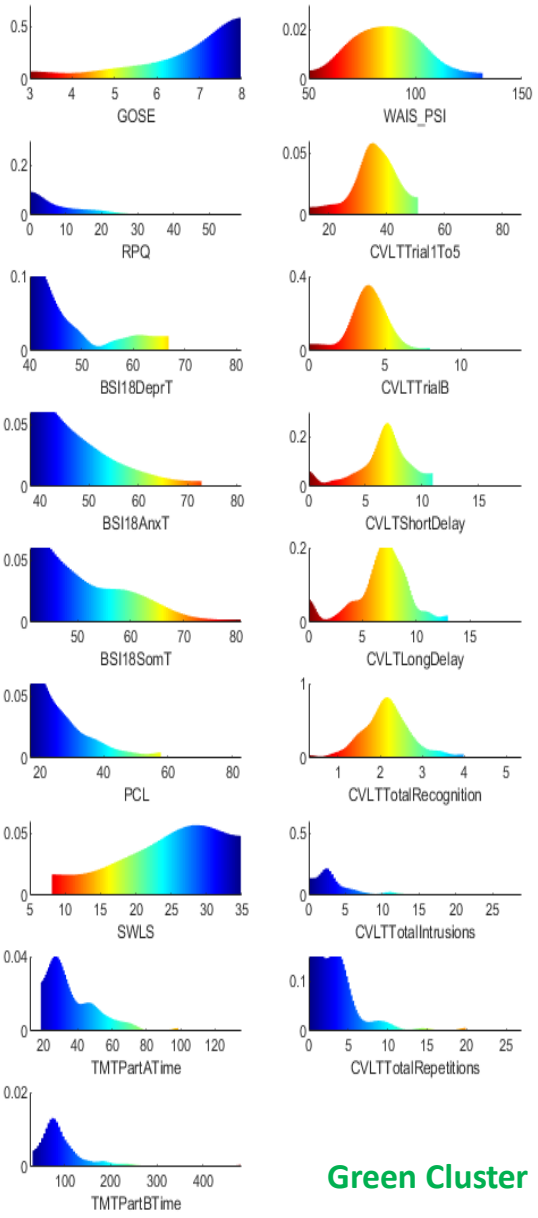


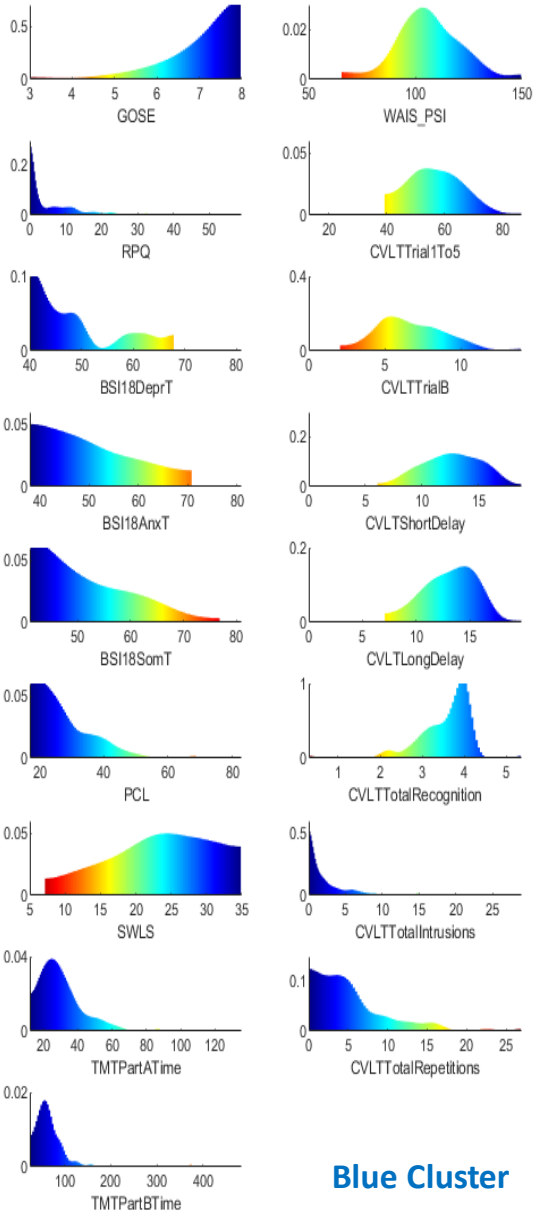


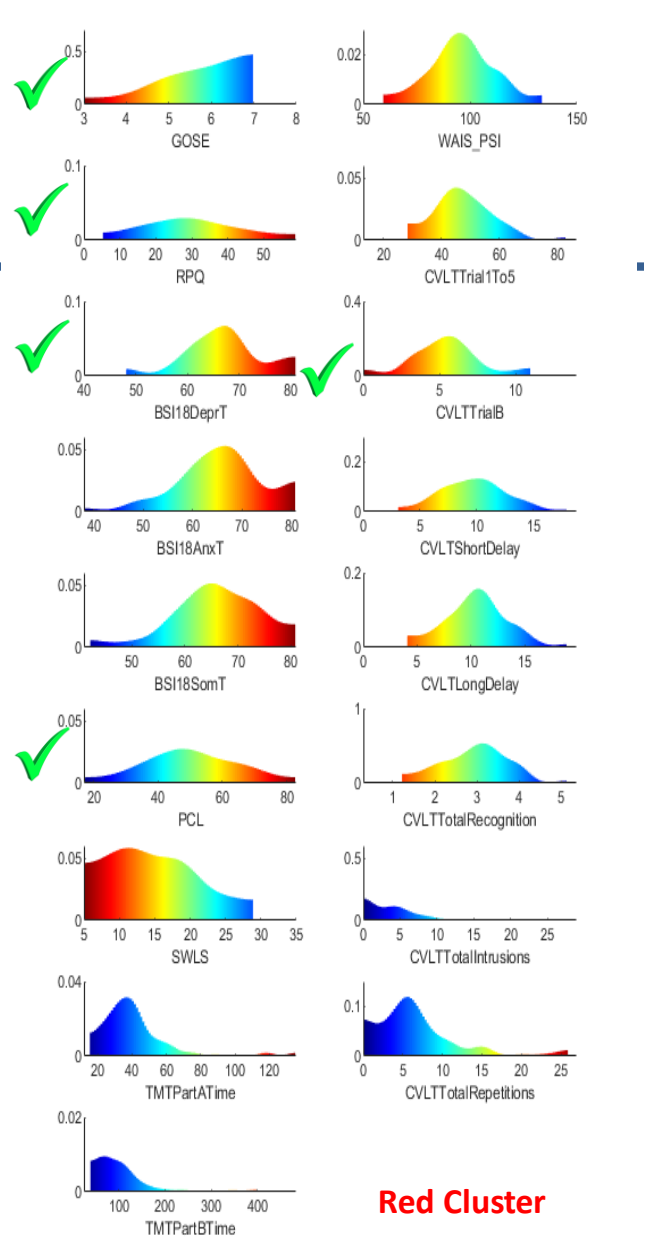




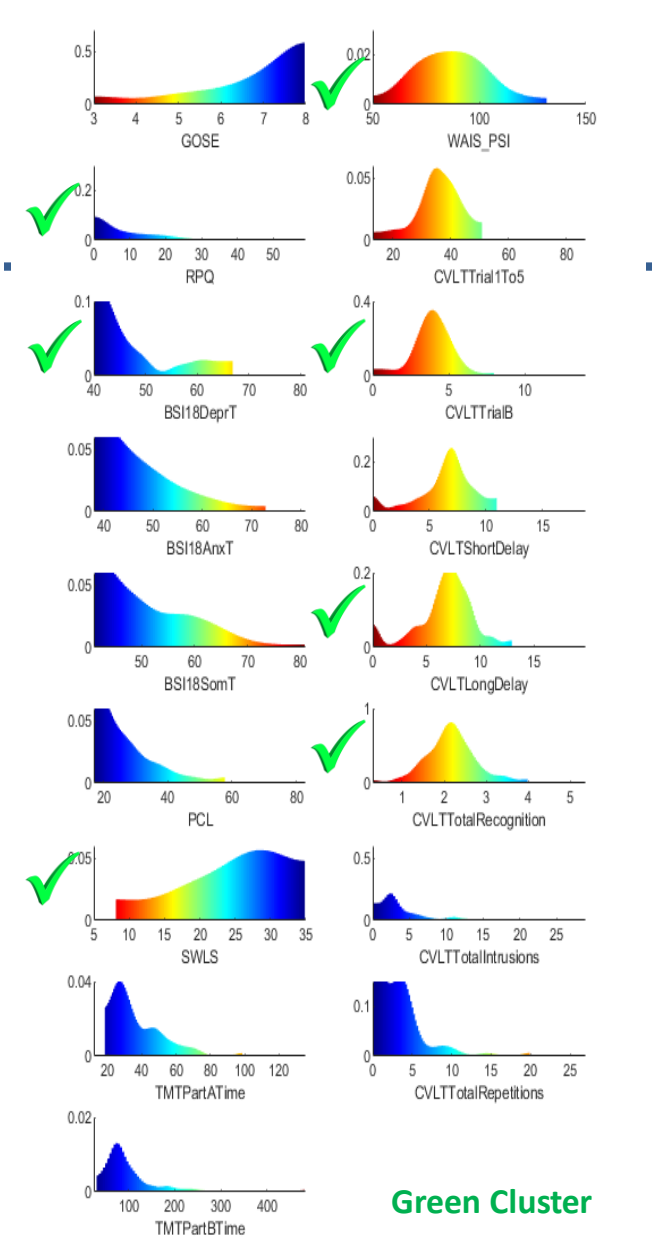

GOSE
WAIS_PSI
RPQ
CVLTTrial1To5
BSI18DeprT
CVLTTrialB
BSI18AnxT
CVLTShortDelay
BSI18SomT
CVLTLongDelay
PCL
CVLTTotalRecognition
SWLS
CVLTTotalIntrusions
TMTPartATime
CVLTTotalRepetitions
TMTPartBTime
Red Cluster


GOSE
WAIS_PSI
RPQ
CVLTTrial1To5
BSI18DeprT
CVLTTrialB
BSI18AnxT
CVLTShortDelay
BSI18SomT
CVLTLongDelay
PCL
CVLTTotalRecognition
SWLS
CVLTTotalIntrusions
TMTPartATime
CVLTTotalRepetitions
TMTPartBTime
Green Cluster

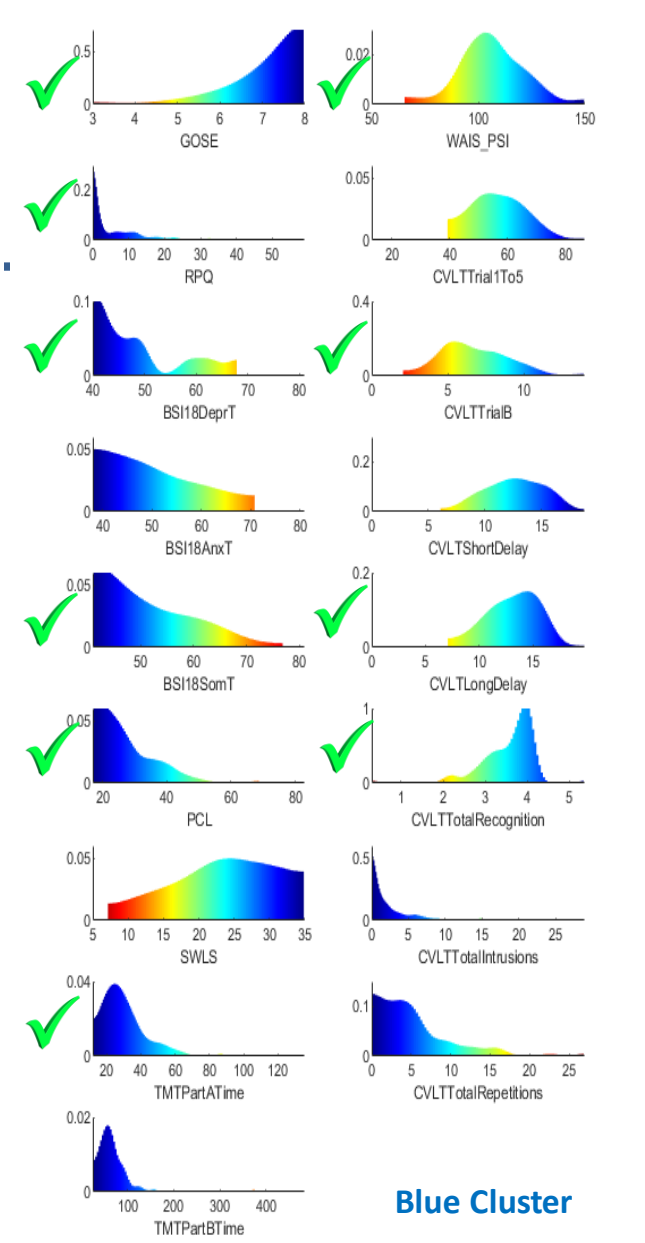

GOSE
WAIS_PSI
RPQ
CVLTTrial1To5
BSI18DeprT
CVLTTrialB
BSI18AnxT
CVLTShortDelay
BSI18SomT
CVLTLongDelay
PCL
CVLTTotalRecognition
SWLS
CVLTTotalIntrusions
TMTPartATime
CVLTTotalRepetitions
TMTPartBTime
Blue Cluster



Red Cluster

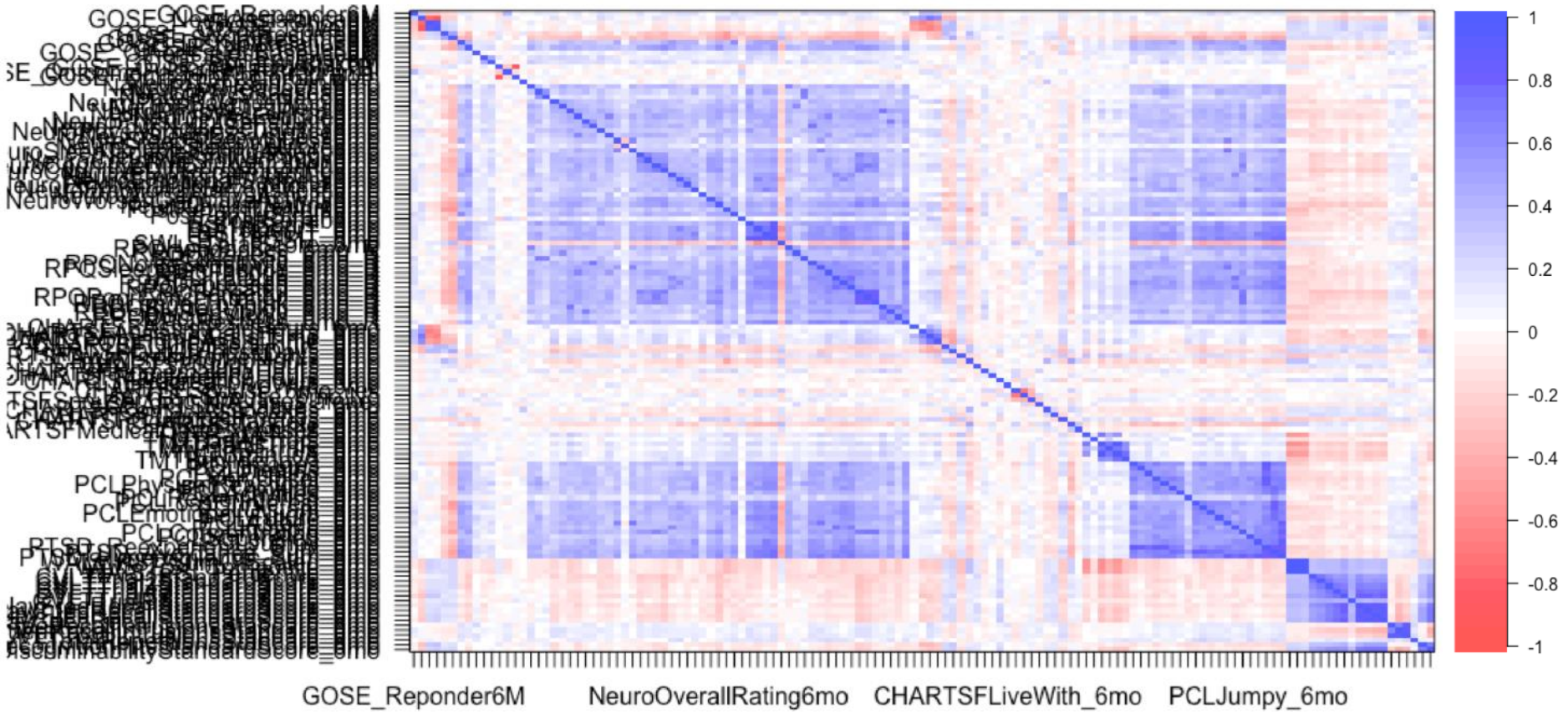


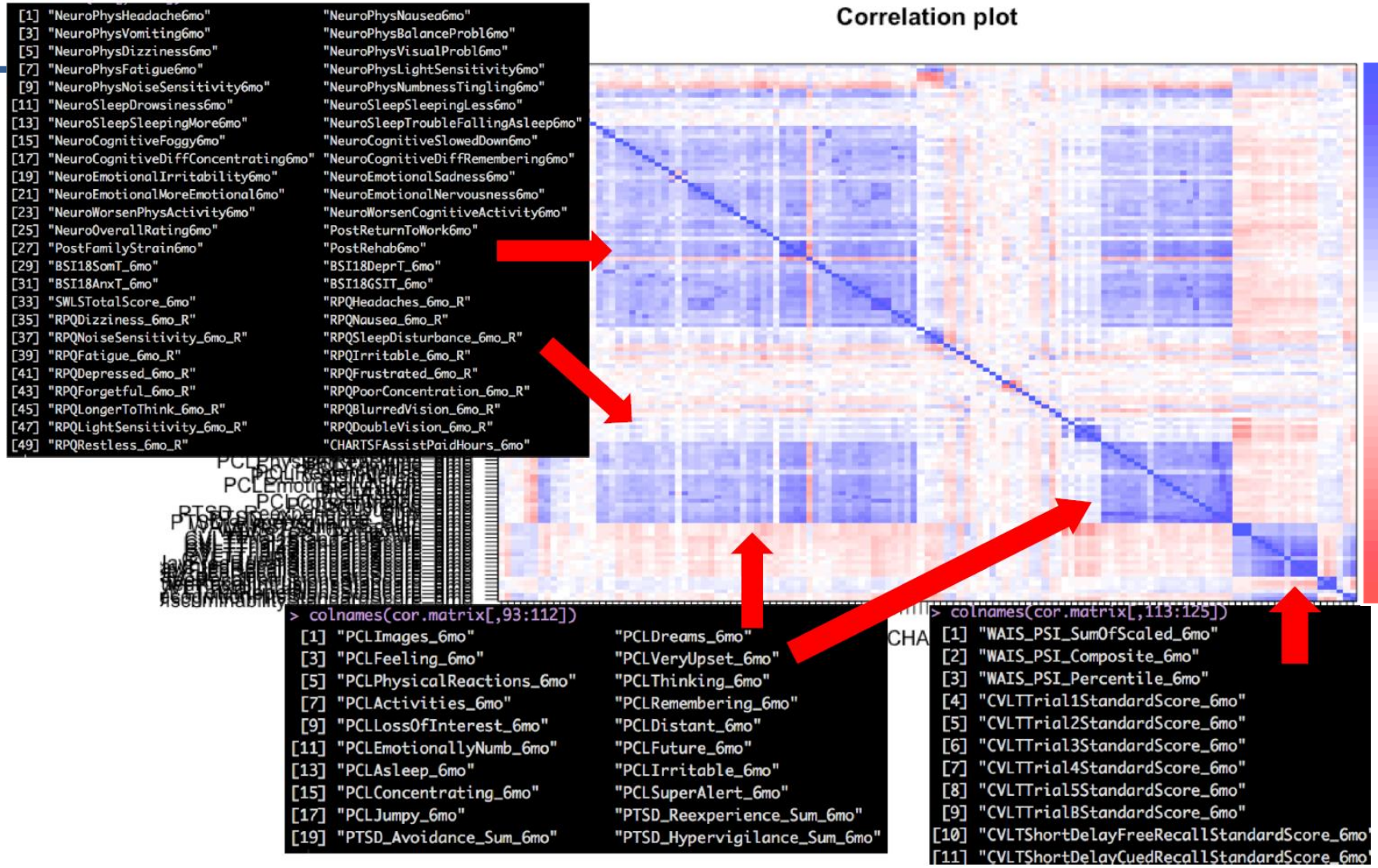
Green Cluster



Blue Cluster

Tbi.6m (132 variables)– correlation plot




[1] "NeuroPhysHeadache6mo" "NeuroPhysNausea6mo"
[3] "NeuroPhysVomiting6mo" "NeuroPhysBalanceProbl6mo"
[5] "NeuroPhysDizziness6mo" "NeuroPhysVisualProbl6mo"
[7] "NeuroPhysFatigue6mo" "NeuroPhysLightSensitivity6mo"
[9] "NeuroPhysNoiseSensitivity6mo" "NeuroPhysNumbnessTingling6mo"
[11] "NeuroSleepDrowsiness6mo" "NeuroSleepSleepingLess6mo"
[13] "NeuroSleepSleepingMore6mo" "NeuroSleepTroubleFallingAsleep6mo"
[15] "NeuroCognitiveFoggy6mo" "NeuroCognitiveSlowedDown6mo"
[17] "NeuroCognitiveDiffConcentrating6mo" "NeuroCognitiveDiffRemembering6mo"
[19] "NeuroEmotionalIrritability6mo" "NeuroEmotionalSadness6mo"
[21] "NeuroEmotionalMoreEmotional6mo" "NeuroEmotionalNervousness6mo"
[23] "NeuroWorsenPhysActivity6mo" "NeuroWorsenCognitiveActivity6mo"
[25] "NeuroOverallRating6mo" "PostReturnToWork6mo"
[27] "PostFamilyStrain6mo" "PostRehab6mo"
[29] "BSI18SomT_6mo" "BSI18DeprT_6mo"
[31] "BSI18AnxT_6mo" "BSI18GSIT_6mo"
[33] "SWLSTotalScore_6mo" "RPQHeadaches_6mo_R"
[35] "RPQDizziness_6mo_R" "RPQNausea_6mo_R"
[37] "RPQNoiseSensitivity_6mo_R" "RPQSleepDisturbance_6mo_R"
[39] "RPQFatigue_6mo_R" "RPQIrritable_6mo_R"
[41] "RPQDepressed_6mo_R" "RPQFrustrated_6mo_R"
[43] "RPQForgetful_6mo_R" "RPQPoorConcentration_6mo_R"
[45] "RPQLongerToThink_6mo_R" "RPQBlurredVision_6mo_R"
[47] "RPQLightSensitivity_6mo_R" "RPQDoubleVision_6mo_R"
[49] "RPQRestless_6mo_R" "CHARTSFAssistPaidHours_6mo"
Correlation plot
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
> colnames(cor.matrix[,93:112])
[1] "PCLImages_6mo" "PCLDreams_6mo"
[3] "PCLFeeling_6mo" "PCLVeryUpset_6mo"
[5] "PCLPhysicalReactions_6mo" "PCLThinking_6mo"
[7] "PCLActivities_6mo" "PCLRemembering_6mo"
[9] "PCLLossOfInterest_6mo" "PCLDistant_6mo"
[11] "PCLEmotionallyNumb_6mo" "PCLFuture_6mo"
[13] "PCLAsleep_6mo" "PCLIrritable_6mo"
[15] "PCLConcentrating_6mo" "PCLSuperAlert_6mo"
[17] "PCLJumpy_6mo" "PTSD_Reexperience_Sum_6mo"
[19] "PTSD_Avoidance_Sum_6mo" "PTSD_Hypervigilance_Sum_6mo"
> colnames(cor.matrix[,113:125])
[1] "WAIS_PSI_SumOfScaled_6mo"
[2] "WAIS_PSI_Composite_6mo"
[3] "WAIS_PSI_Percentile_6mo"
[4] "CVLTTrial1StandardScore_6mo"
[5] "CVLTTrial2StandardScore_6mo"
[6] "CVLTTrial3StandardScore_6mo"
[7] "CVLTTrial4StandardScore_6mo"
[8] "CVLTTrial5StandardScore_6mo"
[9] "CVLTTrialBStandardScore_6mo"
[10] "CVLTShortDelayFreeRecallStandardScore_6mo"
[11] "CVLTShortDelayCuedRecallStandardScore_6mo"

Data Fusion

- Recall, the objective is to

$$\min_{W,H} C(P||WH^t) \coloneqq -\sum_{i,j}[p_{ij}\log\sum_{v}(w_{iv}h_{jv})]$$

subject to $0 \le w_{iv}, h_{jv} \le 1, \forall i,j,v, \sum_{i,v} w_{iv} = 1, \sum_j h_{jv} = 1.$

- Now $P = \sum_{l=1}^{L} \alpha_l P_l, \sum_{l=1}^{L} \alpha_l = 1, \alpha_l \ge 0, \forall l.$

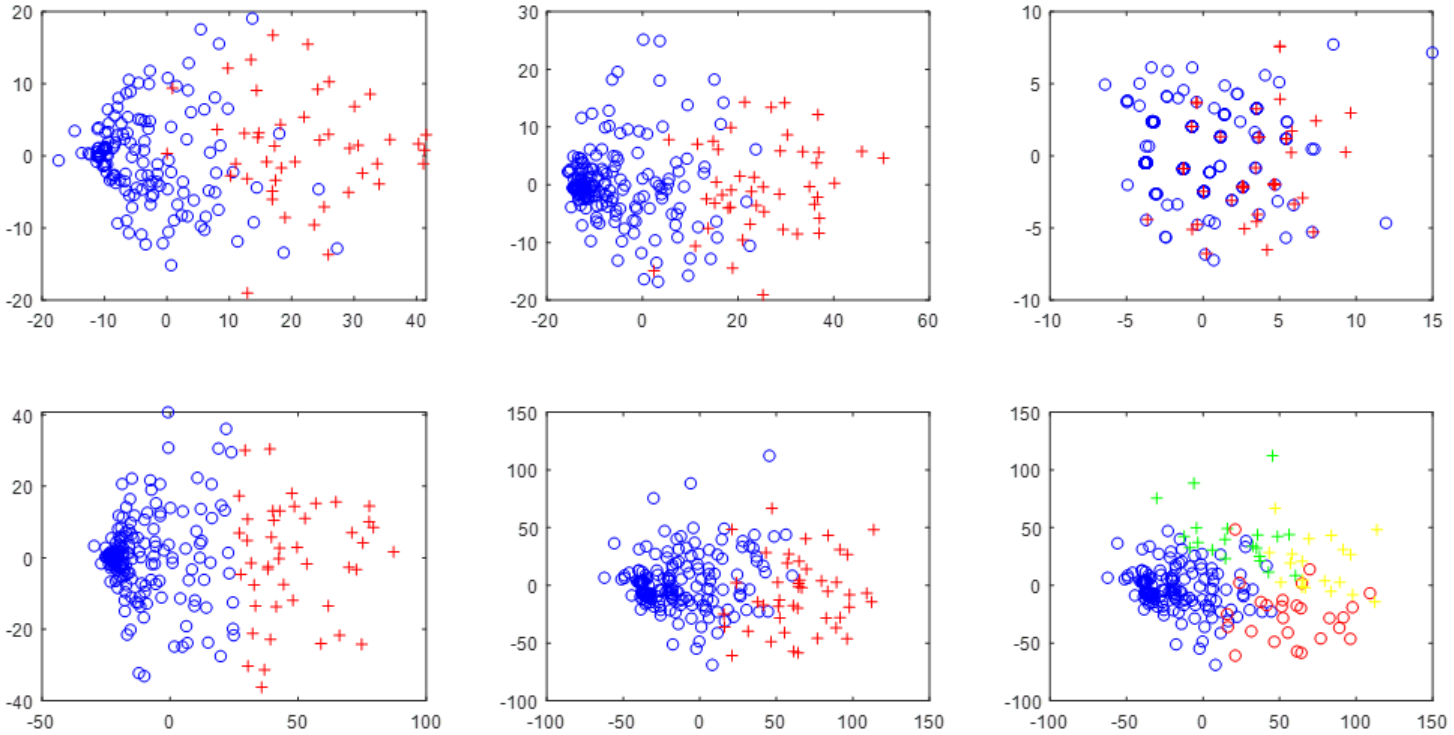
$$\min_{\alpha,W,H} C(P||WH^t) - \gamma En(\alpha) \coloneqq E_\alpha[C(P_l||WH^t)] - \gamma En(\alpha)$$

- Normalized mutual information (NMI) criterion for selecting data for fusion

$$NMI(label_1, label_2) = \frac{2I(label_1, label_2)}{En(label_1) + En(label_2)}$$

	RPQ & GOSE	RPQ & PCL	GOSE & PCL
NMI	0.1036	0.2918	0.0842

Clustering Fusion



- Data from RPQ and PCL can be fused because of their high correlation/NMI. Two sets of labels are generated, one based on both RPQ and PCL (label_1), and the other based on GOSE (label_2).
- (top left) label_1 plotted against RPQ data using MDS. (top middle) label_1 plotted against PCL data using MDS. (top right) label_1 plotted against GOSE data using MDS.
- (bottom left) label_1 plotted against RPQ+PCL data using MDS. (bottom middle) label_1 plotted against RPQ+PCL+GOSE data using MDS. (bottom right) label_1 & label_2 plotted against RPQ+PCL+GOSE data using MDS.

Summary

- Complete integration of the diverse data for TBI diagnosis and patient stratification remains an unmet challenge.
- NMF is applied to patient similarity data for clustering which identifies meaningful phenotypes based on 6-month outcome summary scores
- Step-wise feature selection is used to characterize each cluster
- Future work
  - The proposed framework will be applied to a larger dataset to test if the conclusions hold
  - The data fusion approach will be applied to core variables to characterize the trend of disease development
  - Determine predictors of outcome at 6 months by including pre-injury, acute injury and post injury variables in prediction models.