Sequential Pattern Mining Electronic Health Records for Early Diagnosis of ALS

Cindy Liang, Texas Academy of Mathematics and Science

Lily Sun, Stanford Online High School

William Jin, Community Middle School

Ying Liu, Princeton Pharmatech

Jeffrey Zhang, Princeton Pharmatech

Agenda

- Introduction
- Study Workflow
- Sequence pattern mining with cSPADE
- Results of classification
- Summary

Introduction

- Amyotrophic Lateral Sclerosis (ALS) is a **neurodegenerative disease**.
- ALS affects everything from talking to walking, affecting **all of the muscles** in the human body.
- 50% people diagnosed with ALS **live 3 years or less.**
- Annually, 2:100,000 people are diagnosed with ALS and is **difficult to diagnose.**
- Criteria for the diagnosis is based on the involvement of upper and/or lower motor neurons.
- Typical time between symptom onset and diagnosis is around **nine to twelve months.**

Photo Courtesy of Getty/ Bryan Bedder

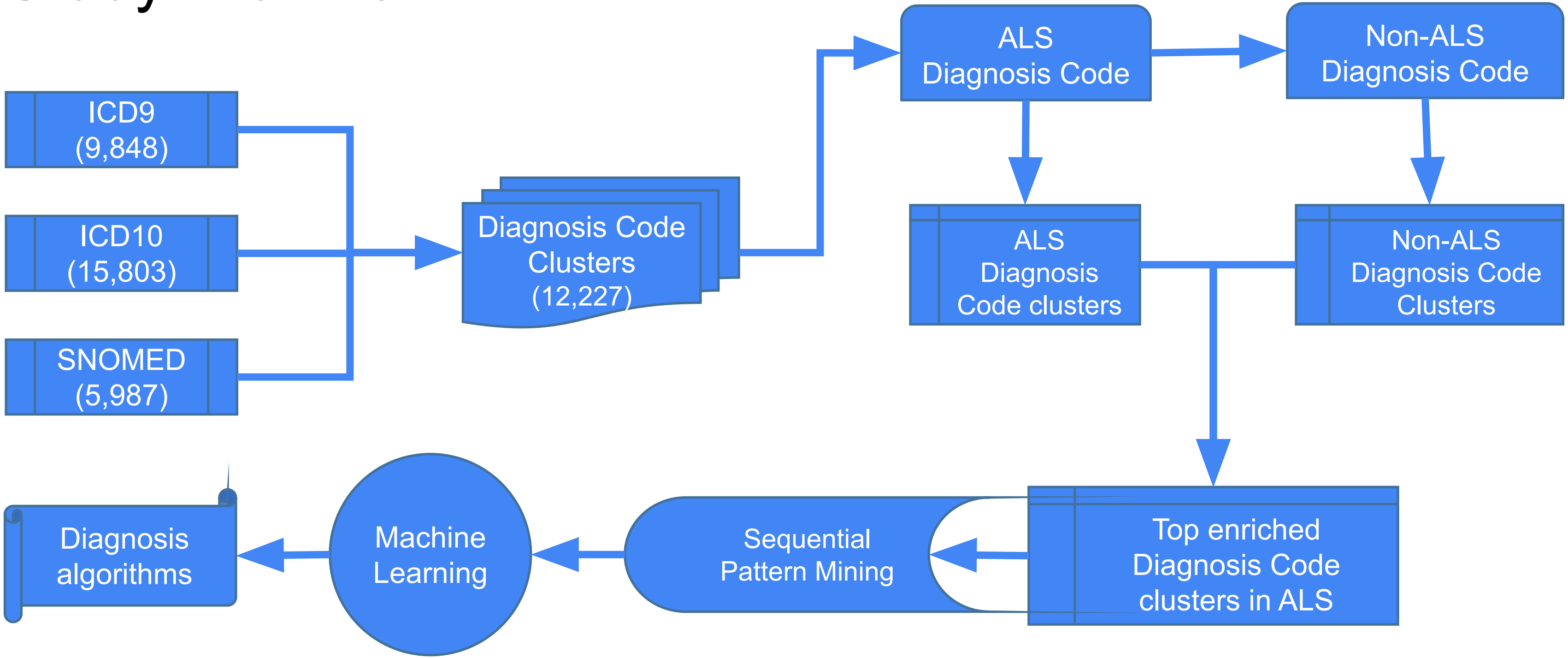
OBJECTIVE: Apply sequential pattern mining algorithm for early diagnosis of ALS disease.

Dataset from Optum’s Database

		ALS		Non-ALS	
AGE	Average	65.7		42.5	
ETHNICITY	Not Hispanic	14,194	85.4%	5,105,634	67.50%
	Hispanic	493	3.0%	548,762	7.30%
	Unknown	1,939	11.7%	1,907,275	25.20%
GENDER	Female	7,394	44.5%	4,077,765	53.90%
	Male	9,219	55.4%	3,466,964	45.80%
	Unknown	13	0.1%	16,942	0.20%
RACE	Caucasian	14,128	85.0%	4,724,569	62.50%
	Asian	162	1.0%	180,958	2.40%
	African American	1,065	6.4%	780,009	10.30%
	Other/Unknown	1,271	7.6%	1,876,135	24.80%
TOTAL		16,626		7,561,671	

Optum’s database is EMR-agnostic. Optum collects, normalizes and integrates provider data from different platforms and from different versions of the same platform.

Study Workflow


ICD9 (9,848)
ICD10 (15,803)
SNOMED (5,987)
Diagnosis Code Clusters (12,227)
ALS Diagnosis Code
Non-ALS Diagnosis Code
ALS Diagnosis Code clusters
Non-ALS Diagnosis Code Clusters
Top enriched Diagnosis Code clusters in ALS
Sequential Pattern Mining
Machine Learning
Diagnosis algorithms

ICD = **I**nternational **S**tatistical **C**lassifications of **D**iseases
SNOMED = **S**ystematized **No**menclature of Human **Med**icine.

Top Enriched Diagnosis Code Clusters for ALS Patients

Diagnosis Code Cluster	ALS		Non-ALS		Odds Ratio
	Yes	No	Yes	No	
ALS	16624	0	0	7561670	
NEURON	5658	10966	34691	7526980	111.95
SPEECH_DISORDER	4128	12496	91770	7469901	26.89
MUSCLE	5984	10640	165052	7396619	25.2
PARALYSIS	528	16096	9888	7551783	25.05
SPONDYLOSIS	2677	13947	70498	7491173	20.4
MUSCULOSKELETAL	4286	12338	153312	7408359	16.79
LACK_OF_COORDINATION	4312	12312	173226	7388445	14.94
DEPENDENCE_ON_WHEELCHAIR	349	16275	11378	7550293	14.23
DYSPHAGIA	4628	11996	202541	7359130	14.02
ABNORMAL_WALKING	3851	12773	160388	7401283	13.91
DEFORMITIES	1510	15114	55760	7505911	13.45
GASTROSTOMY	512	16112	18763	7542908	12.77
DISTURBANCE_OF_SALIVARY_SECRETION	367	16257	16741	7544930	10.17
RESPIRATORY_FAILURE	763	15861	49172	7512499	7.35
PNEUMONITIS	526	16098	36969	7524702	6.65
PULMONARY_DISEASE	201	16423	14113	7547558	6.55

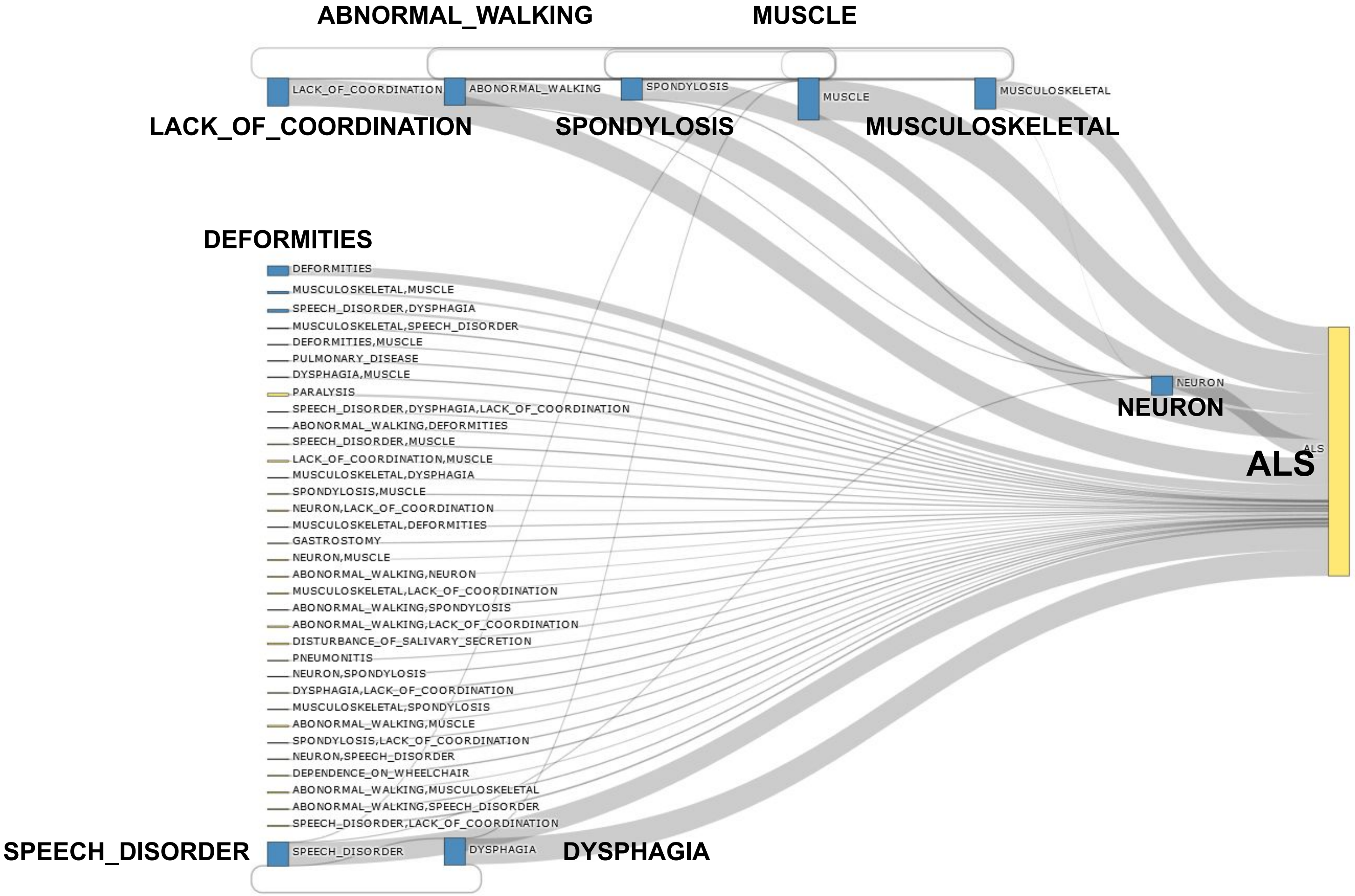
Experiment Setup

- Experiment using R (version 3.6.0) on 8 core (Intel(R) Xeon(R) CPU E5-2660 v3) Linux Server with 64 GB RAM.
- Sequence pattern search using cSPADE algorithm implementation in R
- Classification using CARET package with “GLM”, “MARS”, “svmRadial” algorithms in R
- Classification with 5-fold cross validation
- Data set 60/40 split for training and testing to avoid overfitting.

Sequence Pattern Mining with cSPADE Algorithm

Sequence	Cluster1 →	Cluster2 →	Cluster3 →	Prediction	ALS	Non-ALS
1	DYSPHAGIA	SPEECH_DISORDER	NEURON	ALS	3.10%	0.01%
2	DYSPHAGIA	NEURON	SPEECH_DISORDER	ALS	1.28%	0.01%
3	DEFORMITIES	NEURON		ALS	3.72%	0.02%
4	NEURON	DEFORMITIES		ALS	1.55%	0.02%
5	MUSCULOSKELETAL	NEURON	SPONDYLOSIS	ALS	0.92%	0.02%
6	MUSCLE	NEURON	SPONDYLOSIS	ALS	0.91%	0.02%
7	SPEECH_DISORDER	NEURON		ALS	8.21%	0.03%
8	NEURON	SPEECH_DISORDER		ALS	3.68%	0.03%
9	LACK_OF_COORDINATION	NEURON	MUSCULOSKELETAL	ALS	1.37%	0.03%
10	MUSCLE	NEURON	ABNORMAL_WALKING	ALS	1.57%	0.04%
11	MUSCLE	NEURON	MUSCULOSKELETAL	ALS	1.36%	0.04%
12	ABNORMAL_WALKING	NEURON	MUSCULOSKELETAL	ALS	1.14%	0.04%
13	SPONDYLOSIS	NEURON		ALS	6.26%	0.05%
14	NEURON	SPONDYLOSIS		ALS	2.86%	0.05%
15	NEURON	DYSPHAGIA		ALS	4.37%	0.06%
16	MUSCLE	NEURON		ALS	12.58%	0.08%
17	NEURON	MUSCLE		ALS	5.77%	0.08%
18	NEURON	MUSCULOSKELETAL		ALS	3.80%	0.08%
19	LACK_OF_COORDINATION	NEURON		ALS	10.25%	0.09%
20	NEURON	LACK_OF_COORDINATION		ALS	5.49%	0.09%
21	ABNORMAL_WALKING	NEURON		ALS	8.80%	0.10%
22	NEURON	ABNORMAL_WALKING		ALS	4.82%	0.10%

Sankey Diagram from Sequential Pattern Search


ABNORMAL_WALKING
MUSCLE
LACK_OF_COORDINATION
ABONORMAL_WALKING
SPONDYLOSIS
MUSCLE
MUSCULOSKELETAL
LACK_OF_COORDINATION
SPONDYLOSIS
MUSCULOSKELETAL
DEFORMITIES
DEFORMITIES
MUSCULOSKELETAL,MUSCLE
SPEECH_DISORDER,DYSPHAGIA
MUSCULOSKELETAL,SPEECH_DISORDER
DEFORMITIES,MUSCLE
PULMONARY_DISEASE
DYSPHAGIA,MUSCLE
PARALYSIS
SPEECH_DISORDER,DYSPHAGIA,LACK_OF_COORDINATION
ABONORMAL_WALKING,DEFORMITIES
SPEECH_DISORDER,MUSCLE
LACK_OF_COORDINATION,MUSCLE
MUSCULOSKELETAL,DYSPHAGIA
SPONDYLOSIS,MUSCLE
NEURON,LACK_OF_COORDINATION
MUSCULOSKELETAL,DEFORMITIES
GASTROSTOMY
NEURON,MUSCLE
ABONORMAL_WALKING,NEURON
MUSCULOSKELETAL,LACK_OF_COORDINATION
ABONORMAL_WALKING,SPONDYLOSIS
ABONORMAL_WALKING,LACK_OF_COORDINATION
DISTURBANCE_OF_SALIVARY_SECRETION
PNEUMONITIS
NEURON,SPONDYLOSIS
DYSPHAGIA,LACK_OF_COORDINATION
MUSCULOSKELETAL,SPONDYLOSIS
ABONORMAL_WALKING,MUSCLE
SPONDYLOSIS,LACK_OF_COORDINATION
NEURON,SPEECH_DISORDER
DEPENDENCE_ON_WHEELCHAIR
ABONORMAL_WALKING,MUSCULOSKELETAL
ABONORMAL_WALKING,SPEECH_DISORDER
SPEECH_DISORDER,LACK_OF_COORDINATION
SPEECH_DISORDER
SPEECH_DISORDER
DYSPHAGIA
DYSPHAGIA
NEURON
NEURON
ALS
ALS

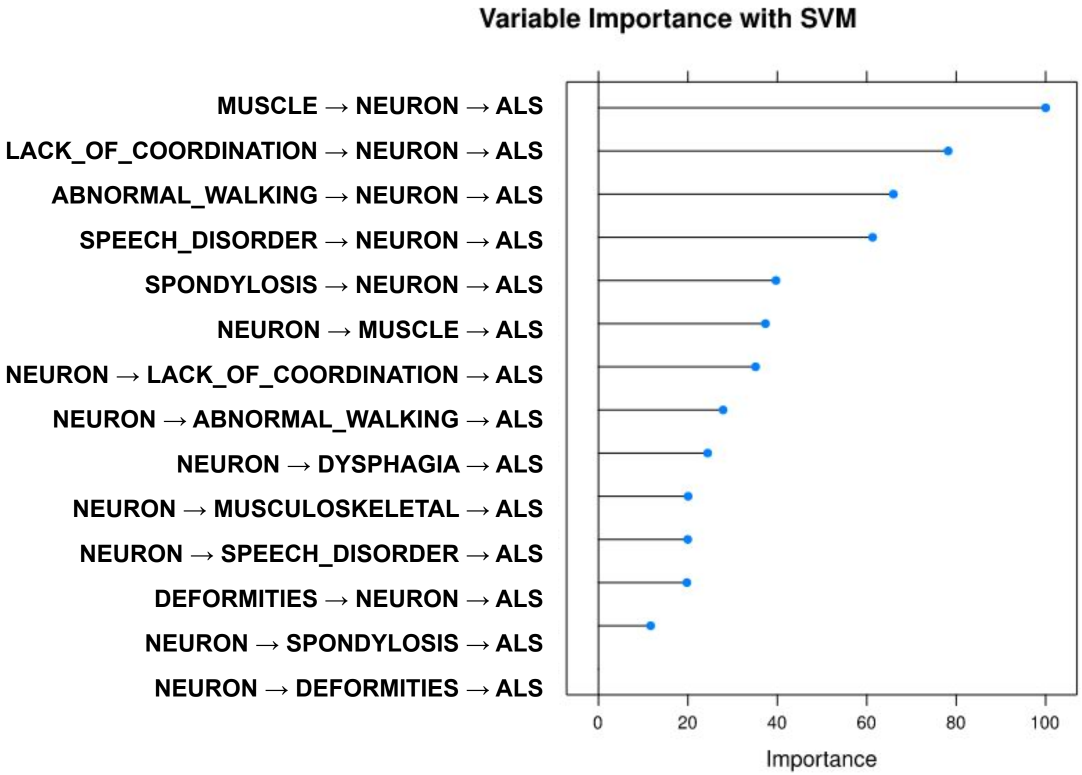
Classification Results

(Training/testing 60:40 split)

Classification Algorithm	Sensitivity	Specificity	True positive	True negative	False positive	False negative
GLM	10.59%	99.9947%	704	3024509	159	5945
MARS	9.23%	99.9962%	614	3024552	116	6035
SVM	24.26%	100%	1613	3024668	0	5036

GLM = Generalized Linear Model
MARS = Multivariate Adaptive Regression Spline
SVM = Support Vector Machines with Radial Basis Function Kernel

Prediction by SVM

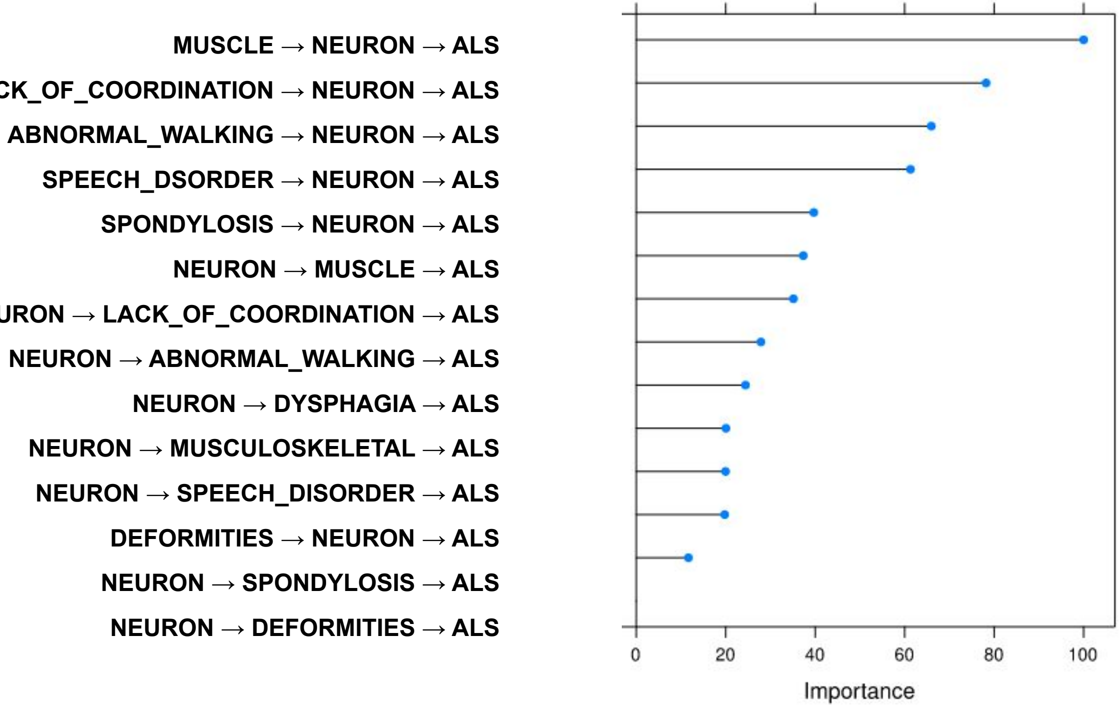


		Prediction	
		ALS	Non-ALS
Diagnosis	ALS	1,613	5,036
	Non-ALS	0	3,024,668

Sensitivity = 24.26%
Specificity = 100%

Average days with one of the 14 sequences before diagnosis of ALS = 169.2 days

Prediction by SVM



		Prediction	
		ALS	Non-ALS
Diagnosis	ALS	1,613	5,036
	Non-ALS	0	3,024,668

Sensitivity = 24.26%
Specificity = 100%

Average days with one of the 14 sequences before diagnosis of ALS = 169.2 days

Summary

- Through this project, we applied sequential pattern mining algorithm for early diagnosis of ALS disease.
- We noticed that the SVM algorithm was able to provide both the highest sensitivity and specificity.
- It is estimated that our algorithm is able to diagnose ALS more than 5 months before diagnosis in health records

Future Plans

- We hope to refine our algorithms to achieve higher sensitivities while keeping the high specificity
- Friendly User Interface for the program

Early Diagnosis of ALS will allow for more effective treatment and the mitigation of the effects of this disease.

Acknowledgements

- Princeton Pharmatech
- ASA JSM