Mitsubishi Project
Environmental Risk Factors for ALS

Claire Wang
Jun 2020 - Aug 2020

Dartmouth
GEISEL SCHOOL OF
MEDICINE

Agenda

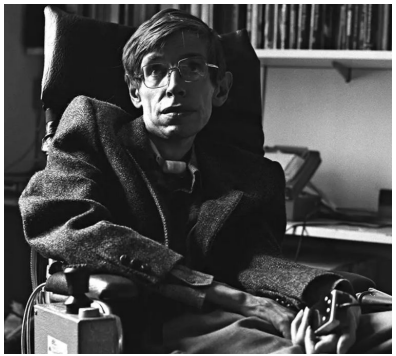
- Background
- Database
- Methods
- Results
- Summary

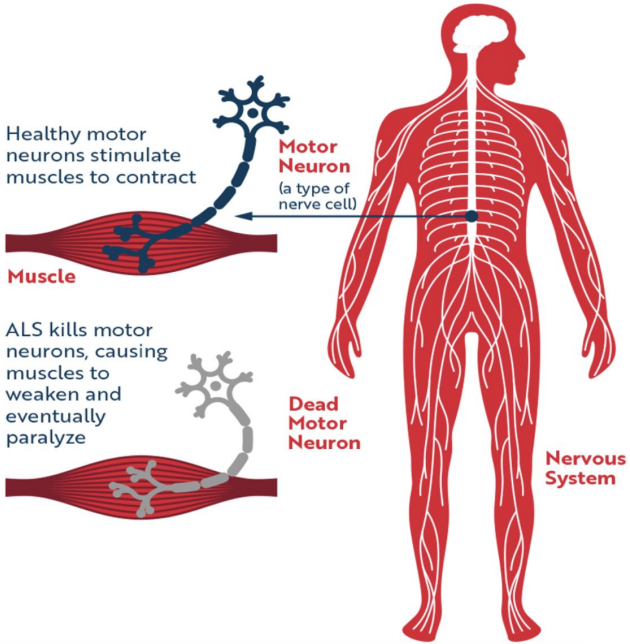
What Is ALS

- Amyotrophic lateral sclerosis
- A nervous system disease in which a person's brain loses connections with the muscles
- Causes weakness, disability, and eventually death from failure of the ventilatory muscles
- One can be diagnosed with ALS after on average of 9 to 12 months since they begin to notice symptoms
- 20% more common in men than in women

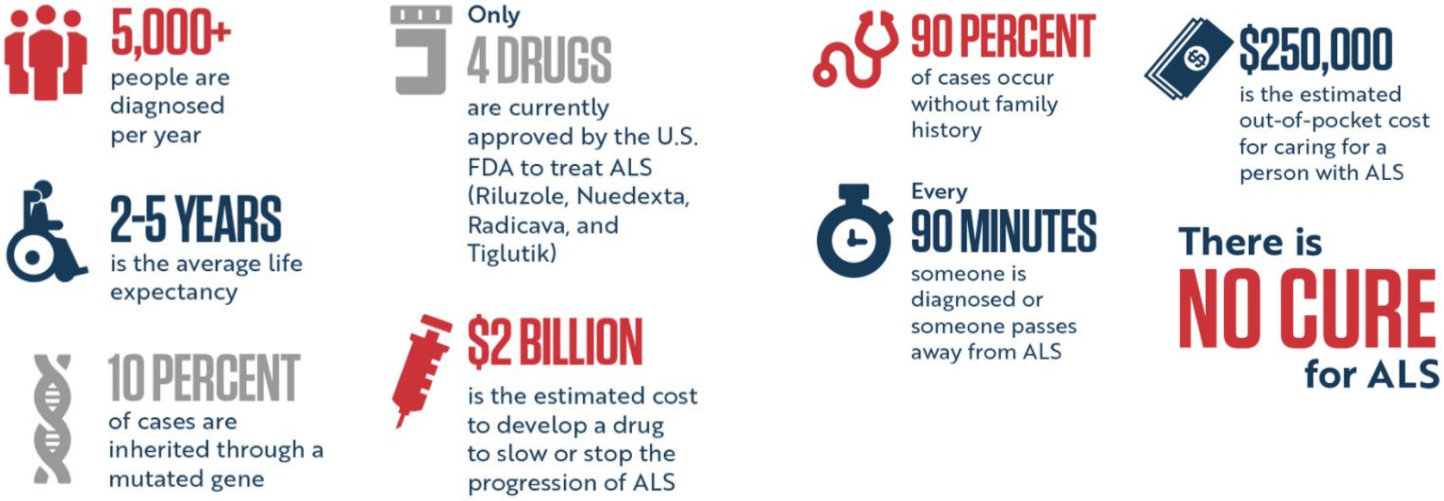
What is ALS

- Stephen Hawking diagnosed with ALS
- Lived with ALS for 55 years
- March 14, 2018



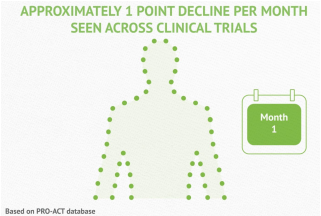


What is ALS

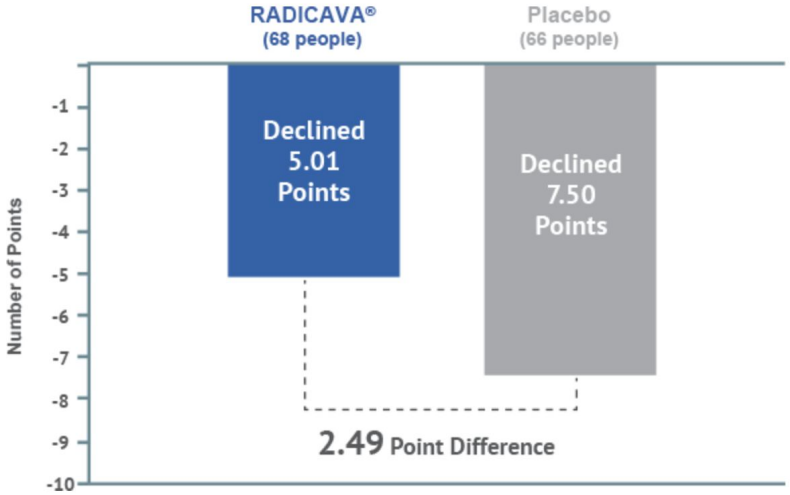

5,000+
people are diagnosed per year
2-5 YEARS
is the average life expectancy
10 PERCENT
of cases are inherited through a mutated gene
Only
4 DRUGS
are currently approved by the U.S. FDA to treat ALS (Riluzole, Nuedexta, Radicava, and Tiglutik)
$2 BILLION
is the estimated cost to develop a drug to slow or stop the progression of ALS
90 PERCENT
of cases occur without family history
Every
90 MINUTES
someone is diagnosed or someone passes away from ALS
$250,000
is the estimated out-of-pocket cost for caring for a person with ALS
There is
NO CURE
for ALS

The Mitsubishi Project - Background

- Mitsubishi Tanabe Pharma America (MTPA) markets Radicava for the treatment of ALS
- Can slow the progress of ALS by 25% - 30%
- None of them cure the disease
- Radicava acts as an antioxidant



The average total score on the ALS Functional Rating Scale–Revised (ALSFRS-R) after 24 weeks



The Mitsubishi Project - Background

Researchers and scientists are seeking a different pathway to improve the performance of the drug

Opportunity:

- Alternative pathway: the ethnobotanical identification of flora with toxic or therapeutic activity (Cox et al. 2016)
- Study has shown that classes of environmental toxins increase the likelihood of ALS, especially in regard to pesticides

Challenge:

- However, people are likely exposed to multiple chemicals
- It is too soon for the scientists to know which individual chemicals, or mixtures of chemicals, lead to motor neuron damage

The Mitsubishi Project - Goal

On the other hand:

- Some of environmental toxicants have synergistic effects
  - BMAA and mercury (Rush et al. 2012)

Project objective:

- Identify environmental pollutants that are risk factors for the development of ALS
- Explore the potential toxicants with synergistic effects
- Propose the results for drug development

My Goal:

- Develop the model tested on the old patients data to predict the future cohort's disease outcome

The Database

Symphony Health
A PRA Health Sciences Company

The Symphony Integrated Dataverse® (IDV®) database:

- The most comprehensive and interconnected source of healthcare data in the industry
- A medical claims database with more over 240M patients

Components of IDV:

- Reclaim resources: practitioner, procedure, diagnosis, patient, and payer information
  - Hospital
  - Medical
  - Prescription
- Point-of-sale prescription data
- Non-retail Invoice data
- Demographics

The Database

Our project data:

- Symphony healthcare claims dataset
- Merged with Air Toxics Release Inventory(TRI) data for chemicals released from large industrial facilities
- Age and gender matched controls for case-control comparison
- 104796 rows, 457 columns
- 78597 controls, 26199 ALS cases

Outcome: ALS_stat	ALS = 1	Control = 0
Location: zip3	Min = 10	Max = 994
Year: dx_year	Min = 2013	Max = 2019
Age: dx_age	Min = 18	Max = 81
Sex: sex	Male	Female
Toxins: (446 types)	From: X2.4.D	To: ZOXAMIDE

The Analytical Approach:

Programming Language:

- R

Data:

- Random assignment
- Train-test split
- Data cleaning/imputation

Methods:

- Pair-wise interaction analysis on the train set (package: Glinternet) ⤶ 1st data filtering
- Logistic regression model with interaction terms on the test set
- Computing false discovery rate given the p-values (package: FDR) ⤶ 2nd data filtering
- Descriptive Analysis on findings of synergistic pairs (package: Tableone)

Step: Random Split

1. Find the year that splits the data to 2:1. Obtain the "old_ALS" data and "recent_ALS" data (package: dyplyr)
2. Random assign the patient data from control group to train and test set with a 0.65:0.35 ratio
3. The "old_ALS" data merged to the train set, "recent_ALS" merged to the test set

dx_year	Als_stat	Counts	Cumsum
2013	1	3537	0.1350052
2014	1	4737	0.3158136
2015	1	4650	0.4933013
2016	1	4195	0.6534219
2017	1	4428	0.822436
2018	1	4407	0.9906485
2019	1	245	1
NA	0	78597	-

Step: Data Cleaning

1. Eliminate the columns(toxins) that have more than 80% NA's
    a. 164 toxins dropped,
    b. 149 toxins have more than 50% NA's out of the rest 282 toxins

2. Convert the data to continuous/categorical
    a. Approach 1: impute NA with smallest non-NA divided by 2
    b. Approach 2: categorize the variables
        1. NA = 0
        2. Values < median of the non-NA's = 1
        3. Values > median of the non-NA's = 2

Step: Pairwise Interaction Analysis

X = 282 predictors(toxins)

Y = ALS_stat

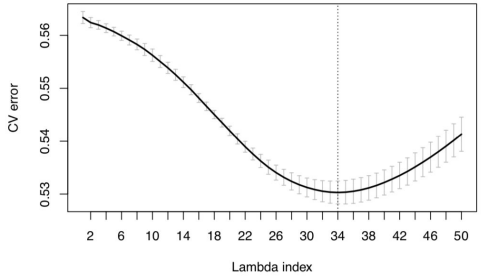
Model: glinternet.cv

- The glinternet model return the main effect coefficients and interaction coefficients for each pair
- Glinternet.cv does k-fold cross validation for glinternet

Step: Pairwise Interaction Analysis

Model 1:

- *numLevels = rep(3, 282) # for categorical variables, 3 levels*
- *glinternet.cv(X, Y, numLevels, nFolds = 3, family = "binomial")*
- X from approach 2 (categorical variables)
- Choose lambda1Std = 28, the largest value of lambda that produces a cv error that is within 1 standard deviation of the minimum cv error



Step: Pairwise Interaction Analysis

- Extract all the coefficients under Lambda1Std, i_1Std = 28
  - *coefs <- coef(gcv_fit$glinternetFit)[[i_1Std]]*
- Build a table *that saves the obtained interactions (603 pairs) with*
  - Indicies of toxins
  - Name of toxins
  - Main effect coefficients - 3 levels
  - Interaction coefficients - 3 x 3 = 9 per pair

```
head(main_effect_table)
```

```
##                [,1]          [,2]          [,3]
## [1,] -9.628271e-05  8.591761e-05  0.0000103682
## [2,] -2.201447e-03  8.146961e-03 -0.0059453096
## [3,]  3.818252e-03 -6.330695e-03  0.0025126298
```

```
coefs$interactionsCoef$catcat[[1]]
```

```
##              cat187_0      cat187_1      cat187_2
## cat1_0 -0.0001602940 -2.314479e-05  0.0001834357
## cat1_1 -0.0002760016 -4.336528e-04  0.0007096513
## cat1_2  0.0004362925  4.567945e-04 -0.0008930901
```

Step: Logistic Regression

Validate with logistic regression model on the test set:

- Extract and save the interactions coefficient and p-values for each pair obtained from the glinternet model
- *summary(glm(Y_test ~ Age + Sex + X_test[,1]* X_test[,187]))$coefficients[6,c(1,4)]*

```
##    Estimate    Pr(>|t|)
## -0.01062673  0.04936902
```

- The final table:

Att_1_index	Att_2_index	Att_1_name	Att_2_name	att1_0_vs_att2_0	att1_0_vs_att2_1	att1_0_vs_att2_2	att1_1_vs_att2_0	att1_1_vs_att2_1	att1_1_vs_att2_2	att1_2_vs_att2_0	att1_2_vs_att2_1	att1_2_vs_att2_2	Estimated_Coef	p_value
1	187	X2.4.D	MYCLOBUTANIL	-0.000160294029054934	-2.31447927396166E-05	0.00018343572230256	-0.000276001581822604	-0.000433652810603422	0.000709651292934036	0.000436292511385547	0.000456794503851049	0.00018343572230256	-0.0106267264851528	0.0493690201355432
2	99	X2.4.DB	FAMOXADONE	-0.000684442704602552	0.000801934379376638	-0.000117502042294 25	-0.000181992138687498	0.00187793178493654	-0.0016959500137692	0.000866424475769887	-0.00267987653183334	-0.000117502042294 25	-0.00839271729890425	0.0467276629099446
2	191	X2.4.DB	NAPROPAMIDE	-0.001118936286441 59	-0.000463578797220171	0.00158250307267035	0.00180883847531622	-0.002301021595707 16	0.000492171109399523	-0.000689914199866043	0.00276458838193591	0.00158250307267035	-0.0161030982630352	8.59931428207112E-05
2	225	X2.4.DB	PYMETROZINE	-0.000759762879754 21	0.000787335606840 48	-2.75834400019118E-05	0.00221294479370996	-0.002717792561113 11	0.000504837054487508	-0.001453192626871 39	0.00193044624135699	-2.75834400019118E-05	-0.0116643100075713	0.00238568209786905
2	246	X2.4.DB	SPIROMESIFEN	-0.000711269067251555	-0.00223057296116114	0.00294180970572544	-0.00259835011307215	-0.003066896938993 64	0.00566521472937853	0.00330958685763644	0.00529743757746753	0.00294180970572544	-0.0134696114590896	0.000536495480233066

Step: Filtering

Filter to obtain rows with p-values <0.05

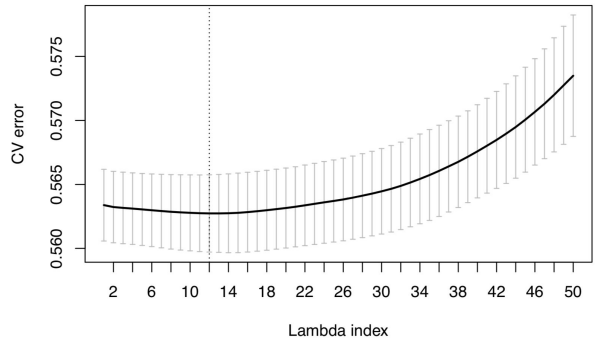
- *coef_table[coef_table$p_value <0.05,]*
- Obtained 235 pairs

Challenge:

- Hard to interpret
- Complicated in terms of further filtering work

Step: Model 2

- *numLevels = rep(1, 282) # 1 for continuous variables*
- *glinternet.cv(X, Y, numLevels,, nFolds = 3, family = "binomial")*
- X from approach 1 (continuous variables)



Step: Model 2

- Found 20 pairs from Glinternet Model
- Filtered to get 8 pairs with p-values < 0.05

	Att_1_index	Att_2_index	Att_1_name	Att_2_name
2	8	226	ACIFLUORFEN	PYRACLOSTROBIN
3	8	238	ACIFLUORFEN	ROTENONE
5	49	164	CLORANSULAM.METHYL	MCPB
12	97	238	ETHOPROPHOS	ROTENONE
14	129	214	FORAMSULFURON	PIPERONYL.BUTOXIDE
15	159	257	MALEIC.HYDRAZIDE	TEBUTHIURON
17	164	238	MCPB	ROTENONE
20	194	207	NICOSULFURON	PETROLEUM.DISTILLATE

	Att_1_main_coefs	Att_2_main_coefs	interaction_coef	p_value
2	-2.251375e-05	-7.348082e-08	-6.105576e-09	1.007386e-02
3	-2.251375e-05	4.367913e-04	6.949447e-07	6.830684e-05
5	-1.151330e-04	-9.258519e-05	9.338570e-06	4.029598e-02
12	4.475148e-07	4.367913e-04	-3.844440e-08	8.350269e-03
14	2.905877e-05	-2.185850e-05	-6.479625e-06	6.410331e-03
15	-1.938417e-06	-5.433801e-06	1.664764e-08	1.180398e-03
17	-9.258519e-05	4.367913e-04	2.114675e-07	3.707409e-03
20	1.671414e-04	1.004280e-06	-4.786234e-08	3.267132e-02

Step: Model 3

1. Glinternet:

	Att_1_index	Att_2_index	Att_1_name	Att_2_name	Att_1_main_coefs	Att_2_main_coefs	interaction_coef	int_coef_glm	p_value
19	3	263	X6.BENZYLADENINE	THIACLOPRID	-0.0404949513371709	0.637291332637374	0.00336265303593312	0.0132621892088431	0.00903288837263681
50	10	19	ALDICARB	BACILLUS.SUBTILIS	0.00764768789941734	-0.0383059287795336	0.0202279875957431	0.00652442657846244	0.0494040444902521
51	10	101	ALDICARB	FENARIMOL	0.00764768789941734	0.0389589171782285	0.0168522077384913	0.00766550284120469	0.0237372536395864
70	14	57	AZADIRACHTIN	CUPROUS.OXIDE	0.0259150790343004	-0.118797402383505	0.0230772982848491	0.0120807348413047	0.00285416773001247
104	18	229	BACILLUS.PUMILIS	PYRETHRINS	1.34950368609536	0.0184025637809732	0.00126492304549287	0.0172245928059466	0.0163292518721553
200	31	156	BUTYLATE	LINDANE	0.268330980202988	-0.0771621263913669	0.0217861317172191	0.0113564854288167	0.0139136997684876
203	31	258	BUTYLATE	TEFLUTHRIN	0.268330980202988	0.0628327470641632	0.0480320348471713	0.00997165884687693	0.0319349383958483

   - 332 pairs
- Filtered by interaction coefficient > 0
   - 27 pairs

	X	Att_1_index	Att_2_index	Att_1_name	Att_2_name	Att_1_main_coefs	Att_2_main_coefs	interaction_coef
51	51	10	101	ALDICARB	FENARIMOL	0.00764768789941734	0.0389589171782285	0.0168522077384913
568	568	111	272	FLONICAMID	TRIASULFURON	-0.0221289721387776	0.169072303579421	0.00643109911694891
684	684	153	198	KAOLIN.CLAY	OXAMYL	-0.112984681672464	-0.0587413747380037	0.0874682157739094
822	822	218	232	PROHEXADIONE	PYRIPROXYFEN	-0.136820166342337	-0.138967605358499	0.0131294500862278

Adjust the p-values from gls:

- Adding the adjusted p-values to the table of 881 pairs

  > library(fdrtool)

  > fdr = fdrtool(table$p_value, statistic="pvalue")

  > table[,'p_adjusted'] <- fdr$pval

- Found 499 pairs that are significant
- Filtering by interaction coefficients > 0
  - Found 56 pairs with synergistic effects
- Matching with main effects > 0
  - Finally obtained 4 pairs with an increasing risk effect on ALS synergistically

att1_main_glm	att2_main_glm	int_coef_glm	p_value	p_adjusted
0.000836394205032688	0.00924589039942902	0.00766550284120469	0.0237372536395864	0.0237372536395864

Results:

- Using the package “tableone” to check how our data performed after the train-test-split
- Our data looks robust, as most of them have more control cases than ALS cases

	Test Set			Train Set		
	Level of predictors	ALS_stat = 0	ALS_stat = 1	Level of predictors	ALS_stat = 0	ALS_stat = 1
Counts		27,496	9,080		51,101	17,119
ALDICARB (%)	0	14150 (51.5)	4526 (49.8)	0	25949 (50.8)	8749 (51.1)
	1	6578 (23.9)	2341 (25.8)	1	12242 (24.0)	4276 (25.0)
	2	6768 (24.6)	2213 (24.4)	2	12910 (25.3)	4094 (23.9)
FLONICAMID (%)	0	14972 (54.5)	4782 (52.7)	0	27981 (54.8)	9345 (54.6)
	1	6252 (22.7)	2137 (23.5)	1	11483 (22.5)	3948 (23.1)
	2	6272 (22.8)	2161 (23.8)	2	11637 (22.8)	3826 (22.3)
KAOLIN.CLAY (%)	0	11350 (41.3)	3445 (37.9)	0	20970 (41.0)	6595 (38.5)
	1	8113 (29.5)	2751 (30.3)	1	15172 (29.7)	5142 (30.0)
	2	8033 (29.2)	2884 (31.8)	2	14959 (29.3)	5382 (31.4)
PROHEXADIONE (%)	0	13941 (50.7)	4522 (49.8)	0	26069 (51.0)	8454 (49.4)
	1	6848 (24.9)	2153 (23.7)	1	12553 (24.6)	4261 (24.9)
	2	6707 (24.4)	2405 (26.5)	2	12479 (24.4)	4404 (25.7)
FENARIMOL (%)	0	10233 (37.2)	3194 (35.2)	0	19187 (37.5)	5884 (34.4)
	1	8670 (31.5)	2800 (30.8)	1	15808 (30.9)	5638 (32.9)
	2	8593 (31.3)	3086 (34.0)	2	16106 (31.5)	5597 (32.7)
TRIASULFURON (%)	0	19738 (71.8)	6264 (69.0)	0	36478 (71.4)	12156 (71.0)
	1	3854 (14.0)	1413 (15.6)	1	7292 (14.3)	2440 (14.3)
	2	3904 (14.2)	1403 (15.5)	2	7331 (14.3)	2523 (14.7)
OXAMYL (%)	0	3832 (13.9)	1145 (12.6)	0	7155 (14.0)	2081 (12.2)
	1	11921 (43.4)	3877 (42.7)	1	21955 (43.0)	7521 (43.9)
	2	11743 (42.7)	4058 (44.7)	2	21991 (43.0)	7517 (43.9)
PYRIPROXYFEN (%)	0	11379 (41.4)	3657 (40.3)	0	20990 (41.1)	6684 (39.0)
	1	8116 (29.5)	2649 (29.2)	1	15028 (29.4)	5227 (30.5)
	2	8001 (29.1)	2774 (30.6)	2	15083 (29.5)	5208 (30.4)

Summary

Advantages:

- The analytical results can drastically save the drug development budget
- 282 toxins = 39621 pairs
- The model efficiently helps to limit the target pairs that worth studying on

Limitations:

- Due to high proportion of missing values:
  - 164 toxins were eliminated and not considered during the analysis
  - < 50% of the 282 toxins have less than 50% NA's
  - Loss of information due to data categorization

Thank you

Mitsubishi Project:

- Professor Jiang Gui
  - Associate Professor of Biomedical Data Science
  - Associate Professor of The Dartmouth Institute
  - Associate Professor of Community and Family Medicine
- Professor Angeline S. Andrew
  - Associate Professor of Neurology

Capstone Program:

- Dr. Jennifer A. Emond
- Dr. Aurora Drew

References

1. What is ALS?, ALS Association, https://www.als.org/understanding-als/what-is-als
2. Su FC, Goutman SA, Chernyak S, et al. Association of Environmental Toxins With Amyotrophic Lateral Sclerosis [published correction appears in JAMA Neurol. 2017 May 1;74(5):612]. *JAMA Neurol*. 2016;73(7):803-811. doi:10.1001/jamaneurol.2016.0594.
3. What is RADICAVA® (edaravone)? https://www.radicava.com/patient/learn/understanding-radicava/
4. Cox PA, Davis DA, Mash DC, Metcalf JS, Banack SA. Dietary exposure to an environmental toxin triggers neurofibrillary tangles and amyloid deposits in the brain. Proc Roy Soc B 2016;283(1823).
5. Haley Otman, Pesticide Exposure May Be ALS Risk Factor, Michigan Health Lab, May 09, 2016, https://labblog.uofmhealth.org/lab-report/pesticide-exposure-may-be-als-risk-factor.
6. Rush T, Liu XQ, Lobner D. Synergistic toxicity of the environmental neurotoxins methylmercury and b-N-methylamino-L-alanine. NeuroReport 2012, 23:216–219.
7. Symphony Integrated Dataverse® (IDV®) https://symphonyhealth.prahs.com/product/idv/