Biomedical and Health Informatics Project:

Searching for Pathology with Metabolomics using Machine Learning with Bayesian Networks

Author: Vincent Preikstas

Advisor: Dr. Isabelle Bichindaritz

6/22/2021

Tables of Contents

Abstract:

The goal of this project was to utilize Bayesian Belief Networks which is a machine learning algorithm to find connections between the heavy metals cadmium, lead and mercury, several health measures, several confounding factors and measured metabolites of various types. These metabolites included various amino acids, glucids and lipids.

This exploration was already begun by a graduate student at SUNY Oswego named Jessica around 2018. She did various statistical analysis of data gathered from a study of the effects of various heavy metals on the blood proteome in children. That study had been conducted from 2016-2018 by Brooks Gump and Kestutis Bendinskas. That same data from Syracuse was used in this project.

Through the work this past year Jessica's data has been summarized and findings from her data were used to create datasets used in machine learning with a tool called Genie. Using Genie the data was normalized and three Bayesian Nets were created, one for each of the metals, that shows links discovered from the data using Bayesian Search.

Introduction:

The toxicity of heavy metals is a widely explored and documented subject within the scientific community. Having high levels of metals such as lead, mercury or cadmium in the blood can have adverse health effects and can lead to disease or other defects. Many of the mechanisms of these different metal toxicities are understood, however there are always deeper levels of association to consider. People can have varying reactions or outcomes related to exposure to heavy metals and this is not always well understood. One area of association that needs further investigation is the relation between the numerous metabolites in the blood and heavy metal toxicity.

Using machine learning techniques, it is possible to see more complex associations between blood levels of metabolites, metals, and pathology than with more traditional statistical analysis. This is an exploratory study that will look to see what relationships exist between various metabolite levels, heavy metal levels and health metrics. Showing these associations is a first step in providing proof of different metabolites either enhancing or inhibiting the toxicity of various heavy metals in humans.

Background and Significance:

Although metal toxicity is generally in the domain of common knowledge, seeking out some literature to corroborate proved important. Along with the basic background of heavy metal's part in disorders, papers that attempt to find mediators or limiting factors for the toxicity were also sought.

A paper by Robert Birdsall called Effects of Lead and Mercury on The Blood Proteome of Children was published in J Proteome Res in 2010. The study was exploring heavy metal exposure in children in association with physiological and neurological problems. The goal was to enhance understanding of biochemical interactions that cause health problems associated with mercury and lead at levels of exposure below CDC guidelines. To this end researchers used proteomics to analyze blood samples from 34 children who were depleted of certain proteins using antibody-based affinity columns. They found that Apolipoprotein E had an inverse significant association with lead concentration, which they state reinforces prior findings that Apolipoprotein E genotype is a mediator of neurobehavioral effects in individuals with lead exposure (Birdsall, 2010).

To establish further foundation for the toxicity of heavy metals an article from Web MD titled Lead Levels Linked to Blood Pressure was reviewed. Written in 2003 about an original study submitted to The Journal of The American Medical Association, it outlines the risk to postmenopausal woman exposed to lead levels even well below the FDA standards for safety to have heightened blood pressure. This in turn can lead to hypertension which can cause death if left untreated. The results of the study showed that lead levels of less than 40ug/dL (US occupational blood lead exposure limit) and even less than 10ug/dL (CDC prevention level for lead poisoning in children) can have effects on blood pressure of older women (WebMD, 2003).

A paper by Julio Chirinos published in the Journal of the American Heart Association looked at possible connections between heart failure and nitric oxide metabolites. They compared NOm between subjects without heart failure, subjects with heart failure and preserved ejection fraction (HFpEF) and subjects with heart failure and reduced ejection fraction (HFrEF). What they found was that HFpEF was associated with reduced plasma NOm, but not HFrEF. This indicated endothelial dysfunctions, enhanced clearance, or deficient ingestion of inorganic nitrate in the diet. They used an ANOVA test to compare their groups. These kinds of findings that measure the absence or presence of metabolites in different amounts related to measurable health markers is exactly the kind of conclusions our project is probing for (Chirinos, 2016).

The paper by Jidapa Kraisangka and Marek Druzdzel and the article posted on MachineLearningMaster.com by Jason Brownlee both provide information about the workings of Bayesian Networks and were mainly included for reference while learning and deploying our own Bayesian Network.

Keeping these outside sources in mind our study could find new correlations and more complex relationships than simple regression which could be novel information. Using more advanced machine learning techniques such as Bayesian Nets presents a lot of new opportunities to fill gaps as well as enhance understandings we glean from more traditional statistics by taking a step further.

Specific Aims:

First Project Aim:

The first goal is a summary of Jessica's findings. She did several statistical analyses looking for correlations. These included Numeric and Logistic Regressions.

Second Project Aim:

The second goal of the project will be to do additional analysis with a focus on applied machine learning to see associations or dependencies between health metrics, heavy metal levels and other metabolites. To this end a Bayesian network will be used to find conditional dependencies and conditionally independent relationships which will give some clues as to the relationships between different metabolites and metal toxicity which is being explored.

Approach/Methods:

First Goal Method:

The word document analysisResultsFinalV5(1).docx was obtained from a Google Drive folder called Metabolomic that contains various data and files pertaining to the statistical work Jessica did on the Syracuse dataset. The document analysisResultsFinalV5(1), is the most complete document in the folder containing all correlation analysis results done with SPSS. This document was reviewed, and a summary of the correlations found is available in the results section of this document. Our method here was simply to review the document for comprehension and the found correlations between health measures and metals from the various analysis were used to create the more compound datasets used with the Bayesian Search Algorithm with this project's analysis.

Second Goal Method:

This goal was the main focus of the project; using Genie to create Bayesian networks by learning from the Syracuse dataset. The first task was research into similar topic areas and understanding the dataset. The Syracuse dataset consisted of only 299 subject entries (rows), but had over 3000 attributes (columns). This meant the dataset was extremely information rich for each subject. The contents of the dataset were health measures (many different heart health measures, food intake…, mental health metrics, etc.), confounding factors (smoking surveys, socioeconomic score, demographic information), targeted metabolomic data (counts of concentration of amino acids, lipids, glucids), large amounts of untargeted metabolomic data (metabolite compounds not readily named but identified) and of course metal levels.

The first task was pruning the dataset into smaller subsets pertaining to each metal to do learning on as using all 3000+ attributes was not practical. The subsequent datasets contained all confounding factors, health measures for which the metal and a given measure had correlation (taken from Jessica's work), the metal measure itself and the targeted metabolites. Below you can find each dataset and the attributes that were included.

Lead Dataset: Subnum, Pb, "gender 1 = Male, 2 = Female", "race 1 = African American, 2 = Caucasion", childage, sesscore, rawbmi, SMK1c, SMK1, TCmgdL (Total cholesterol measured in milligrams per decilitre), HDLmgdL (High density lipoprotiens measured in milligrams per decilitre), nonHDLmgdL (Non high density lipoproteins measured in milligrams per decilitre), GlumgdL (Glucose measured in milligrams per decilitre), Alanine, Sarcosine, Glycine, alpha_aminoisobutyric_acid, Valine, Leucine, Isoleucine, Threonine, Serine, Proline, Asparagine, Aspartic_Acid, Methionine, Glutamic_Acid, Phenylalanine, Glutamine, Ornithine, Lysine, Histidine, Tyrosine, Typtophan, Stearicacid, Oleicacid, Hexoseglucoseetc, Palmiticacid, Creatine, Hypoxanthine, Gluconate, IMP, AMP, UDPNacetylDglucosamine, AMP_A, FAD, FBP, NADH, NADP, Phosphoenolpyruvate, s7P, SUCResults, MALResults, @6PGResults, CITICITResults, ADPResults, ATPResults, Cortisol, Cortisone, @11Deoxycortisol, Corticosterone, @18Hydroxycortisol, RVDiasDimen, LVsSysDimen, LVSysVolum, dninevlf, meanvlf, CMcynicismChild, CMtotalchild, DBDODDscore, dbdhyperscore, DBDadhdhyp, RBPptotal, CSIparentTot, CSdef, N_Disorder, CSIparentND, MirrorPer

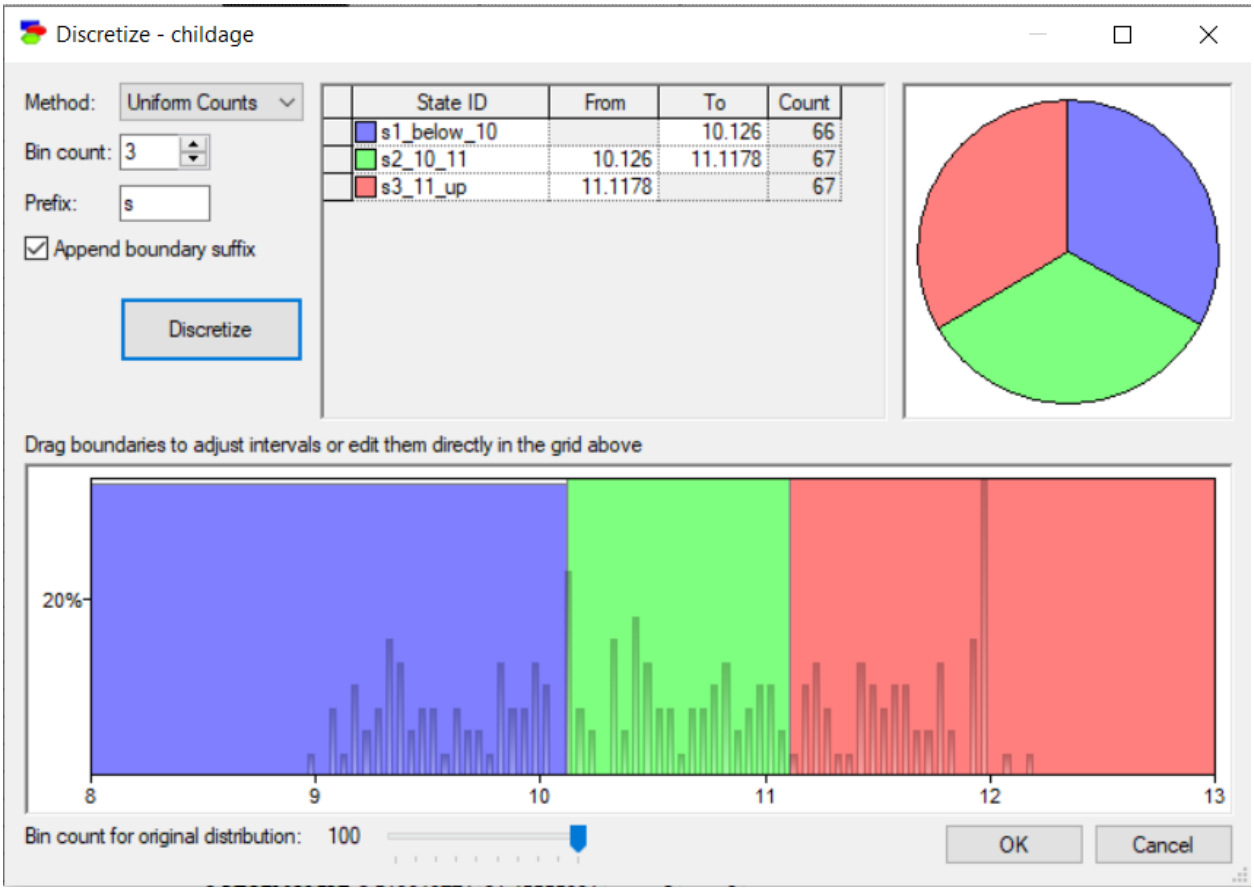
Mercury Dataset: Subnum, Hg, "gender 1 = Male, 2 = Female", "race 1 = African American, 2 = Caucasion", childage, sesscore, rawbmi, SMK1c, SMK1, TCmgdL (Total cholesterol measured in milligrams per decilitre), HDLmgdL (High density lipoprotiens measured in milligrams per decilitre), nonHDLmgdL (Non high density lipoproteins measured in milligrams per decilitre), GlumgdL (Glucose measured in milligrams per decilitre), Alanine, Sarcosine, Glycine, alpha_aminoisobutyric_acid, Valine, Leucine, Isoleucine, Threonine, Serine, Proline, Asparagine, Aspartic_Acid, Methionine, Glutamic_Acid, Phenylalanine, Glutamine, Ornithine, Lysine, Histidine, Tyrosine, Typtophan, Stearicacid, Oleicacid, Hexoseglucoseetc, Palmiticacid, Creatine, Hypoxanthine, Gluconate, IMP, AMP, UDPNacetylDglucosamine, AMP_A, FAD, FBP, NADH, NADP, Phosphoenolpyruvate, s7P, SUCResults,

MALResults, @6PGResults, CITICITResults, ADPResults, ATPResults, Cortisol, Cortisone, @11Deoxycortisol, Corticosterone, @18Hydroxycortisol, SystolicBP, AoAnnDiam, Eem1, mrhr, ninehr, Carotid_AIX, irratio, drtlf, drthf, drttotal, dbtnn50cnt, meanNUlf, RoundCBCLwdrwnmean, FATimeAvg

Cadmium Dataset: Subnum, Cd, "gender 1 = Male, 2 = Female", "race 1 = African American, 2 = Caucasion", childage, sesscore, rawbmi, SMK1c, SMK1, TCmgdL (Total cholesterol measured in milligrams per decilitre), HDLmgdL (High density lipoprotiens measured in milligrams per decilitre), nonHDLmgdL (Non high density lipoproteins measured in milligrams per decilitre), GlumgdL (Glucose measured in milligrams per decilitre), Alanine, Sarcosine, Glycine, alpha_aminoisobutyric_acid, Valine, Leucine, Isoleucine, Threonine, Serine, Proline, Asparagine,Aspartic_Acid, Methionine, Glutamic_Acid, Phenylalanine, Glutamine, Ornithine, Lysine, Histidine, Tyrosine, Typtophan, Stearicacid, Oleicacid, Hexoseglucoseetc, Palmiticacid, Creatine, Hypoxanthine, Gluconate, IMP, AMP, UDPNacetylDglucosamine, AMP_A, FAD, FBP, NADH, NADP, Phosphoenolpyruvate, s7P, SUCResults, MALResults, @6PGResults, CITICITResults, ADPResults, ATPResults, Cortisol, Cortisone, @11Deoxycortisol, Corticosterone, @18Hydroxycortisol, HeartRate, EAratio, irestHR, irestSV, irestLVET, mrhr, rtsbp, ninehr, dfnineDBP, dfninetpr, dfninemap, meanaci, dbtsdsd, dbtrmssd, dmtlfhf, drtvlf, regdrthf, zregdrthf, zregdmthf, PSp, RoundCBCLattmean, CBCLtscore, DEP_AffR, MirrorPer, FACntAvg, rthr, RoundCBCLtpmean

The data was normalized using both Microsoft Excel as well as the Genie software. The Genie software is a tool by BayesFusion LLC and the version used in this project was their academic version of the software downloaded from their website: https://www.bayesfusion.com/.

Several subjects in the study were missing the majority of their metabolomic data, these rows were deleted in excel along with some health measures that had most of the data missing. Once deletions were complete the preprocessing of the data continued in the Genie tool after exporting the excel sheets to CSV files. Genie has a number of powerful data normalization tools that include filling in missing variables as well as discretization. Most Bayesian Network algorithms require data to be discrete so each dataset had all continuous data discretized. The standard was each attribute had missing values replaced with the average value (except for smoking survey in which all missing values were replaced with “doesn’t smoke”) and the discretization was done with binning into three bins with uniform count. Below you can see the GUI interface in Genie showing discretization of ‘childage’ attribute.



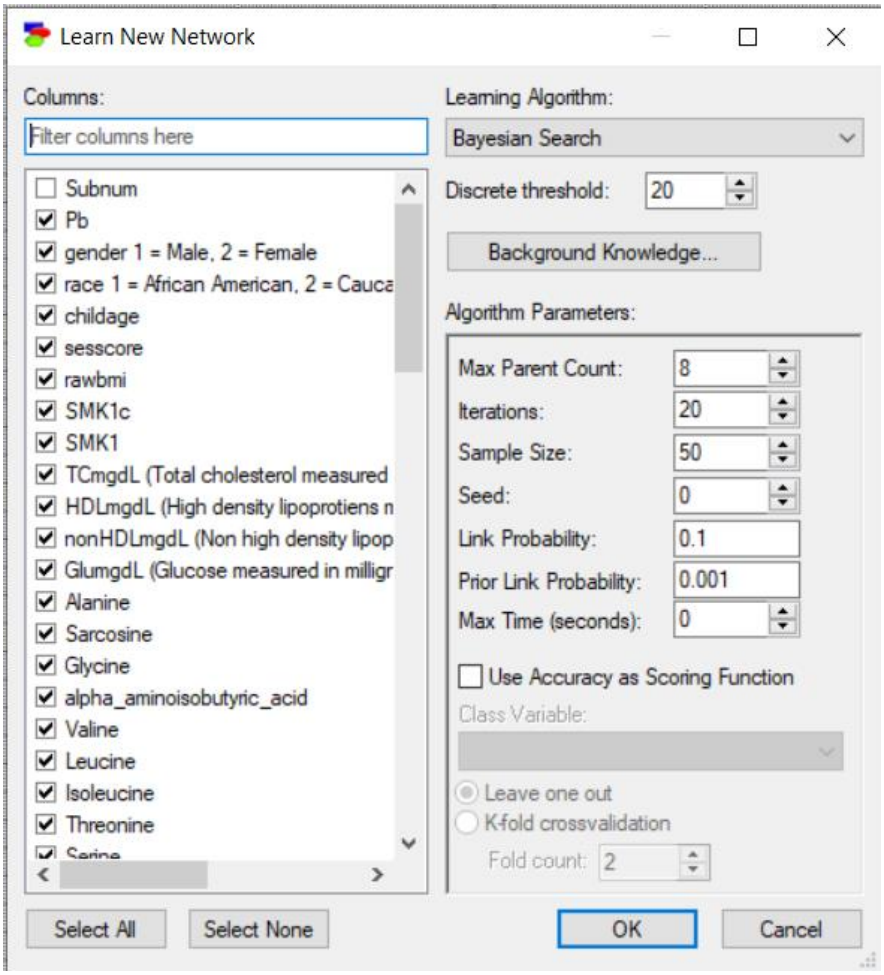
This process was somewhat time intensive as Genie lacks a method to perform this function on multiple attributes at once. Below you can see the results of this normalization and how it appears in the data table. The example is a section of the normalized Lead model data.

Tyrosine	Typtophan	Stearicacid	Oleicacid	Hexoseglucoseetc	Palmiticacid	Creatine
s3_62_up	s3_50_up	s2_53975532_70449192	s1_below_220410832	s1_below_15469916	s1_below_138099712	s2_7755159_9461317
s2_50_62	s3_50_up	s3_70449192_up	s3_332619456_up	s3_19624456_up	s3_178151120_up	s2_7755159_9461317
s3_62_up	s3_50_up	s3_70449192_up	s2_220410832_332619456	s3_19624456_up	s2_138099712_178151120	s2_7755159_9461317
s3_62_up	s3_50_up	s3_70449192_up	s2_220410832_332619456	s1_below_15469916	s2_138099712_178151120	s2_7755159_9461317
s1_below_50	s1_below_40	s3_70449192_up	s3_332619456_up	s3_19624456_up	s3_178151120_up	s3_9461317_up
s3_62_up	s3_50_up	s2_53975532_70449192	s2_220410832_332619456	s1_below_15469916	s2_138099712_178151120	s3_9461317_up
s2_50_62	s3_50_up	s3_70449192_up	s1_below_220410832	s3_19624456_up	s2_138099712_178151120	s2_7755159_9461317
s3_62_up	s3_50_up	s3_70449192_up	s3_332619456_up	s1_below_15469916	s3_178151120_up	s3_9461317_up
s3_62_up	s3_50_up	s1_below_53975532	s1_below_220410832	s3_19624456_up	s2_138099712_178151120	s3_9461317_up
s3_62_up	s3_50_up	s2_53975532_70449192	s3_332619456_up	s3_19624456_up	s3_178151120_up	s2_7755159_9461317
s1_below_50	s2_40_50	s1_below_53975532	s1_below_220410832	s1_below_15469916	s1_below_138099712	s3_9461317_up
s1_below_50	s1_below_40	s1_below_53975532	s1_below_220410832	s2_15469916_19624456	s1_below_138099712	s3_9461317_up
s2_50_62	s3_50_up	s3_70449192_up	s1_below_220410832	s3_19624456_up	s2_138099712_178151120	s3_9461317_up
s1_below_50	s2_40_50	s3_70449192_up	s2_220410832_332619456	s3_19624456_up	s2_138099712_178151120	s3_9461317_up
s3_62_up	s1_below_40	s3_70449192_up	s3_332619456_up	s3_19624456_up	s3_178151120_up	s3_9461317_up
s1_below_50	s1_below_40	s3_70449192_up	s3_332619456_up	s3_19624456_up	s3_178151120_up	s3_9461317_up
s2_50_62	s3_50_up	s2_53975532_70449192	s1_below_220410832	s3_19624456_up	s1_below_138099712	s2_7755159_9461317
s3_62_up	s3_50_up	s3_70449192_up	s3_332619456_up	s1_below_15469916	s3_178151120_up	s3_9461317_up
s3_62_up	s3_50_up	s2_53975532_70449192	s2_220410832_332619456	s3_19624456_up	s2_138099712_178151120	s2_7755159_9461317
s3_62_up	s3_50_up	s2_53975532_70449192	s2_220410832_332619456	s1_below_15469916	s2_138099712_178151120	s1_below_7755159
s3_62_up	s1_below_40	s1_below_53975532	s2_220410832_332619456	s3_19624456_up	s2_138099712_178151120	s2_7755159_9461317
s3_62_up	s1_below_40	s1_below_53975532	s1_below_220410832	s3_19624456_up	s1_below_138099712	s2_7755159_9461317
s3_62_up	s3_50_up	s2_53975532_70449192	s2_220410832_332619456	s3_19624456_up	s2_138099712_178151120	s2_7755159_9461317
s2_50_62	s1_below_40	s1_below_53975532	s2_220410832_332619456	s3_19624456_up	s2_138099712_178151120	s2_7755159_9461317
s1_below_50	s1_below_40	s2_53975532_70449192	s2_220410832_332619456	s1_below_15469916	s2_138099712_178151120	s3_9461317_up
s3_62_up	s1_below_40	s3_70449192_up	s3_332619456_up	s3_19624456_up	s3_178151120_up	s2_7755159_9461317
s1_below_50	s1_below_40	s3_70449192_up	s3_332619456_up	s3_19624456_up	s3_178151120_up	s2_7755159_9461317
s2_50_62	s1_below_40	s1_below_53975532	s1_below_220410832	s1_below_15469916	s1_below_138099712	s1_below_7755159
s3_62_up	s3_50_up	s2_53975532_70449192	s2_220410832_332619456	s3_19624456_up	s1_below_138099712	s2_7755159_9461317
s2_50_62	s1_below_40	s2_53975532_70449192	s1_below_220410832	s3_19624456_up	s1_below_138099712	s2_7755159_9461317
s1_below_50	s2_40_50	s1_below_53975532	s1_below_220410832	s1_below_15469916	s1_below_138099712	s2_7755159_9461317
s3_62_up	s3_50_up	s2_53975532_70449192	s1_below_220410832	s3_19624456_up	s2_138099712_178151120	s3_9461317_up
s1_below_50	s2_40_50	s3_70449192_up	s2_220410832_332619456	s1_below_15469916	s2_138099712_178151120	s2_7755159_9461317
s2_50_62	s1_below_40	s2_53975532_70449192	s2_220410832_332619456	s1_below_15469916	s2_138099712_178151120	s1_below_7755159
s3_62_up	s3_50_up	s2_53975532_70449192	s2_220410832_332619456	s3_19624456_up	s2_138099712_178151120	s3_9461317_up
s2_50_62	s2_40_50	s1_below_53975532	s1_below_220410832	s1_below_15469916	s1_below_138099712	s2_7755159_9461317
s3_62_up	s3_50_up	s2_53975532_70449192	s2_220410832_332619456	s1_below_15469916	s2_138099712_178151120	s3_9461317_up
s1_below_50	s1_below_40	s3_70449192_up	s3_332619456_up	s3_19624456_up	s3_178151120_up	s2_7755159_9461317
s2_50_62	s2_40_50	s2_53975532_70449192	s2_220410832_332619456	s3_19624456_up	s2_138099712_178151120	s2_7755159_9461317
s2_50_62	s1_below_40	s2_53975532_70449192	s1_below_220410832	s3_19624456_up	s2_138099712_178151120	s3_9461317_up
s2_50_62	s2_40_50	s3_70449192_up	s3_332619456_up	s1_below_15469916	s3_178151120_up	s3_9461317_up
s3_62_up	s3_50_up	s3_70449192_up	s2_220410832_332619456	s3_19624456_up	s2_138099712_178151120	s3_9461317_up
s3_62_up	s2_40_50	s2_53975532_70449192	s2_220410832_332619456	s1_below_15469916	s2_138099712_178151120	s2_7755159_9461317
s1_below_50	s1_below_40	s3_70449192_up	s3_332619456_up	s1_below_15469916	s1_below_138099712	s1_below_7755159
s1_below_50	s2_40_50	s2_53975532_70449192	s3_332619456_up	s2_15469916_19624456	s3_178151120_up	s3_9461317_up
s3_62_up	s3_50_up	s3_70449192_up	s3_332619456_up	s3_19624456_up	s3_178151120_up	s3_9461317_up
s2_50_62	s3_50_up	s3_70449192_up	s3_332619456_up	s3_19624456_up	s3_178151120_up	s2_7755159_9461317

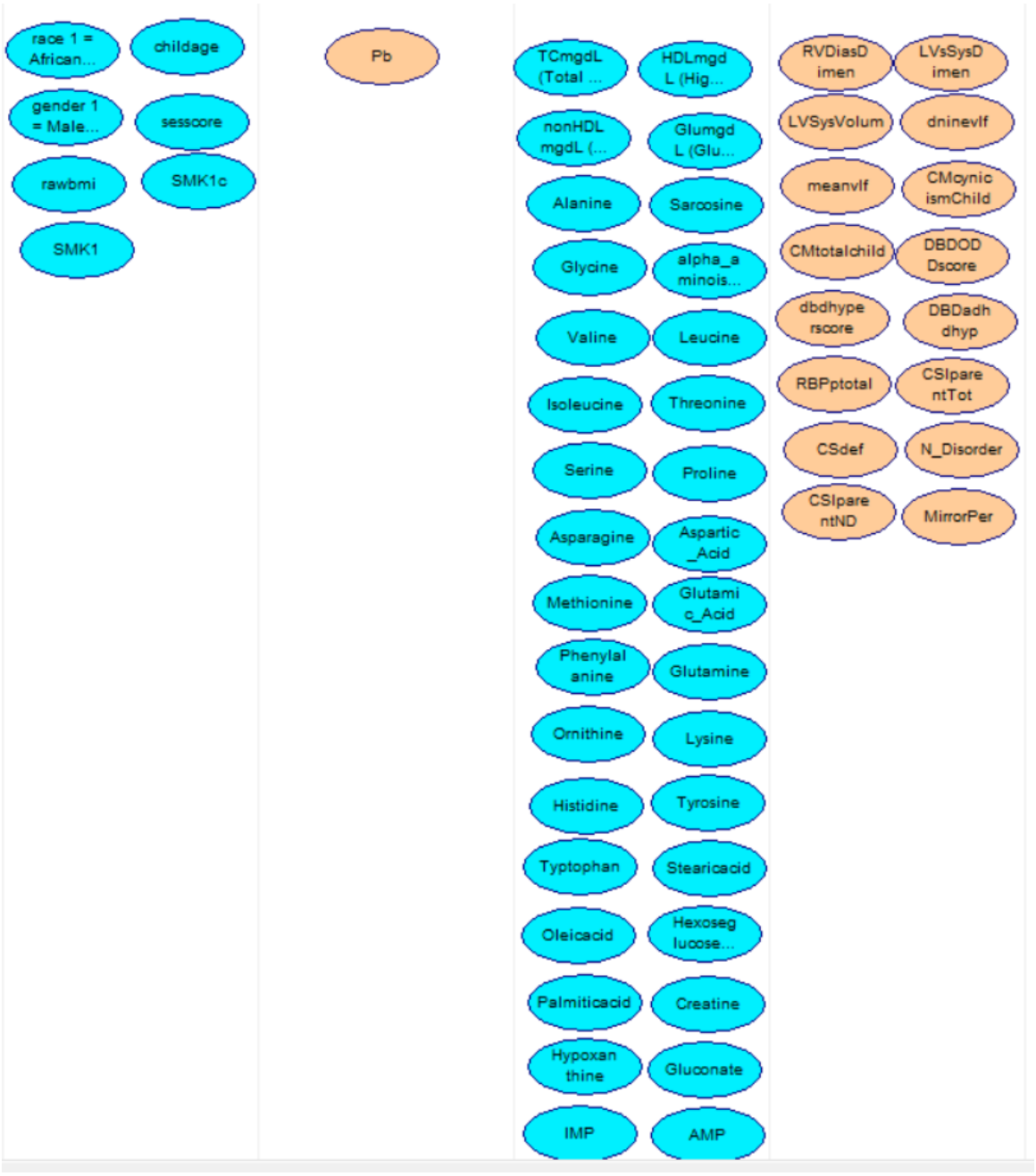
Genie automatically creates ranges for the user and although the labels created are not very readable, they are functional.

The next step of the process was now to create the Bayesian Belief Network. A Bayesian net is an acyclic directional graph of probability tables that can be used to model any kind of data or situation. Another way to think of it is as a probabilistic graphical model with nodes that have directed edges, these patterns model both conditionally dependent and conditionally independent relationships between variables. The real power that a Bayesian Net has over more traditional correlation is that the probabilities and connections are causal and not simply correlation. Calculating these nets and the tables at each node can become daunting so computer programs are used to both do the calculations and apply certain optimizations/ assumptions (Brownlee, 2019). For our project we wanted to learn the network from the data to hopefully find interesting connections between the metals, metabolites, health measures and confounding factors that show causation. To this end we used the Bayesian Search

algorithm which is the default network learning algorithm in Genie. We used all default parameters which you can see below.



We also provided background knowledge for each of the three models which all looked similar to an example you can see below.



This allowed for some control over the directionality of the graph by enforcing a timing element. The idea being that the earlier temporal tiers come first. In our case it was confounding factors, metal levels, metabolites and then health measures. This means that each temporal tier can have a directional relationship with a later tier and its own tier, but it cannot point to a previous tier. This is the best we could do to enforce our prior knowledge and assumptions about the relationship between the various types of attributes.

Having set the prior knowledge and algorithm for learning the program was allowed to learn from the data and the resulting networks can be found in the results section. As a final step the networks were organized visually to reflect the background knowledge.

Results:

First Project Goal Result:

Jessica calculated and recorded various metrics that can be found in the analysisResultsFinalV5(1) document. These include things such as basic descriptive statistics on many attributes (means, std. Deviation, etc) as well as different types of box plots and visuals. This document will not restate these sorts of results from Jessica's work, this doc will instead list attributes for various tests of correlation. These were important to this study as well as being the larger part of Jessica's work. For all detailed information untruncated with SPSS output images please see analysisReesultsFinalV5(1). Only results that were significant ($p < 0.05$) are included here. Significant attributes are included in the following format < <attribute> (<significance>) >.

Chi Squared Analysis:

Race and Lead (0.000), Race and Mercury (0.001)

ANOVA Analysis:

Binned Lead and SES Score (0.001)

Correlation Analysis:

Cadmium Correlations (with extra variables race, gender, sesscore, bmipct, childage and SMK1c): HeartRate (0.047), EAratio (0.027), irestHR (0.008), irestSV (0.020), irestLVET (0.002), mrhr (0.041), ninehr (0.007), dfnineDBP (0.048), dfninetpr (0.034), meanaci (0.029), lasted12mnths (0.009), gumchewing (0.005), dbtsdsd (0.040), dbtrmssd (0.049), dmtlfhf (0.049), drtvlf (0.043), regdrthf (0.046), dsci1pnn50 (0.000), dsci2pnn50 (0.000), dsci2lfhf (0.019), dsci3pnn50 (0.001), dsci3lfhf (0.042), dsci4se (0.015), dsci4nn50cnt (0.043), dsci4pnn50 (0.018), dsci4lf (0.012), dsci4hf (0.017), dsci4total (0.010), zregdrthf (0.046), zregdmthf (0.046), PSp (0.039), RoundCBCLattmean (0.015), CBCLtscore (0.031), DEP_AffR (0.029), MirrorPer (0.019), FACntAvg (0.045), Glutamic_Acid (0.049), Ornithine (0.042), Lysine (0.031), IMP (0.039), UDP-N-acetyl-D-glucosamine (0.021), FBP (0.022), MAL Results (0.032), ATP Results (0.030), rawbmi (0.002).

Lead Correlations (with extra variables race, gender, sesscore, bmipct, childage and SMK1c): Pb (0.000), RVDiasDimen (0.007), LVsSysDimen (0.015), Filterswater (0.006), oilpaint (0.045), howoften (0.028), dninevlf (0.045), dsci2nn50cnt (0.018), dsco2pnn50 (0.005), dsci2vlf (0.020), dsci3nn50cnt (0.043), dsci3pnn50 (0.033), dsci3lf (0.028), dsci4nn50cnt (0.022), dsci4vlf (0.042), meanvlf (0.034), PerTotal (0.000), CMcynicismChild (0.002), CMtotalchild (0.015), DBDODDscore (0.041), DBDodd (0.021), DBDadhdhyp (0.022), RBPptotal (0.016), CSIparentND (0.000), CSIparentTot (0.004), CSdef (0.024), MirrorPer (0.002), N_Disorder (0.019), Cortisol (0.008), Cortisone (0.035), Corticosterone (0.017), SMK1p (0.000), SMK6ap (0.036), SMK6bp (0.000), SMK7p (0.019), SMK9p (0.003).

Mercury Correlation (with extra variables race, gender, sesscore, bmipct, childage and SMK1c): SystolicBP (0.040), AoAnnDiam (0.038), Eem1 (0.000), mrhr (0.035), ninehr (0.031), coughsyrup (0.015), Carotid_AIX (0.033), irratio (0.021), drtlf (0.42), drthf (0.005), drtotal (0.010), dsci2nn50cnt (0.040), meanNUlf (0.028), RoundCBCLwdrwnmean (0.040), FATimeAvg (0.037), Tyrosine (0.045), Hexose (glucose etc.) (0.013), Hypoxanthine (0.032), Coricosterone (0.038).

Numeric Regression Analysis:

Cadmium Linear Regression: rawbmi (0.002), HeartRate (0.047), EAratio (0.027), irestHR (0.008), irestSV (0.020), irestLVET (0.002), mrhr (0.041), rthr (0.022), ninehr (0.007), dfnineDBP (0.048), dfninetpr (0.034), meanaci (0.029), lasted12mnths (0.009), gumchewing (0.005), dbtsdsd (0.040), dbtrmssd (0.049), dmtlfhf (0.049), drtvlf (0.043), regdrthf (0.046), dsci1pnn50 (0.000), dsci2nn50cnt (0.012), dsci2pnn50 (0.000), dsci2lfhf (0.019), dsci3pnn50 (0.001), dsci3lfhf (0.042), dsci4se (0.015), dsci4nn50cnt (0.043), dsci4pnn50 (0.018), dsci4lf (0.012), dsci4hf (0.017), dsci4total (0.010), zregdrthf (0.046), zregdmthf (0.046), PSp (0.039), RoundCBCLattmean (0.015), CBCLtscore (0.031), DEP_AffR (0.029), MirrorPer (0.019), FACntAvg (0.045), Glutamic_Acid (0.049),Ornithine (0.042), Lysine (0.031), IMP (0.039), UDP-N-acetyl-D-glucosamine (0.021), FBP (0.022), MAL Results (0.032), ATP Results (0.030), SVstrokevolume (0.027), TPRtotalperipheralresistance (0.045) and many more untargeted metabolites, please see analysisResultsFinalV5(1) for the full table.

Lead Linear Regression: Cd (0.001), RVDiasDimen (0.007), LVsSysDimen (0.015), LVSysVolum (0.012), Filterswater (0.006), oilpaint (0.045), howoften (0.028), dninevlf (0.045), dsci2nn50cnt (0.018), dsci2pnn50 (0.005), dsci2vlf (0.020), dsci3nn50cnt (0.043), dsci3pnn50 (0.033), dsci3lf (0.028), dsci4nn50cnt (0.022), dsci4vlf (0.042), meanvlf (0.034), PerTotal (0.000), CMcynicismChild (0.002), CMtotalchild (0.015), DBDODDscore (0.010), dbdhyperscore (0.041), DBDodd (0.021), DBDadhdhyp (0.022), RBPptotal (0.016), CSIparentND (0.000), CSIparentTot (0.004), CSdef (0.024), MirrorPer (0.002), N_Disorder (0.019), Cortisol (0.008), Cortisone (0.035), Corticosterone (0.017), SMK1p (0.000), SMK6ap (0.036), SMK6bp (0.000), SMK7p (0.019) and many more untargeted metabolites, please see analysisResultsFinalV5(1) for the full table.

Logistic Regression Analysis (Hg, Pb and Cd quartiled to contain equal parts participants with Test of Between Subjects Effects):

Cd Logistic Regression: rawbmi (0.004), IVSSysthick (0.001), LVPWSysThick (0.005), LVMass (0.020), Height (0.027), Weight (0.010), BSA (0.007), irestHR (0.000), irestSV (0.003), irestLVET (0.000), sciHR (0.019), mrhr (0.000), rthr (0.001), ninehr (0.000), lasted12mnths (0.011), gumchewing (0.039), IRNN50cnt (0.017), dbtsdnn (0.024), dbtsdsd (0.028), dbtrmssd (0.033), drtvlf (0.027), dsci1pnn50 (0.028), dsci2nn50cnt (0.008), dsci2pnn50 (0.029), dsci3nn50cnt (0.021), dsci3pnn50 (0.028), dsci4sdnn (0.028), dsci4se (0.002), dsci4nn50cnt (0.013), dsci4hf (0.038), meansdsd (0.043), ShiftRS (0.022), ShiftUS (0.022), SAPQ_Total (0.029), ERtotal (0.036), RoundCBCLanxmean (0.001), RoundCBCLsommean (0.015), RoundCBCLtpmean (0.008), RoundCDCLattmean (0.004), CBCLinternal (0.004), CBCLcompC (0.012), CBCLtotalsc (0.013), CBCLtscore (0.001), CBCLtotalTsc (0.005), CShyp (0.010), Ornithine (0.042), UDP-N-acetyl-D-glucosamine (0.008) and many more untargeted metabolites, please see analysisResultsFinalV5(1) for the full table.

Pb Logistic Regression: Cd (0.001), hg2 (0.003), RVDiasDimen (0.025), IVSysthick (0.023), Filterswater (0.030), dsci2pnn50 (0.028), dsci3pnn50 (0.035), PerTotal (0.005), CMcynicismChild (0.003), CMtotalchild (0.018), Income (0.040), DBDODDscore (0.026), DBDodd (0.017), DBDadhdhyp (0.045), RBPptotal (0.032), MEIM_EXPL (0.019), MEIM_Total (0.040), ADP Results (0.005), Corticosterone (0.035), SMK1p (0.001), SMK6ap (0.003), SMK6bp (0.023), SMK6dp (0.030), SMK7p (0.015), SMK9p (0.042) and many more untargeted metabolites, please see analysisResultsFinalV5(1) for the full table.

Hg Logistic Regression: SystolicBP (0.009), Eem2 (0.002), Eem1 (0.003), irestLVET (0.034), meannormlf (0.006), meannormhf (0.006), meanratio (0.010), irratio (0.037), dratio (0.030), drtlf (0.020), drthf (0.001), drttotal (0.003), dsci2nn50cnt (0.025), meantotal (0.042), meanlfhf (0.016), meanNUlf (0.028), meanhf (0.015), dninelfhf (0.003), DERSstrategies (0.031), FATimeAvg (0.011), childsesladder2 (0.038), Threonine (0.024), Serine (0.037), Phenylalanine (0.033), Tyrosine (0.018), SUC

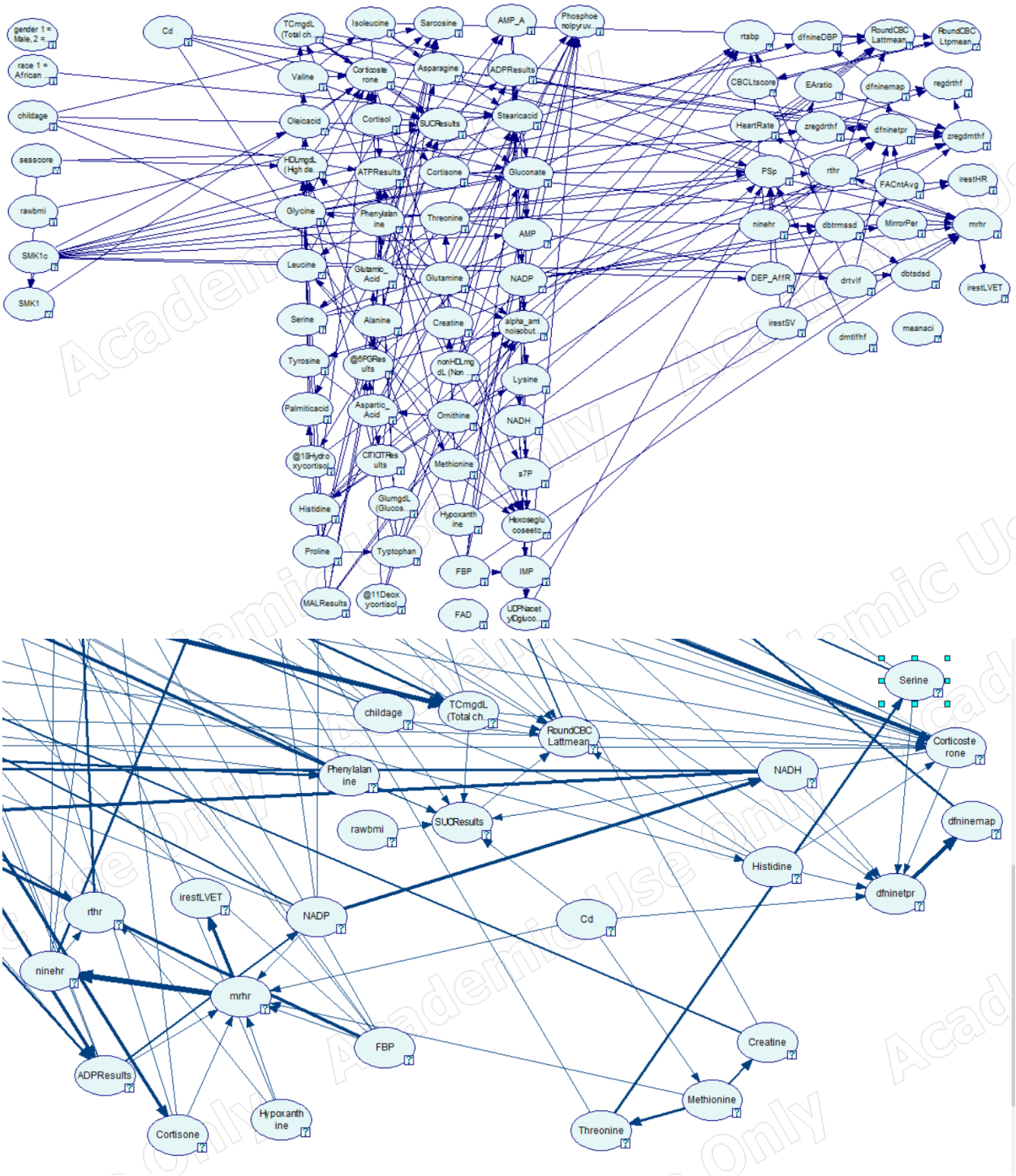
Results (0.015), SMK6bp (0.033) and many more untargeted metabolites, please see analysisResultsFinalV5(1) for the full table.

There were also sections with variables tested using Parameter Estimates for Cd, Pb and Hg. For these results see analysisResultsFinalV5(1) after each metals logistic regression test of between subjects while quartiled results. For the purpose of creating Bayesian Networks the results from the Numeric Regression and Correlation Analysis were used to select health measures for the datasets that were used to do learning.

Second Project Goal Result:

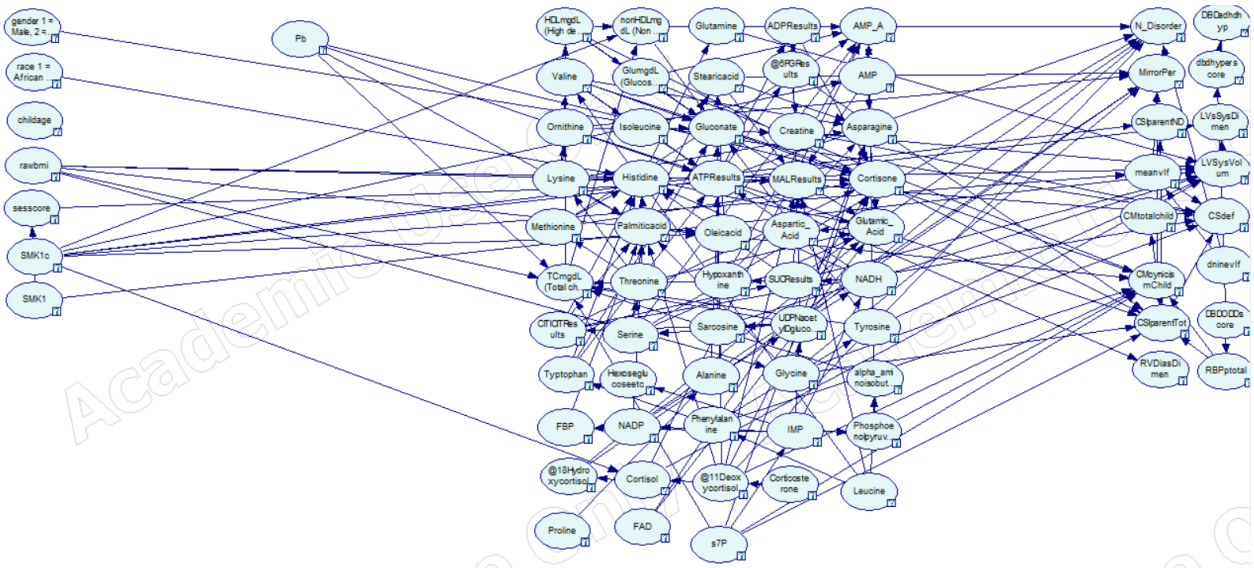
Three networks were created, one for each of the metals of interest. For each there will be a first image showing the whole network (note the network is organized by temporal tier) and then an image focused on the metal with the metal's connections and the metal's connections' connections showing influence strength.

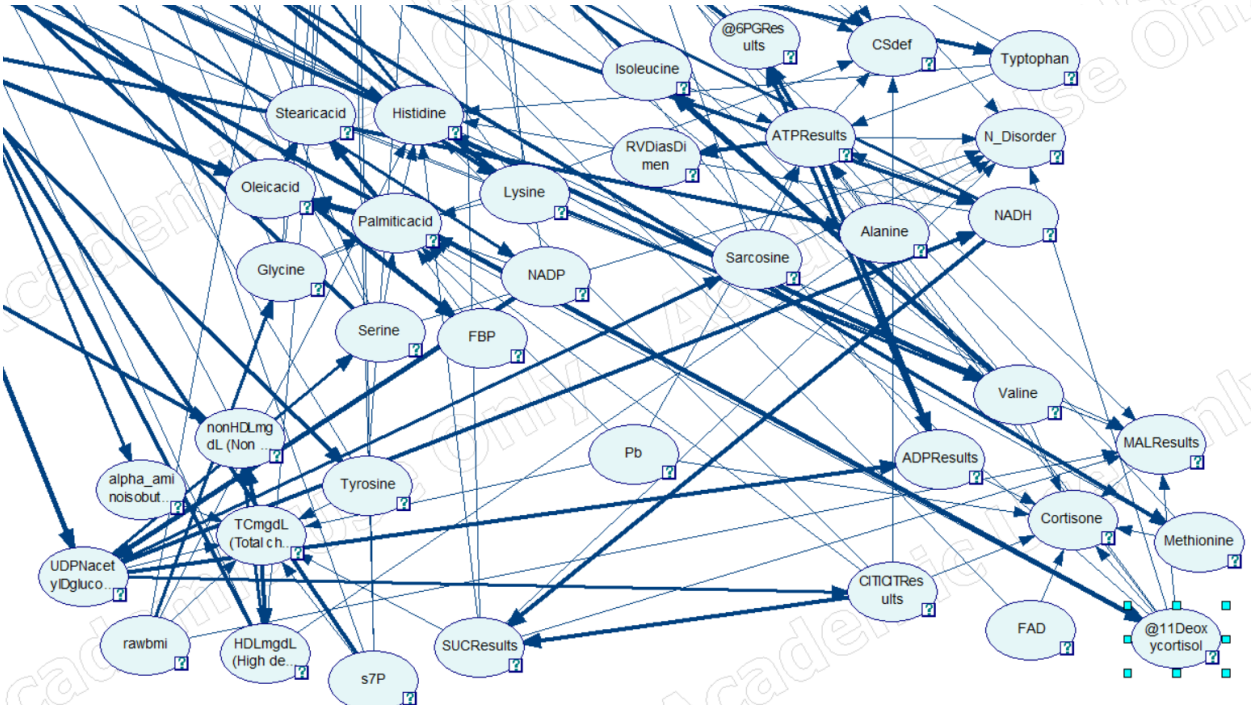
Cadmium Network:



We can see connections between Cadmium and the metabolites Methionine and SUCResults as well as with the health measures mrhr and dfninetpre.

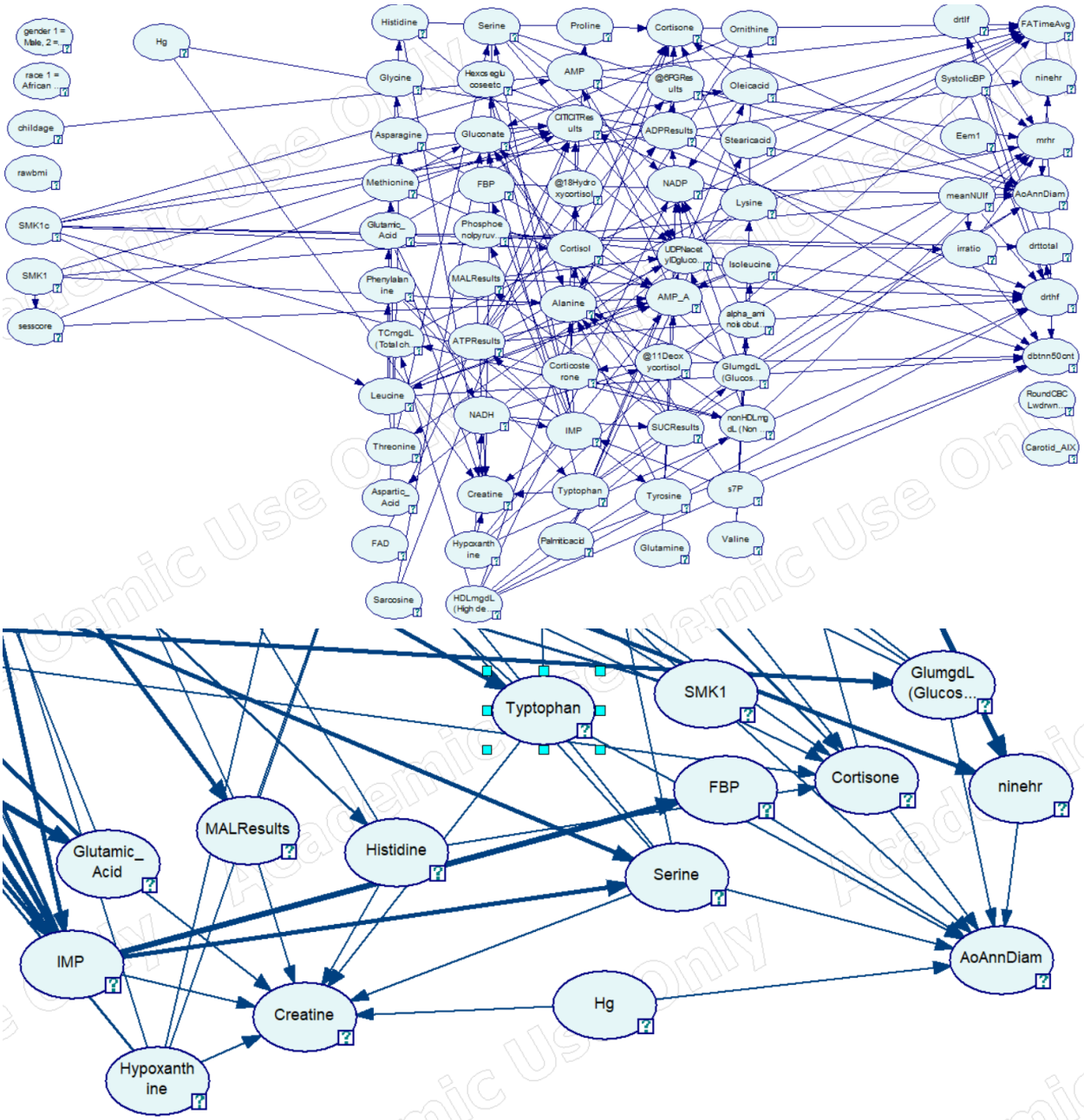
Lead Network:





Here Lead is connected with four different metabolites, being TCmgdl, Palmiticacid, ATPResults and Cortisone.

Mercury Network:



We see that Mercury is linked to the metabolite Creatine and the health measure AoAnnDiam.

Discussion:

Cadmium Network: The Cadmium network showed several interesting connections. The relationship between Cd predicting both mrhr and Methionine and Methionine predicting mrhr is a relationship that merits more investigation. Although these connections don't have high strength, they still could be indicative of the sort of limiting relationships between metabolites, metals and health measures we were hoping to find. Of further interest would be Cd's connections to either dfninetpr and mrhr and their connections to any metabolites as possible places to look. This network shows us a few interesting connections which we could make assumptions or theories about, it narrows down places to look deeper with further attempts at networks with fewer variables and different prior knowledge setups to try. This is true in general for all three metal networks. Other interesting relationships to focus on would be Cd -> mrhr -> ninehr and irestLVET; Cd -> dfninetpr -> dfninemap. Another interesting relationship is rawbmi, SUCResults and Cd. this combination of a confounding factor, health measure and metal could be a place to find a limiting situation.

Lead Network: In this network Lead is connected directly with four Metabolites so we look to those metabolites' connections with health measures for the relationships we are most interested in. We see that ATP Results is predictive of N_Disorder, CSdef and RVDiasDimen. Also @11Deoxycortisol predicts cortisone as well as N_Disorder. These relationships that are closely related to Lead through one or two connections which could be more of the limiting and interconnection we are looking to find. All the relationships depicted could be interesting although most of them are inter metabolite influence, which is interesting, but we are more interested in influences of health measures. Of interest as well is rawbmi which predicts TCmgdltotal as well as several other attributes in this grouping around Lead.

Mercury Network: The Mercury network has one particularly interesting relationship group being Hg -> AoAnnDiam and several Metabolites also predicting AoAnnDiam. Being a cardiovascular health measure, this is particularly in sync with predictions we had about the possible connections we would find. SMK1 which was child smoking as a confounding factor also points to AoAnnDiam which could be another possible interplay. Many of these connections do not have very thick connections denoting high strength, but they are still worth investigating in greater detail. The connection to Creatine with Hg is also interesting as several other metabolites are also predictive of Creatine.

Conclusion:

Only a few of the connections seen in the networks created were mentioned but most of the connections could be worth looking into. We focused more on interconnections between the metals, metabolites and health measures and were more dismissive of only metabolite webs but using this knowledge as basis of further inquiry could still be very valuable if for instance, we looked into what health measures are known to be influenced by certain metabolites linked with a heavy metal that were not included in this study. These networks contain many connections that constitute possible partial information in future setups, networks or inquires.

Works Cited

Birdsall, R., Kiley, M., Segu, Z., Palmer, C., Madera, M., Gump, B., MacKenzie, J., Parsons, P., Mechref, Y., Novotny, M., Bendinskas, K. (2010). Effects of lead and Mercury on the Blood Proteome of Children. *Journal of Proteome Research*, 4443-4453.

Lead Levels Linked to Blood Pressure. (2003, March 25). Retrieved from https://www.webmd.com/hypertension-high-blood-pressure/news/20030325/lead-levels-linked-to-blood-pressure#

Chirinos, J., Akers, S., Trieu, L., Ischiropoulos, H. (2016). Heart Failure, Left Ventricular Remodeling, and Circulating Nitric Oxide Metabolites. Journal of the American Heart Association, 5(10). doi:https://doi/10.1161/JAHA.116.004133

Brownlee, J. (2019, October 11). A gentle introduction to bayesian belief networks. Retrieved February 07, 2021, from https://machinelearningmastery.com/introduction-to-bayesian-belief-networks/

Kraisangka, Jidapa & Druzdzel, Marek. (2014). Discrete Bayesian Network Interpretation of the Cox's Proportional Hazard Model. LNCS. 8754. 10.1007/978-3-319-11433-0_16.

Appendix

**Variables pseudo data dictionary for separated datasets by type (metals, cofoundingfactors, health measures, metabolites targeted, metabolites untargeted)*

Variable List

Metals

Patient ID	(Subnum)	1
Metal	(Cd-Hg)	2-4

CofoundingFactors

Patient ID	(Subnum)	1
Gender, Race, Age, SES Score, Height, Weight		2-7
BMI	(bmipct-rawbmi)	8-9
SmokingSurvey	(SMK1c-SMK9)	10-25

HealthMeasures

Patient ID	(Subnum)	1
Echo Results	(PCAnum – BSA)	2-31
GumpCVRVariables	(Order- Meandfsv)	32-125
Hefferman Data_Final	(Carotid_IMT-Aoritic_AiX_atHR75)	126-131
HRVScoredData	(IRSDNN-Dninetotal)	132-246
PsychosocialVars	(daysbetween-N_Discord)	247-381

MetabolitesTargetted

Patient ID	(Subnum)	1
Alere Data	(TCmgdl – GlumgdL)	2-5
SLS_amino_acids_quantitative (Amino Acids)	(Alanine – Typtophan)	6-26
SLS_GLY_TCA_additional_SEMIquantitative (Glucids)	(Stearicacid - s7P)	27-43
SLS_GLY_TCA_quantitative (Lipids)	(SUCResults-ATPResults)	44-49
SLS_Steroids_quantitative_modified	(Cortisol-@18Hydroxycortisol)	50-54

MetabolitesUntargetted

Patient ID	(Subnum)	1
UntargettedMetabolites	(CE171MNH4@10.93 - @977.2973@22.69408)	2-2530

**For original pseudo data dictionary see OriginalDataDiction.docx in Documents folder in the main project folder or analysisResultsFinalV5(1).docx*

**For all datasets see Datasets folder in the main project folder*

**For all Genie networks see main project folder for background knowledge files for networks see Datasets folder in the main project folder*