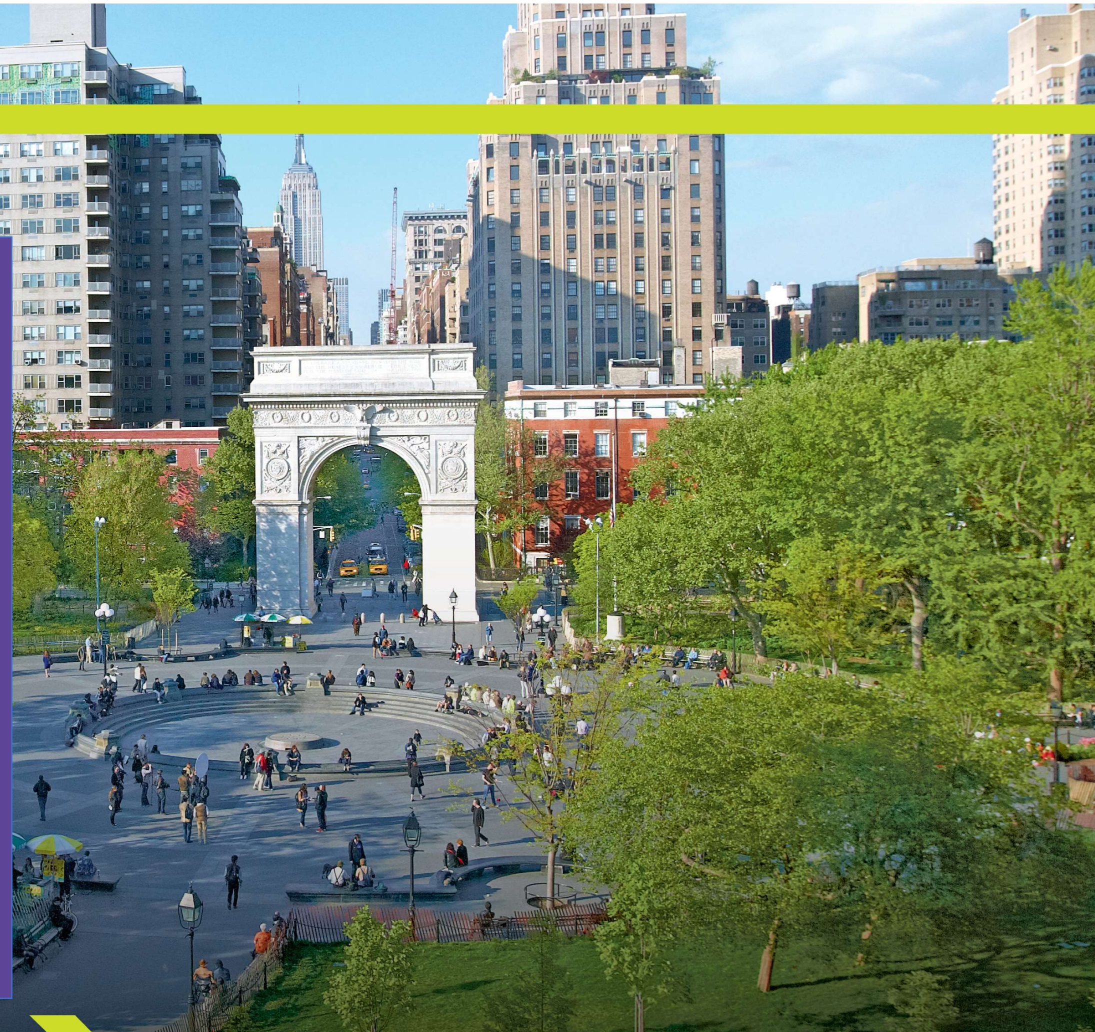
Generating models before generating parameters: A Bayesian network approach

Ray Niaura, PhD

CAsToR Symposium 2021
Simulation Modeling in Tobacco Regulatory Science: Where are we and where should we go next?
June 7 — 9, 2021



reinvent THE PUBLIC HEALTH PARADIGM

DISCLOSURES

Past 3 years: I do work with the FDA CTP via contractual mechanisms with Westat. I am a co-Investigator on several NIH grant awards. During 2019, I reviewed grant proposals related to tobacco harm reduction studies funded by the Foundation for a Smoke Free World. I did not receive any compensation for this work. I quit this activity in early 2020.

This talk has nothing to do with any of the above, my opinions are my own, and I only represent myself and not any other entities, human or otherwise.

OUTLINE

- Running example: Dealing with high dimension data in the PATH Study youth sample – ecig->cig pathway
- Propensity score balancing
- Intro to Bayesian Networks (BN), information entropy
- BN structure learning
- Confounding or what?

Relationships Between E-cigarette Use and Subsequent Cigarette Initiation Among Adolescents in the PATH Study: An Entropy Balancing Propensity Score Analysis

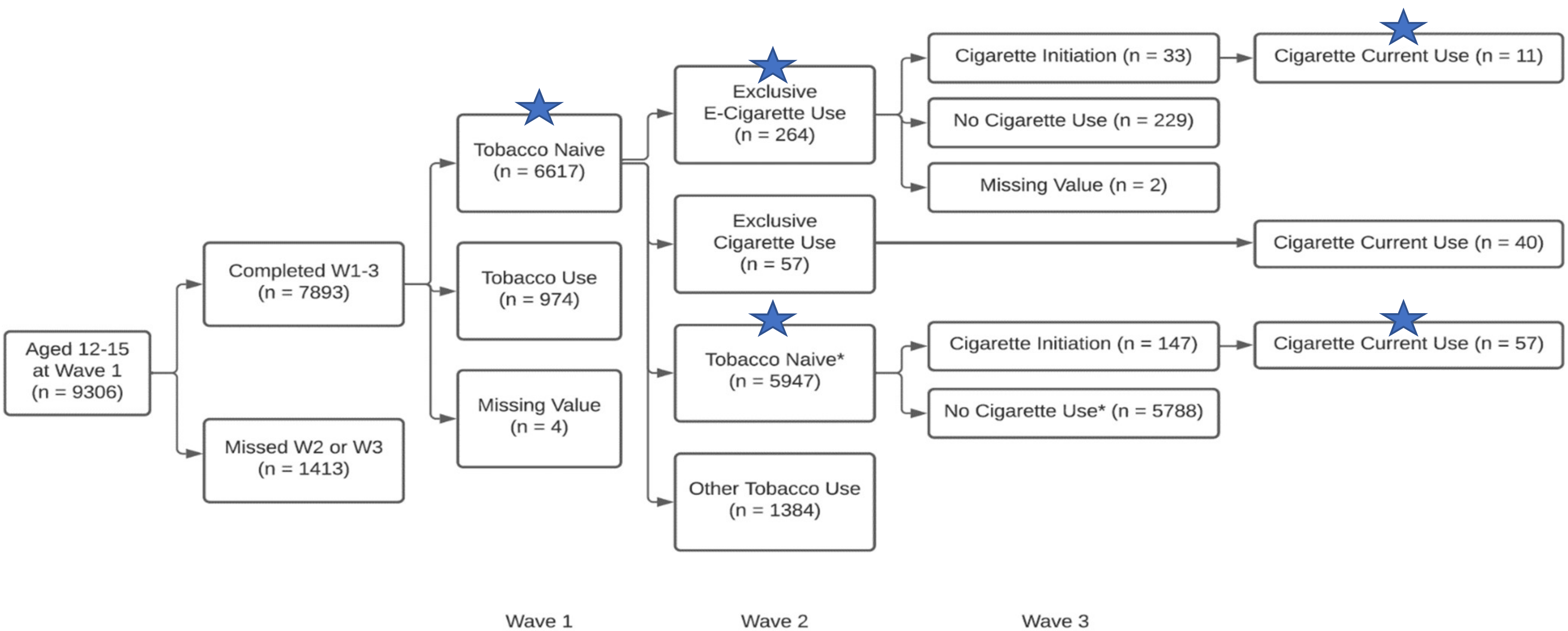
Authors and affiliations:

Shu Xu, PhD[1], Donna L. Coffman, PhD[2], Bin Liu, MPH[1], Yifan Xu, MPH[1], Jiarui He, MPH[1], Raymond Niaura, PhD[1]

[1] School of Global Public Health, New York University

[2] College of Public Health, Temple University

PATH Study, Youth Sample, Waves 1-3: New User Design



Public use data.

MODEL VARIABLES

Exposure**.** Ever used e-cigarettes (Sample 1) or combustible cigarettes (Sample 2) exclusively at Wave 2 (2014 – 2015).

Outcome**.** Combustible cigarette past 30-day (P30D) use assessed at Wave 3 (2015 – 2016).

Pre-exposure confounders**.** 55 pre-exposure variables (Wave 1, 2013 - 2014): self- or parent-report on socio-demographic (e.g., sex, race), interpersonal (e.g., depression, impulsivity, medical history), behavioral (e.g., alcohol use, drug use) and social environmental (e.g., living with smokers, parental monitoring) factors.

Entropy balancing (EB): A multivariate reweighting method which adjusts the weight of each participant so that the covariate distributions in the reweighted data achieve balance (i.e., mean and variance). Obviates the need for PS matching. Survey weights were included in the model for estimating EB weights.

Entropy balancing of covariates

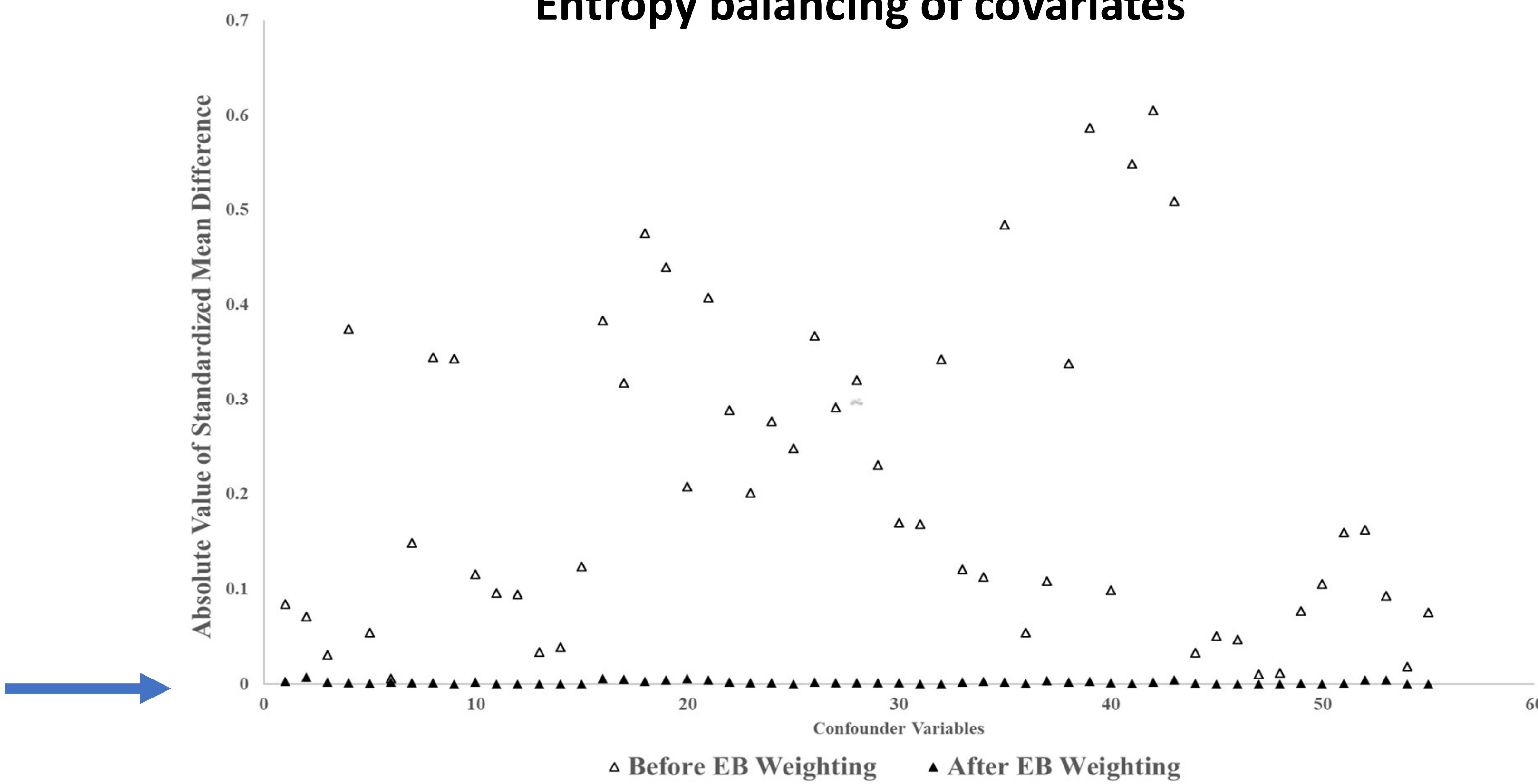


Figure 1. Standardized mean differences (SMD) of pre-exposure confounders before and after entropy balancing. All SMDs of the 55 confounder variables after EB were close to 0.

Effect of Initial *E-cigarette Exposure* on P30-day Cigarette Use

Table 2

The Effects of E-cigarette Ever Use on Subsequent Combustible Cigarette Use

Outcome	Sampling Weights Only			EB + Sampling Weights		
Sample 1	B	SE	OR (95% CI)	B	SE	OR (95% CI)
Cigarette Initiation	1.79	0.25	5.99 (3.66, 9.78)	1.17	0.34	3.22 (1.65, 6.33)
Past 30-day Cigarette Use	1.64	0.34	5.16 (2.64, 10.03)	1.30	0.51	3.67 (1.35, 10.06)
Sample 2						
Past 30-day Cigarette Use	3.84	0.30	47.23 (26.14 85.34)	3.09	0.41	21.98 (9.86, 49.43)

Notes. EB = Entropy balancing; B: unstandardized regression coefficient; SE = standard error; OR = adjusted odds ratio. The first set of ORs was based on a model adjusted for sampling weights only. The second set of ORs was based on a model adjusted for sampling weights and entropy balancing weights, where the entropy balancing model also used sampling weights.

Effect of Initial *Cigarette* Exposure on P30-day Cigarette Use

Table 2

The Effects of E-cigarette Ever Use on Subsequent Combustible Cigarette Use

Outcome	Sampling Weights Only			EB + Sampling Weights		
Sample 1	B	SE	OR (95% CI)	B	SE	OR (95% CI)
Cigarette Initiation	1.79	0.25	5.99 (3.66, 9.78)	1.17	0.34	3.22 (1.65, 6.33)
Past 30-day Cigarette Use	1.64	0.34	5.16 (2.64, 10.03)	1.30	0.51	3.67 (1.35, 10.06)
Sample 2						
Past 30-day Cigarette Use	3.84	0.30	47.23 (26.14 85.34)	3.09	0.41	21.98 (9.86, 49.43)

Notes. EB = Entropy balancing; B: unstandardized regression coefficient; SE = standard error; OR = adjusted odds ratio. The first set of ORs was based on a model adjusted for sampling weights only. The second set of ORs was based on a model adjusted for sampling weights and entropy balancing weights, where the entropy balancing model also used sampling weights.

Bayesian Network

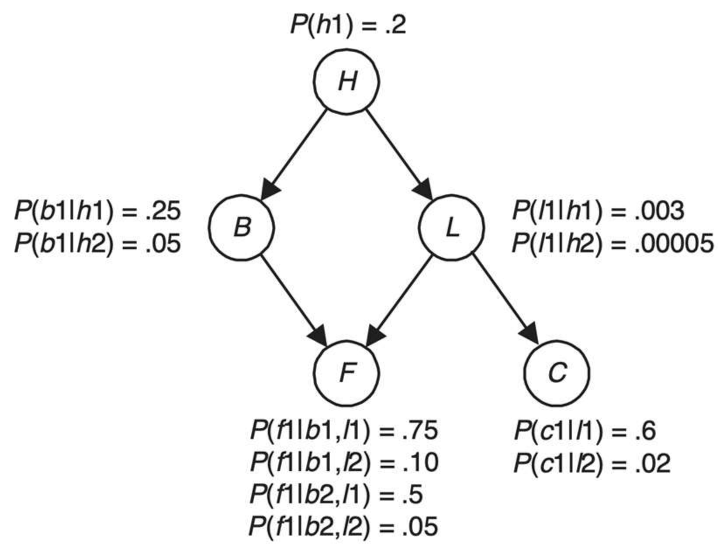


Figure 4.13: A Bayesian nework.

Variable	Value	When the Variable Takes this Value
H	$h1$	Patient has a smoking history
	$h2$	Patient does not have a smoking history
B	$b1$	Patient has bronchitis
	$b2$	Patient does not have bronchitis
L	$l1$	Patient has lung cancer
	$l2$	Patient does not have lung cancer
F	$f1$	Patient is fatigued
	$f2$	Patient is not fatigued
C	$c1$	Patient has a positive chest X-ray
	$c2$	Patient has a negative chest X-ray

$$P(x_1,x_2,x_3,x_4,x_5) = P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_1)P(x_4 \mid x_2,x_3)P(x_5 \mid x_4)$$

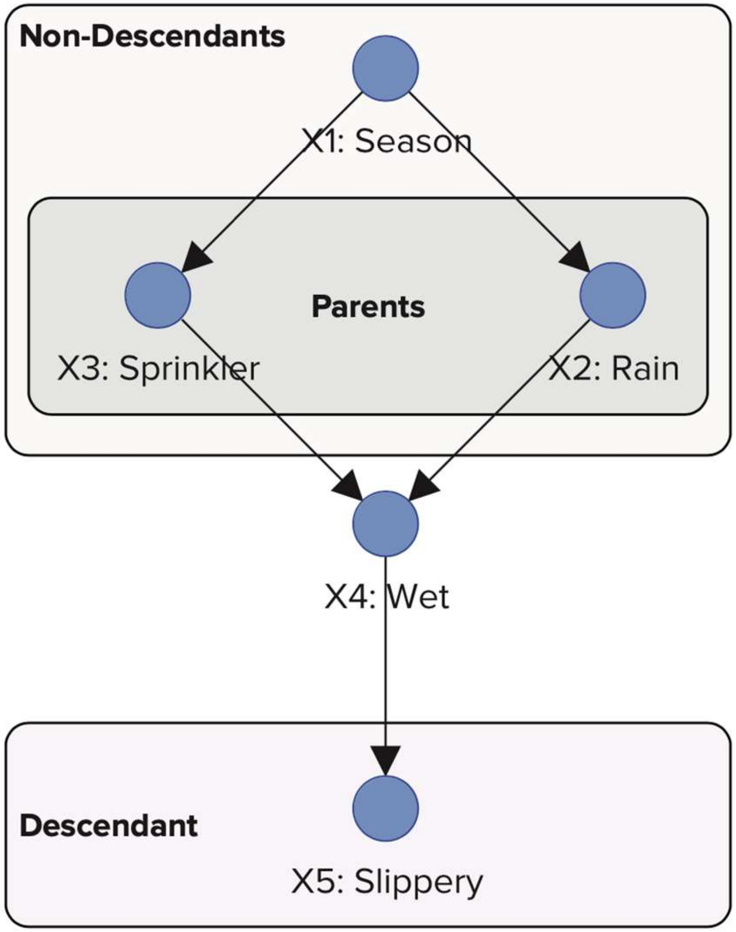


Figure 2.4

For example, the probability that the *Sprinkler* is on given that the *Pavement* is slippery is:

$$\begin{aligned} P(X_3 = on \mid X_5 = true) &= \frac{P(X_3 = on, X_5 = true)}{P(X_5 = true)} \\ &= \frac{\sum_{x_1,x_2,x_4} P(x_1,x_2,X_3 = on,x_4,X_5 = true)}{\sum_{x_1,x_2,x_3,x_4} P(x_1,x_2,x_3,x_4,X_5 = true)} \\ &= \frac{\sum_{x_1,x_2,x_4} P(x_1)\,(x_2 \mid x_1)\,P(X_3 = on \mid x_1)\,P(x_4 \mid x_2, X_3 = on)\,P(X_5 = true \mid x_4)}{\sum_{x_1,x_2,x_3,x_4} P(x_1)\,P(x_2 \mid x_1)\,P(x_3 \mid x_1)\,P(x_4 \mid x_2,x_3)\,P(X_5 = true \mid x_4)} \end{aligned} \tag{2.5}$$

BN software

BAYESIA

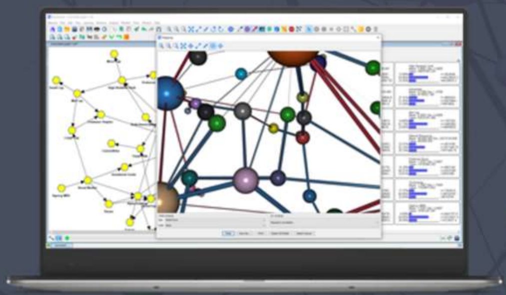
The Bayesia Product Portfolio | BayesiaLab Knowledge Hub & Library | Courses & Events | BayesiaLab Store

BayesiaLab 9

The Leading Desktop Software for Bayesian Networks.
Artificial Intelligence for Research, Analytics, and Reasoning

Built on the foundation of the Bayesian network formalism, BayesiaLab is a powerful desktop application (Windows, macOS, Linux/Unix) with a highly sophisticated graphical user interface. It provides scientists a comprehensive "lab" environment for machine learning, knowledge modeling, diagnosis, analysis, simulation, and optimization. With BayesiaLab, it has become feasible for applied researchers in many fields, rather than just computer scientists, to take advantage of the Bayesian network formalism.

Learn More



Information Entropy

Definition

Entropy, denoted $H(X)$, is a key metric in BayesiaLab for measuring the uncertainty associated with the probability distribution of a variable X.

Entropy is expressed in bits and defined as follows:

$$H(X) = -\sum_{x \in X} p(x) \log_2 (p(x))$$

The **Entropy** of a variable X can also be understood as the sum of the **Expected Log-Losses** of its states.

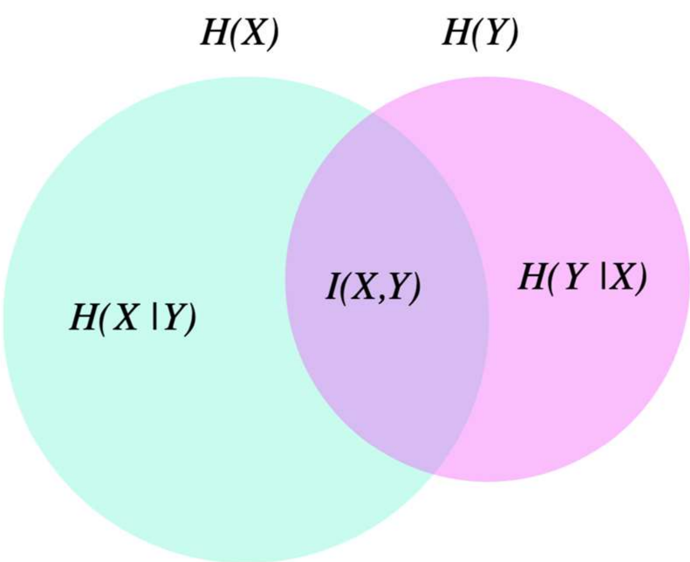
Mutual Information

Definition

The **Mutual Information** $I(X, Y)$ measures the amount of information gained on variable X (the reduction in the **Expected Log-Loss**) by observing variable Y:

$$I(X,Y) = H(X) - H(X|Y)$$

The **Venn Diagram** below illustrates this concept:

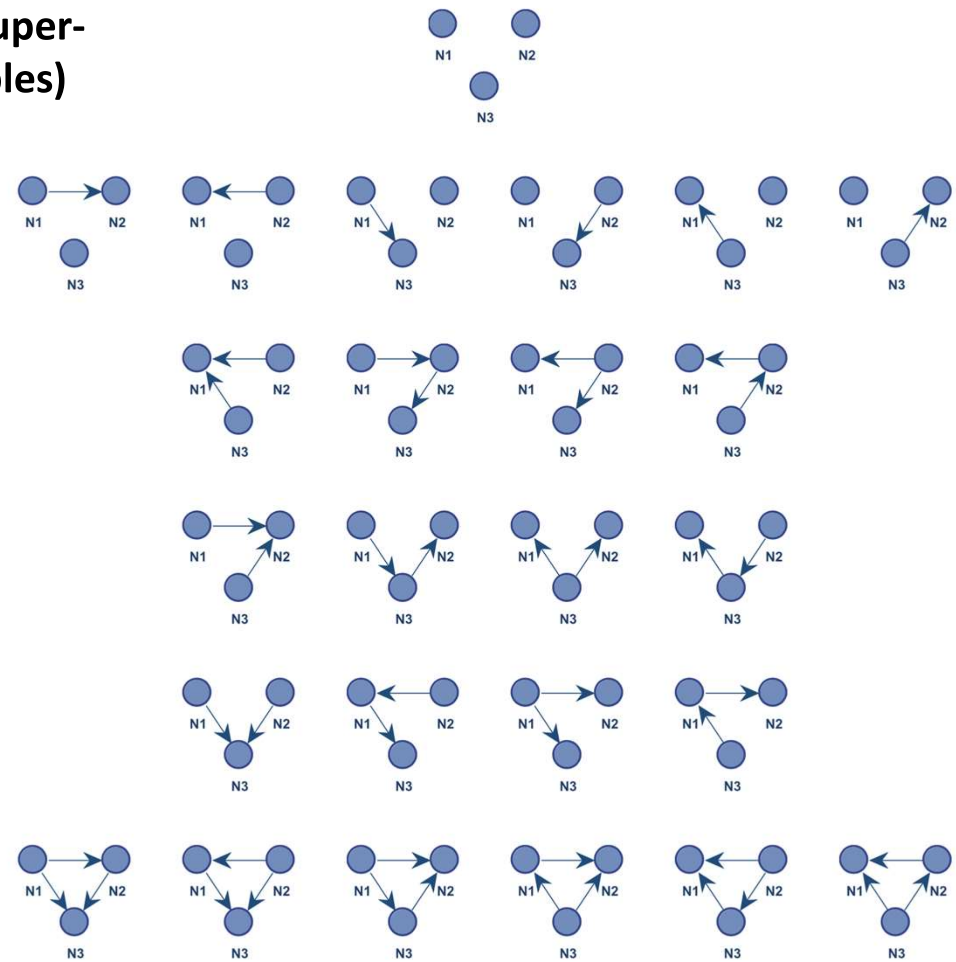


The **Conditional Entropy** $H(X|Y)$ measures, in bits, the **Expected Log-Loss** associated with variable X once we have information on variable Y:

$$H(X|Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log_2 (p(x|y))$$

Number of possible networks (models) grows super-exponentially with the number of nodes (variables)

Number of Nodes	Number of Possible Networks
1	1
2	3
3	25
4	543
5	29281
6	3.7815×10^{6}
7	1.13878×10^{9}
8	7.83702×10^{11}
9	1.21344×10^{15}
10	4.1751×10^{18}
...	...
47	8.98454×10^{376}



Learning Bayesian Network Structure

Score-based algorithms, based on a metric (MDL) that measures the quality of candidate networks with respect to the observed data. Trades off network complexity against the degree of fit to the data, which is typically expressed as the likelihood of the data given the network.

Easy to encode prior knowledge in network form, either by fixing portions of the structure, forbidding relations, or by using prior distributions over the network parameters.

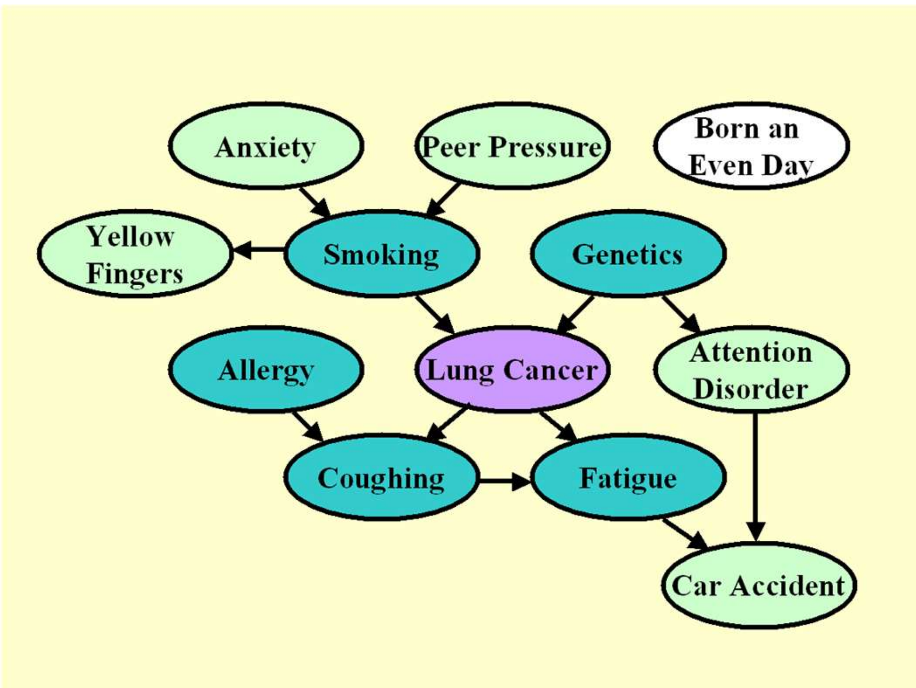
Minimum Description Length (MDL) Score

$$MDL(B,D) = \alpha DL(B) + DL(D \mid B), \tag{8.1}$$

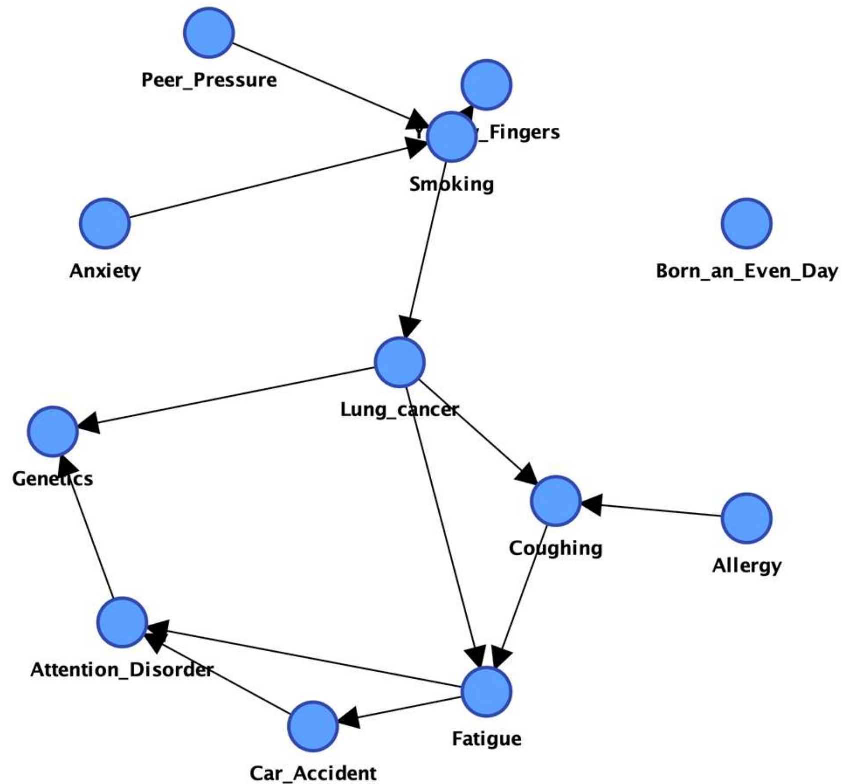
where:

- α represents BayesiaLab's **Structural Coefficient** (the default value is 1), a parameter that permits changing the weight of the structural part of the MDL Score (the lower the value of α, the greater the complexity of the resulting networks),
- $DL(B)$ the number of bits to represent the Bayesian network B (graph and probabilities), and
- $DL(D|B)$ the number of bits to represent the dataset D given the Bayesian network B (likelihood of the data given the Bayesian network).

Causal Challenge



Original Model

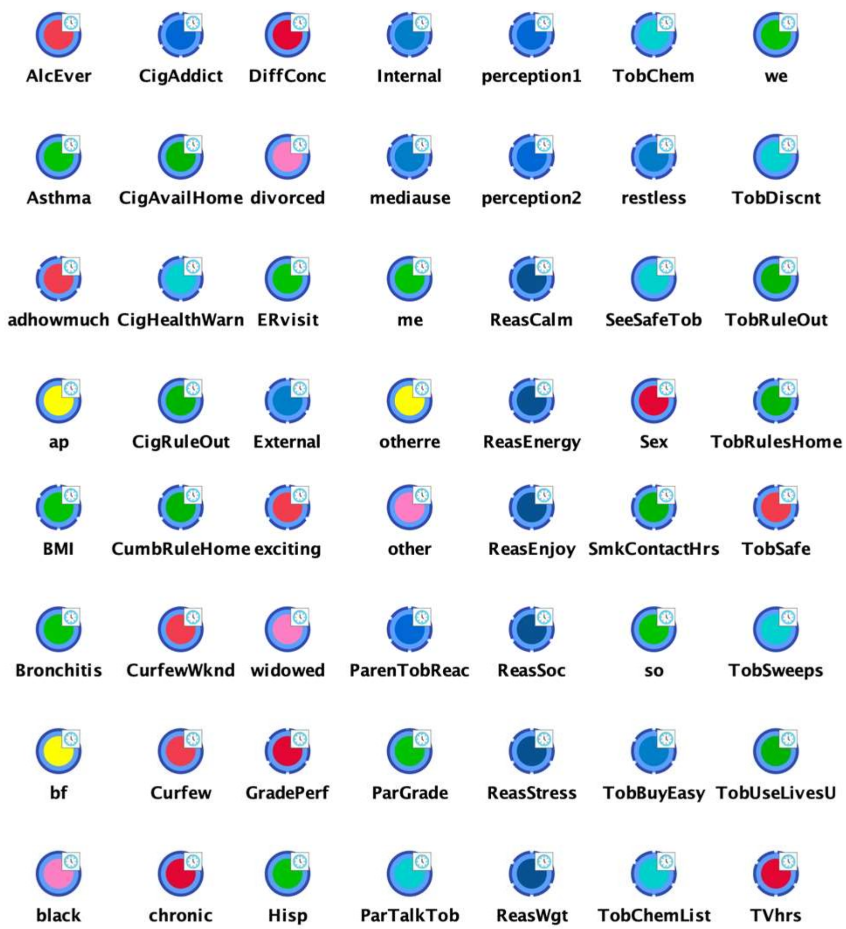


Learned Model


CAUTION
WORK
IN PROGRESS

Pre-Exposure Covariates, Exposure and Outcome (temporal sequence)

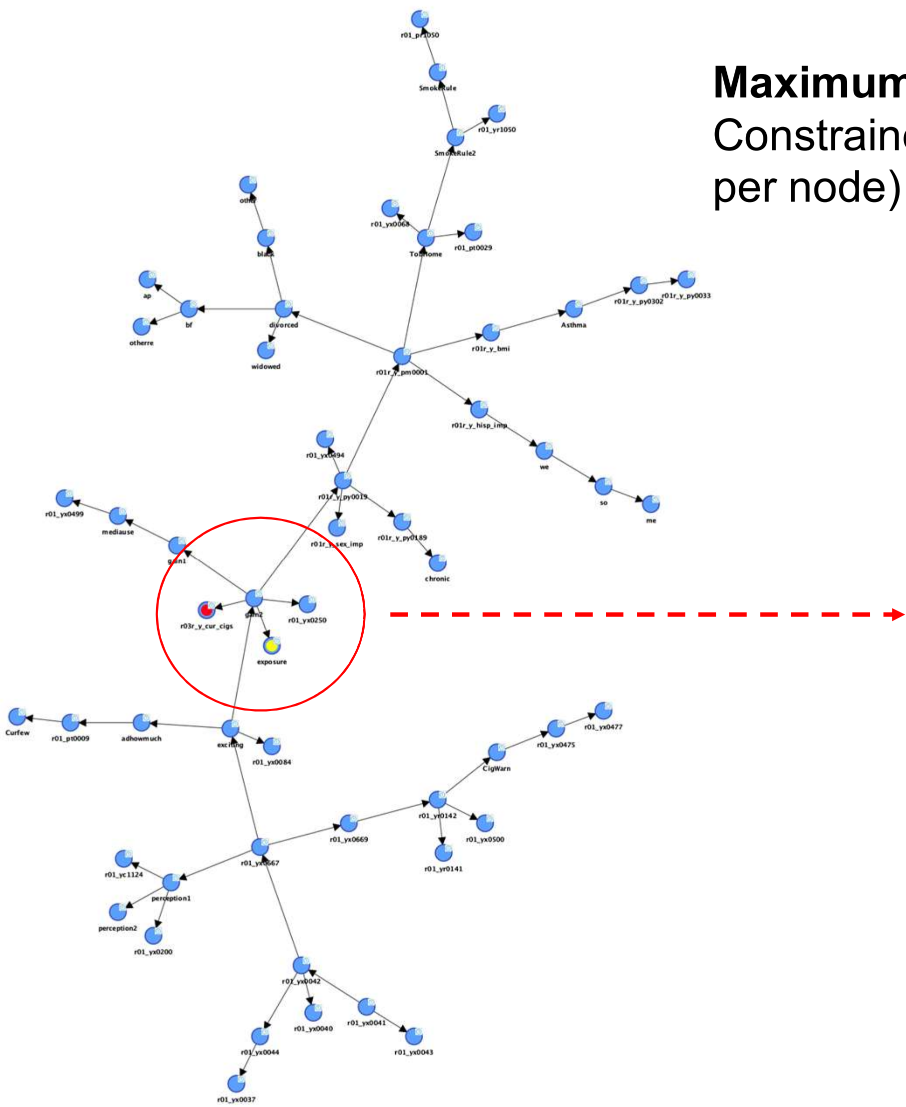
Wave 1

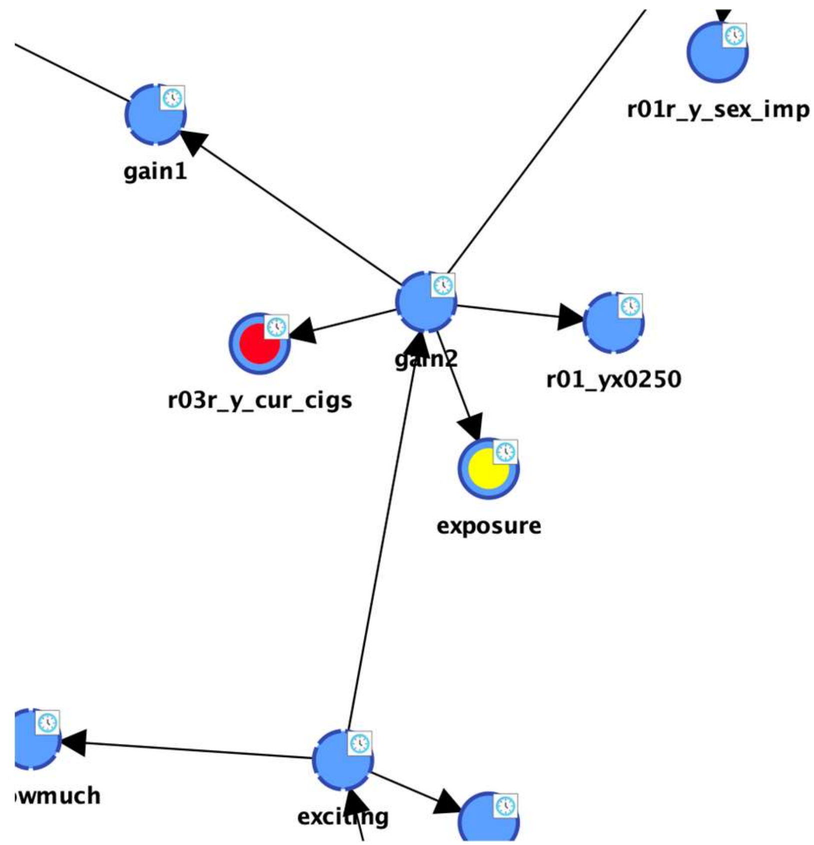


Wave 2 Wave 3

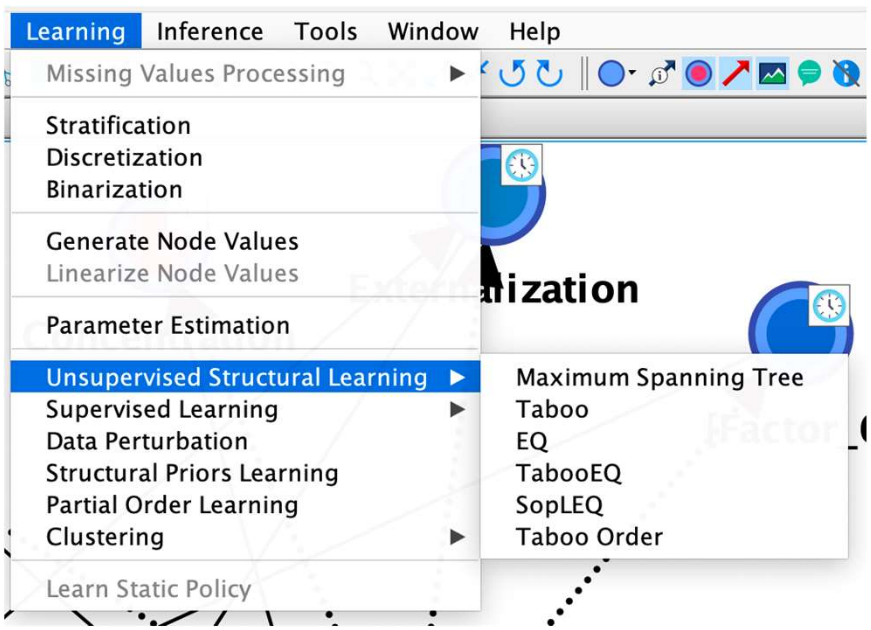
Number of Nodes	
Total	58
Discrete	31
Continuous	27
Constraint	0
Decision	0
Utility	0
Function	0

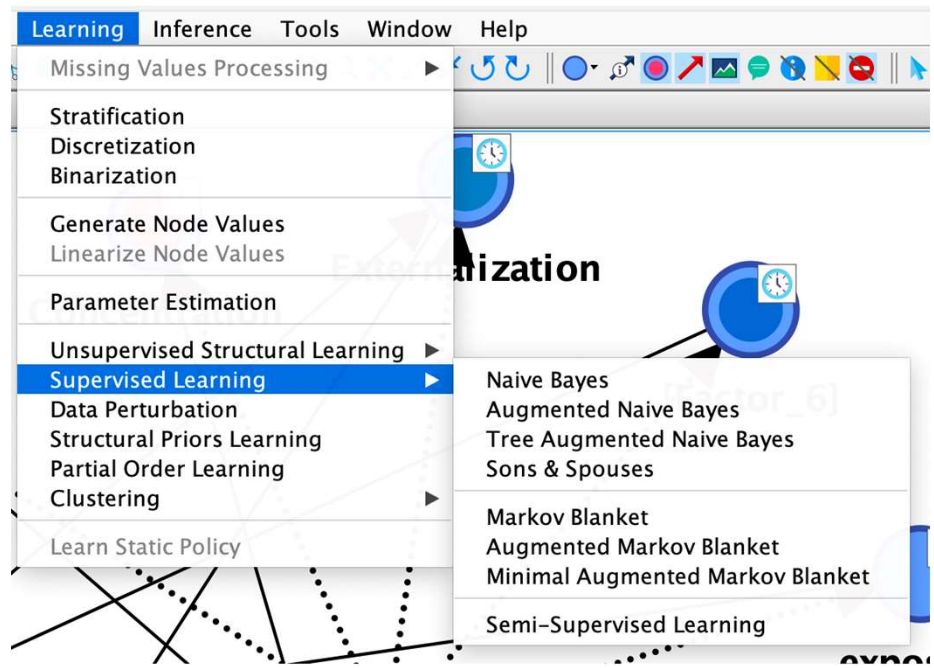
Maximum Weight Spanning Tree (MWST).
Constrained to learning a tree structure (one parent per node)



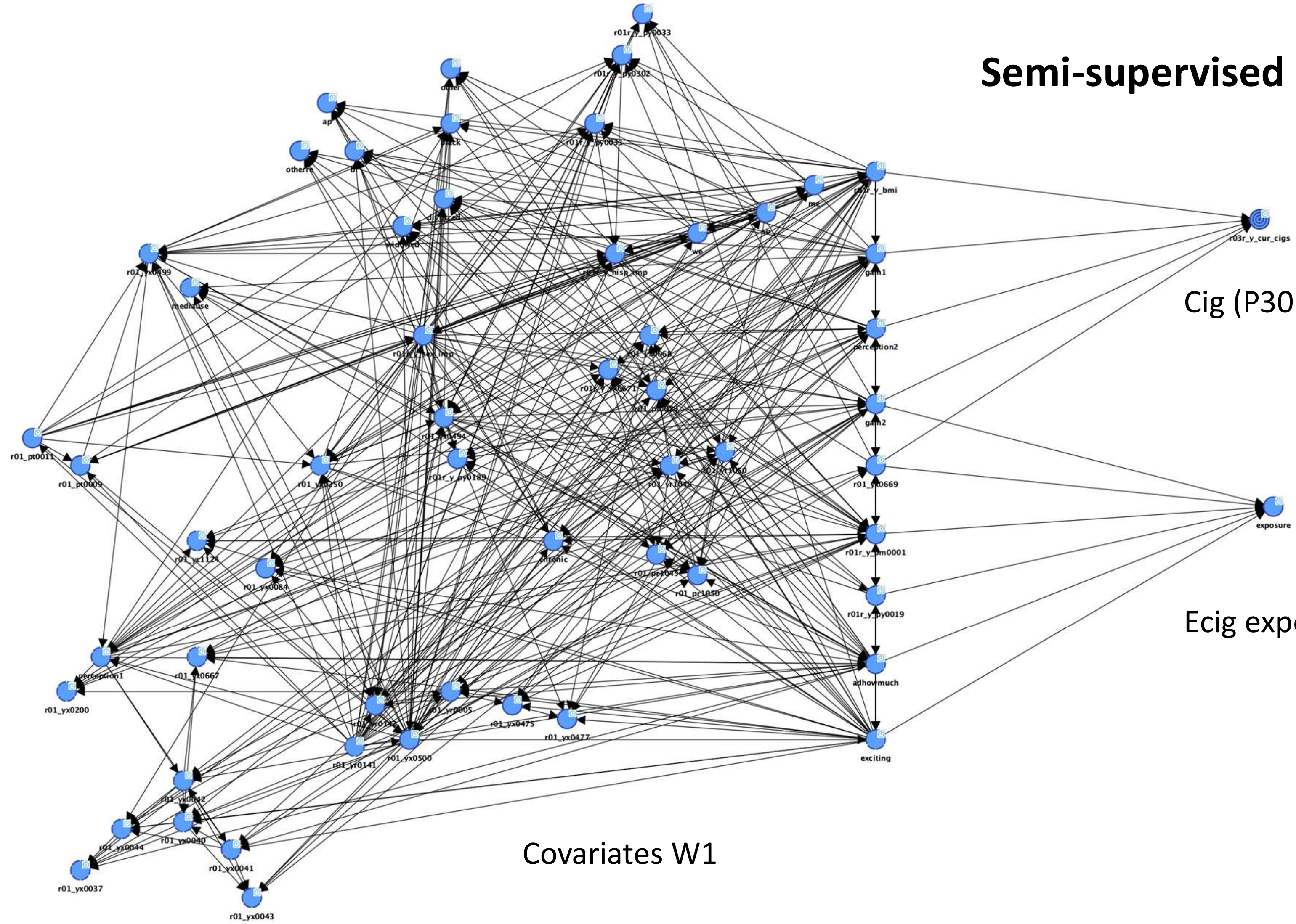


BN Learning algorithms





Semi-supervised learning

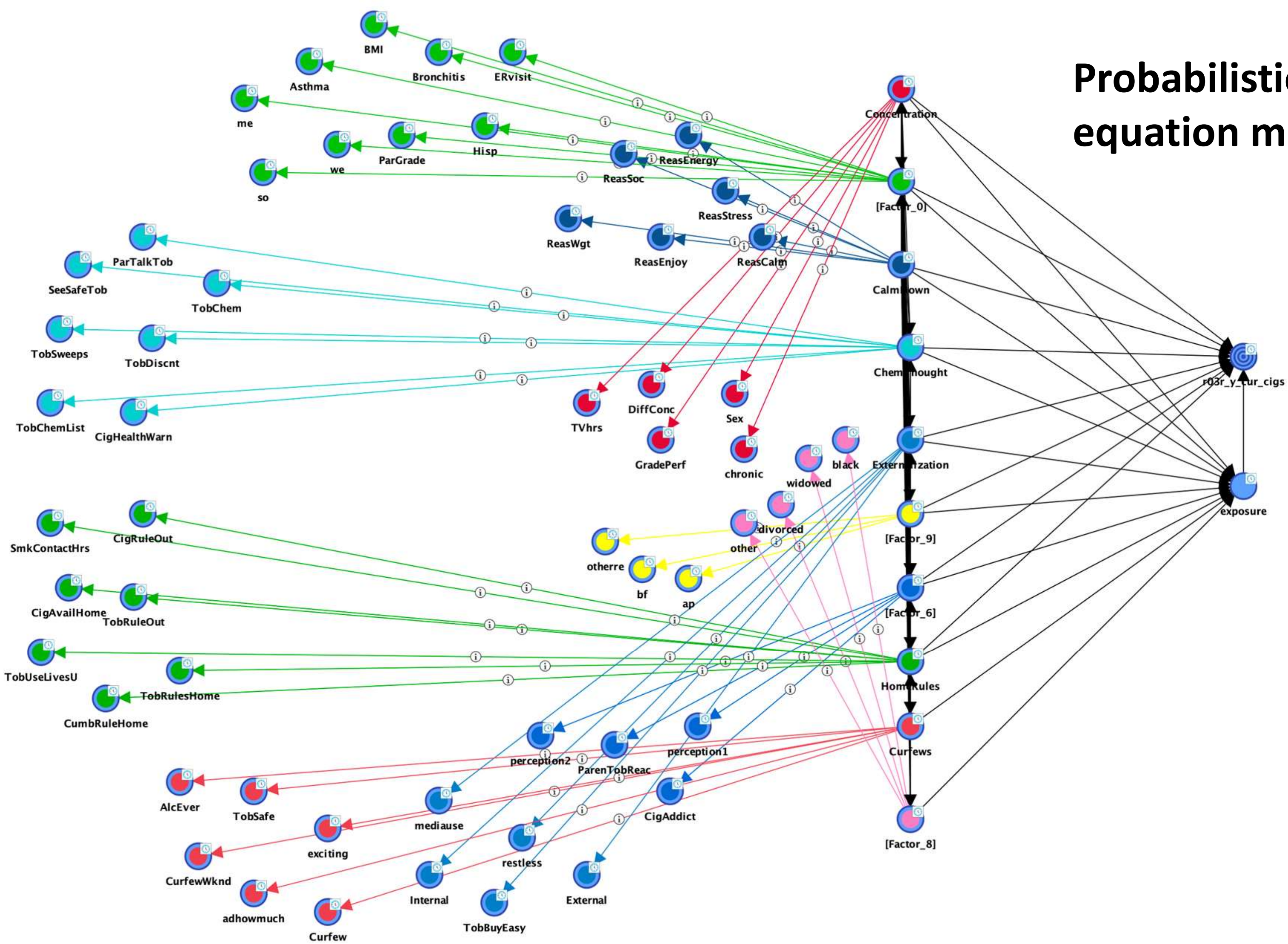


Cig (P30D) W3

Ecig exposure W2

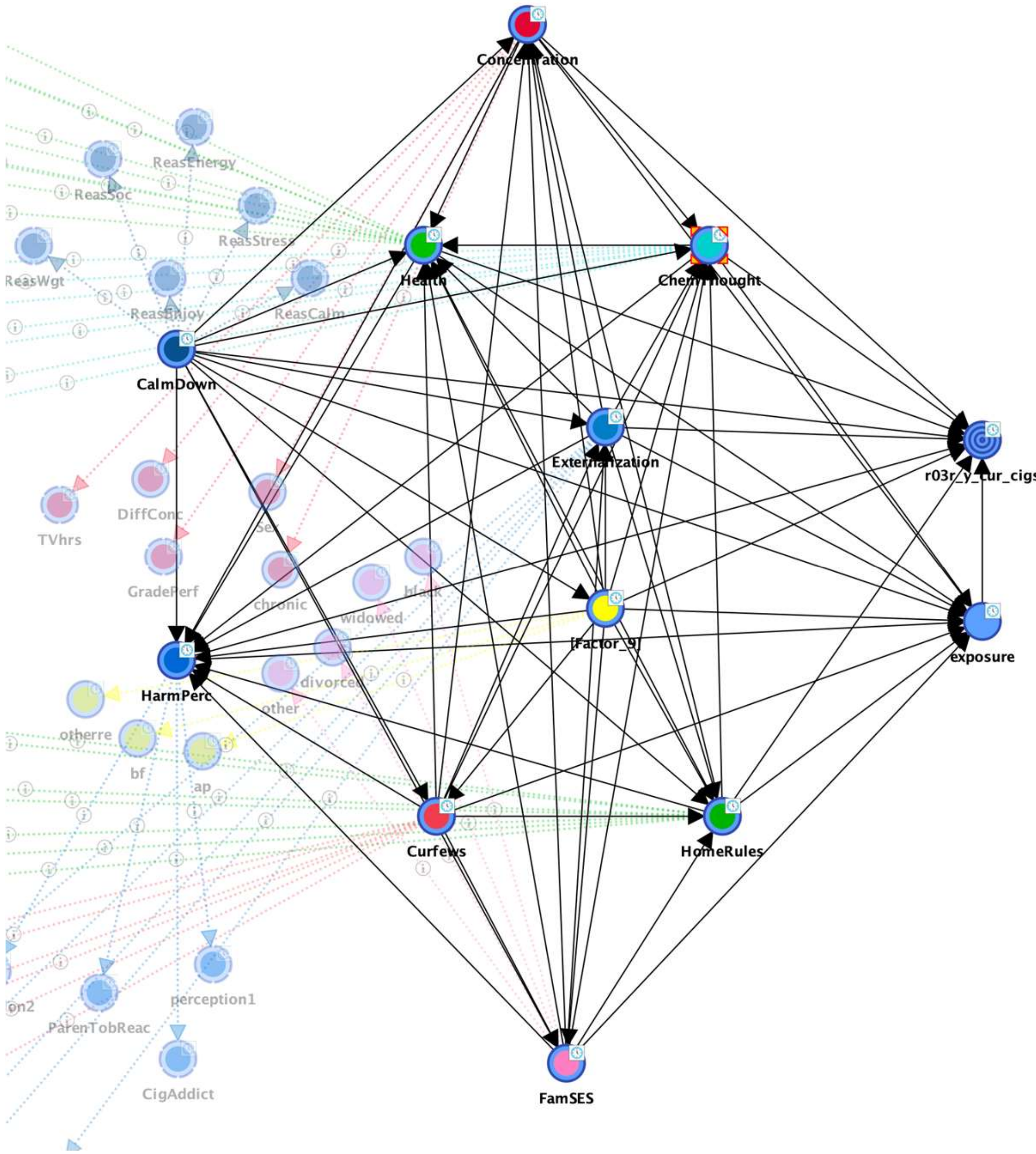
Covariates W1

Probabilistic structural equation model



Cig (P30D) W3

Ecig exposure W2


Concentration
Health
ChemThought
CalmDown
Externalization
r03r_y_cur_cigs
Factor_9
exposure
HarmPerc
Curfews
HomeRules
FamSES
ReasEnergy
ReasSoc
ReasStress
ReasWgt
ReasEnjoy
ReasCalm
TVhrs
DiffConc
GradePerf
chronic
black
widowed
divorced
other
otherre
bf
ap
perception1
ParenTobReac
CigAddict

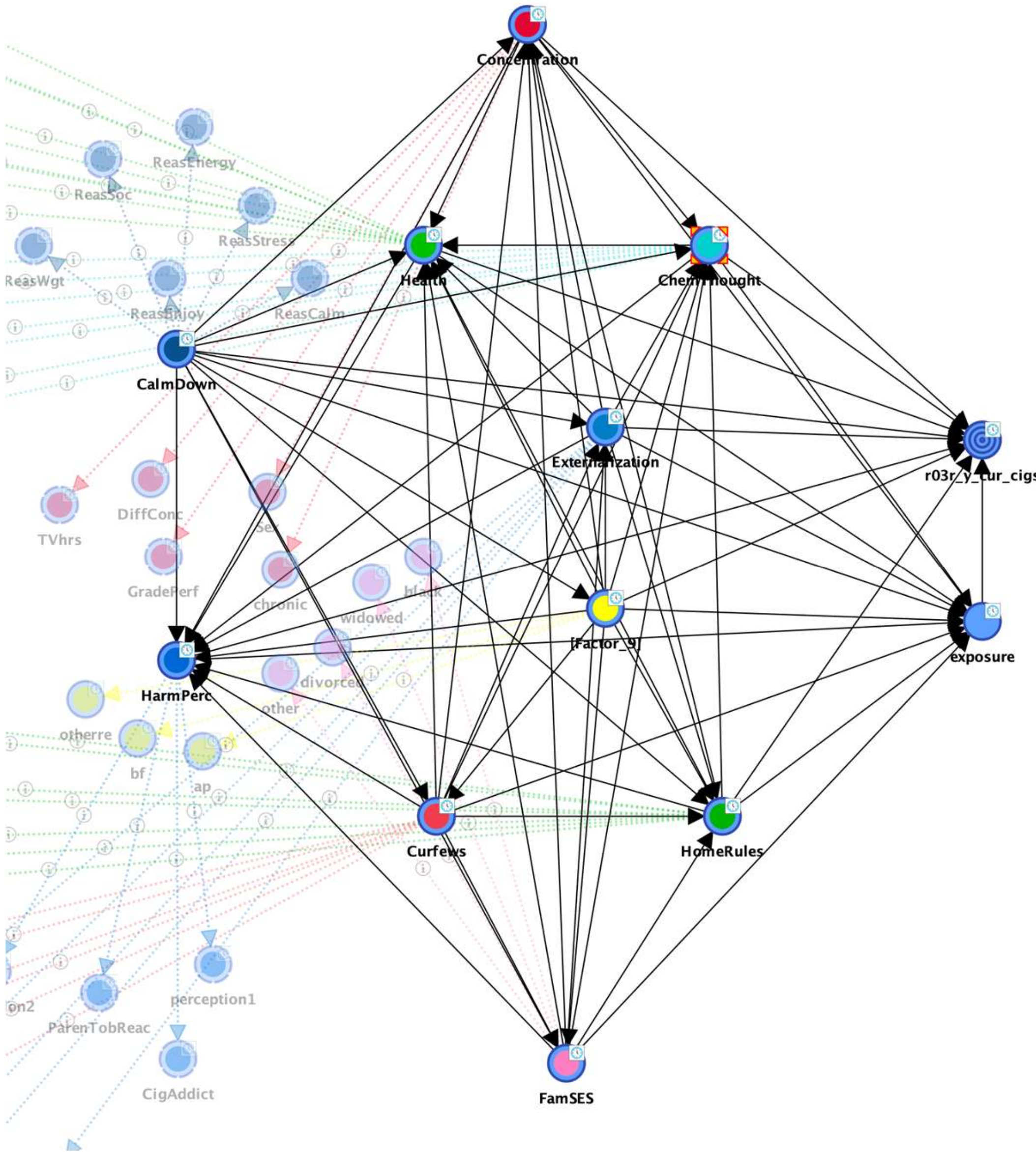
European Journal of Epidemiology (2019) 34:211–219
https://doi.org/10.1007/s10654-019-00494-6

ESSAY

Principles of confounder selection

Tyler J. VanderWeele[1]

- Control for each covariate that is a cause of the exposure, or of the outcome, or of both;
- Exclude from this set any variable known to be an instrumental variable;
- Include as a covariate any proxy for an unmeasured variable that is a common cause of both the exposure and the outcome.



Concentration
Health
ChemThought
CalmDown
Externalization
r03r_y_cur_cigs
Factor_9
exposure
HarmPerc
Curfews
HomeRules
FamSES
ReasEnergy
ReasSoc
ReasStress
ReasWgt
ReasEnjoy
ReasCalm
TVhrs
DiffConc
GradePerf
chronic
widowed
black
divorced
other
otherre
bf
ap
ParenTobReac
perception1
CigAddict

Confusion Matrix

Occurrences | Reliability | Precision

Value	0 (13053434.85)	1 (146010.98)
0 (13150545.09)	99.9348%	72.3359%
1 (48900.73)	0.0652%	27.6641%

Overall Precision: 99.1354%   Mean Precision: 63.7995%

Overall Reliability: 99.0133%   Mean Reliability: 90.8991%

Gini Index: 1.0529%   Relative Gini Index: 95.1818%

Lift Index: 1.0108   Relative Lift Index: 99.9724%

ROC Index: 97.5910%

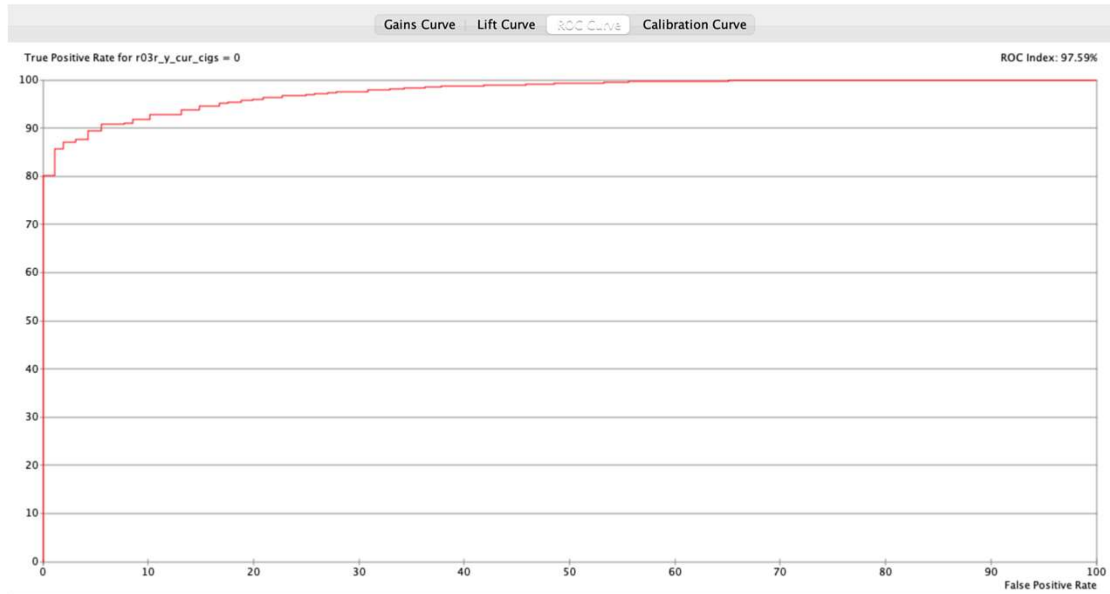
Calibration Index: 89.9159%

Binary Log-Loss: 0.0279

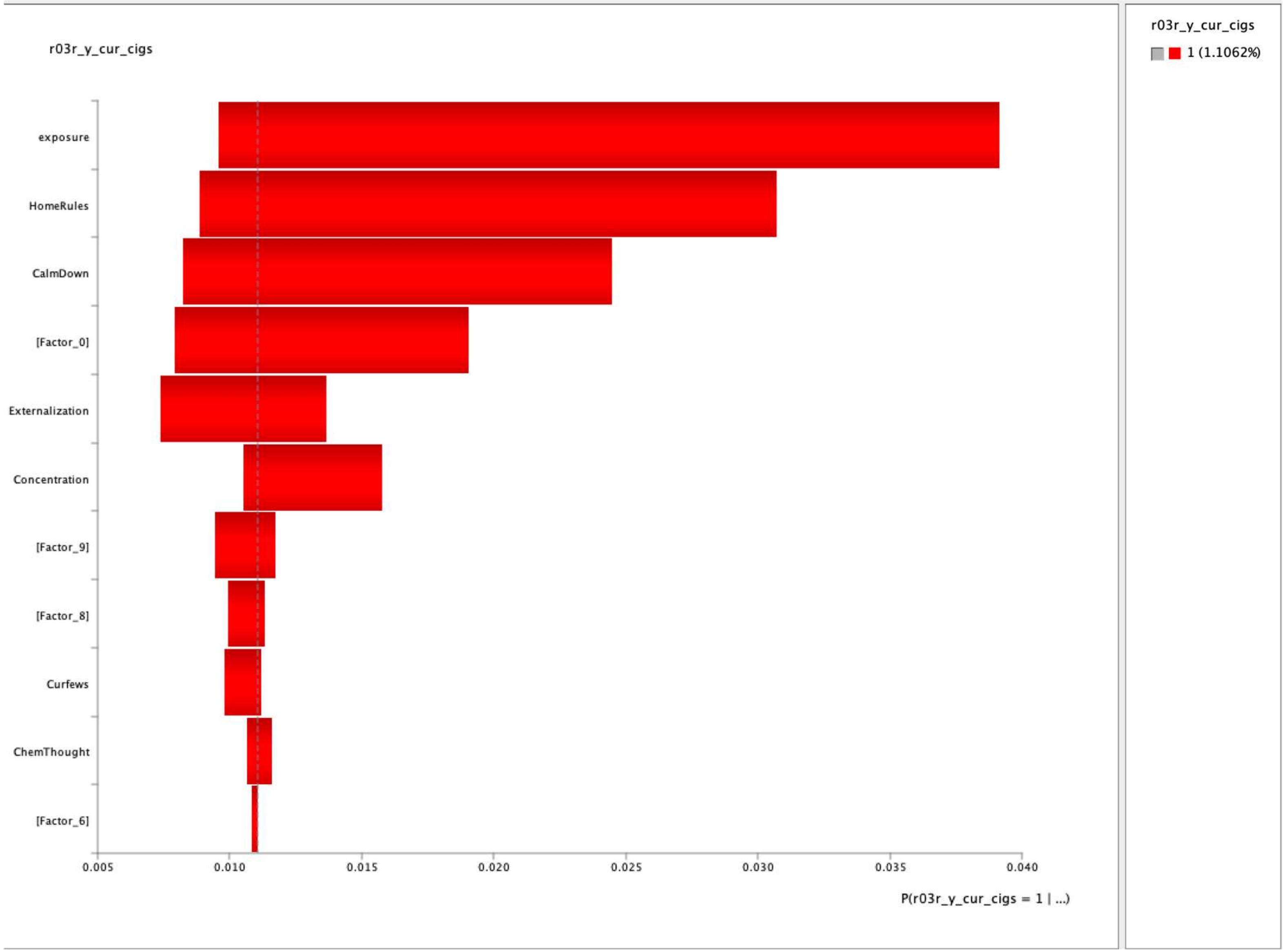
R: 0.5843   RMSE: 0.0849

R2: 0.3414   NRMSE: 8.4883%

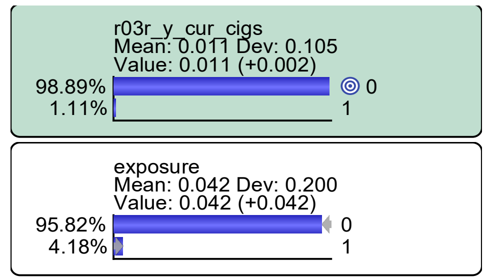
Acceptance Threshold: Maximum Likelihood



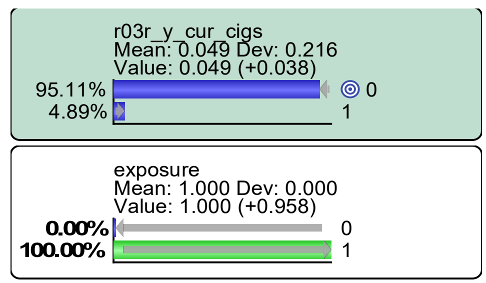
Overall Analysis with r03r_y_cur_cigs											
Node	**Mutual Information**	**Normalized Mutual Information**	**Relative Mutual Information**	**Relative Significance**	**Prior Mean Value**	**G-test**	**df**	**p-value**	**G-test (Data)**	**df (Data)**	**p-value (Data)**
exposure	0.0023	0.2326%	2.6506%	1.0000	0.0418	42,561.1980	1	0.0000%	42,561.1980	1	0.0000%
r01_yx0042_(6)	0.0018	0.1799%	2.0500%	0.7734	3.6475	32,917.2012	3	0.0000%	32,917.2012	3	0.0000%
gain2_(5)	0.0016	0.1632%	1.8593%	0.7015	2.9621	29,854.7689	1	0.0000%	29,854.7689	1	0.0000%
r01_pr1045_(7)	0.0014	0.1412%	1.6090%	0.6071	1.4505	25,837.0445	2	0.0000%	25,837.0445	2	0.0000%
r01r_y_py0189_(5)	0.0005	0.0500%	0.5692%	0.2148	0.9676	9,140.1774	1	0.0000%	9,140.1774	1	0.0000%
we_(9)	0.0003	0.0340%	0.3876%	0.1462	0.7493	6,223.7931	3	0.0000%	6,223.7931	3	0.0000%
bf_(3)	0.0002	0.0225%	0.2563%	0.0967	0.2236	4,114.7563	1	0.0000%	4,114.7563	1	0.0000%
perception1_(4)	0.0002	0.0193%	0.2197%	0.0829	3.2350	3,528.1259	1	0.0000%	3,528.1259	1	0.0000%
r01_yr0142_(7)	0.0002	0.0168%	0.1917%	0.0723	1.6666	3,077.6779	1	0.0000%	3,077.6779	1	0.0000%
other_(4)	0.0000	0.0040%	0.0457%	0.0172	0.2215	733.2377	2	0.0000%	3,270.9457	2	0.0000%
r01_pt0009_(6)	0.0000	0.0009%	0.0101%	0.0038	0.9026	162.8814	1	0.0000%	951.6565	1	0.0000%


r03r_y_cur_cigs
exposure
HomeRules
CalmDown
[Factor_0]
Externalization
Concentration
[Factor_9]
[Factor_8]
Curfews
ChemThought
[Factor_6]
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
P(r03r_y_cur_cigs = 1 | ...)
r03r_y_cur_cigs
1 (1.1062%)

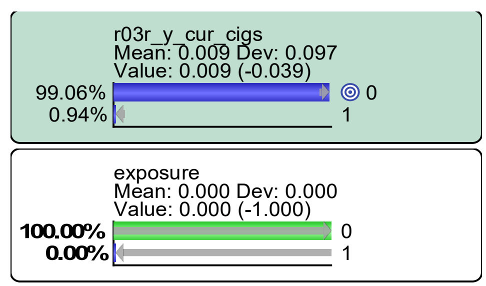
Effects of Ecig Exposure on Cigarette Smoking



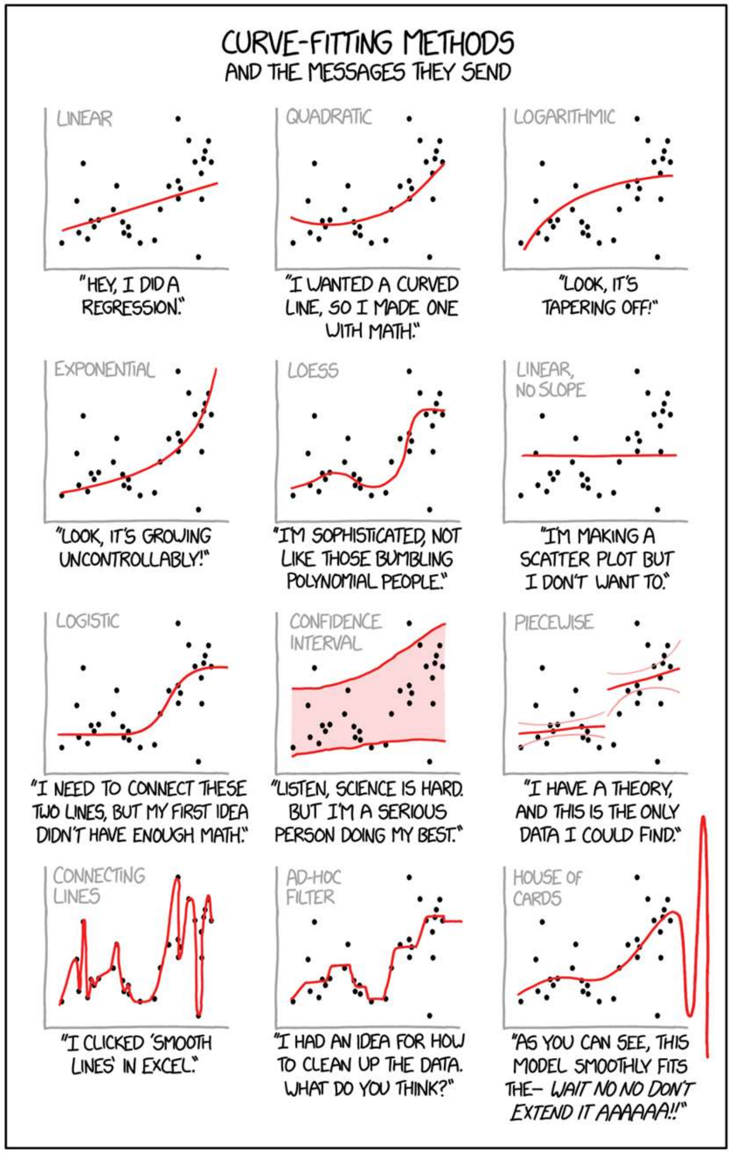
Observation



Ecig use set to 100%



Ecig use set to 0%


CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND
LINEAR
"HEY, I DID A REGRESSION."
QUADRATIC
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPERING OFF!"
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY!"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
LINEAR, NO SLOPE
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."
LOGISTIC
"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."
CONFIDENCE INTERVAL
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."
PIECEWISE
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."
CONNECTING LINES
"I CLICKED 'SMOOTH LINES' IN EXCEL."
AD-HOC FILTER
"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"
HOUSE OF CARDS
"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE— WAIT NO NO DON'T EXTEND IT AAAAAA!!"


THANK
YOU!