

Mining Data on Traumatic Brain Injury with Reconstructability Analysis

Martin Zwick

Systems Science Program
Portland State University
Portland Oregon

zwick@pdx.edu

IEEE SSCI2017 (Honolulu, Nov 27, 2017)

- ABSTRACT: This paper reports the analysis of data on traumatic brain injury using a probabilistic graphical modeling technique known as reconstructability analysis (RA). The study shows the **flexibility, power, and comprehensibility of RA modeling**, which is well-suited for mining biomedical data.
- One finding of the analysis is that education is a **confounding** variable for the Digit Symbol Test in discriminating the severity of concussion; another – and **anomalous** – finding is that previous head injury predicts improved performance on the Reaction Time test. This analysis was exploratory, so its findings require follow-on **confirmatory** tests of their generalizability.

1. Exploratory modeling with RA (Occam)

2. Results on Preece data set

Project information

- Brain Trauma Evidence-based Consortium: BTEC
- Funded by DoD via Brain Trauma Foundation & Stanford

Coauthors:

- **Nancy Carney**: OHSU, BTEC founder/previous head
- **Tracie Nettleton**: Research assistant
-
- Wayne Wakeland: PI of PSU BTEC effort
- Forrest Alexander, Peter Olson: Programmers

1. Exploratory modeling with RA (Occam)

- Most biomedical data analyses are **confirmatory**, testing only **specific** hypotheses. Since studies are expensive & time-consuming, it is useful to explore what else might be **discovered** in the data.
- **Exploratory** studies can find **unexpected** effects, especially **non-linear & many-variable interactions** (which should, however, then be tested in confirmatory mode with new data).
- Exploratory studies (by data analysts) are **unbiased**.

Why RA & Occam software

- Explicitly designed for **exploratory** modeling
 - Analyzes both **nominal** & **continuous** (binned) variables
 - **Easily interpretable; standard text input; web-accessible, emails** results to user
- Other statistical & machine-learning methods (log-linear, logistic regression, Bayesian networks, classification trees, support vector machines, neural nets) **not well designed for exploration, or have limited model types, or have difficulty** with **nominal** variables or with **stochasticity**

What RA is

- **Reconstructability Analysis (RA)** = Information theory + Graph theory
- **RA model** = a **hypergraph** applied to **data**
= a (joint or conditional) probability distribution **simpler** (fewer df) than the data, **capturing much** of the **information** in the data

Two types of RA explorations

- ***Neutral search*** (clustering): find relations among all variables
(not discussed here)
- ***Directed search*** (classification): **predict** DV from IVs. Want:
 - **High accuracy** (information captured) (**low error**) measured by
 - $\% \Delta H$ = % reduction of uncertainty (*like variance*)
 - $\% c$ = % correct in prediction (*a general measure*)
 - **High model simplicity** (**low complexity**) = low Δdf
 - Model selection criteria **trade off** these two objectives

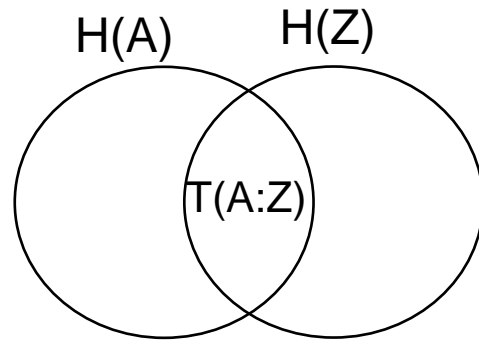
Model selection criteria

Tradeoff between accuracy & simplicity (error & complexity)

- *Conservative:* Bayesian Information Criterion (BIC)
- *Aggressive:* Akaike Information Criterion (AIC)
Incremental p-value (IncrP)
- AIC & BIC: linear combinations of error & complexity; BIC penalizes more for complexity: weights it by $\ln(N)$
- IncrP uses Chi-square p-values to pick models whose difference from -- & every incremental step from -- independence is statistically significant

Uncertainty reduction: primary measure

- Reduction of uncertainty (Shannon entropy), a simple example



	Z_0	Z_1	
A_0	$.67 * .5$	$.33 * .5$	$.5$
A_1	$.33 * .5$	$.67 * .5$	$.5$
df=3	$.5$	$.5$	

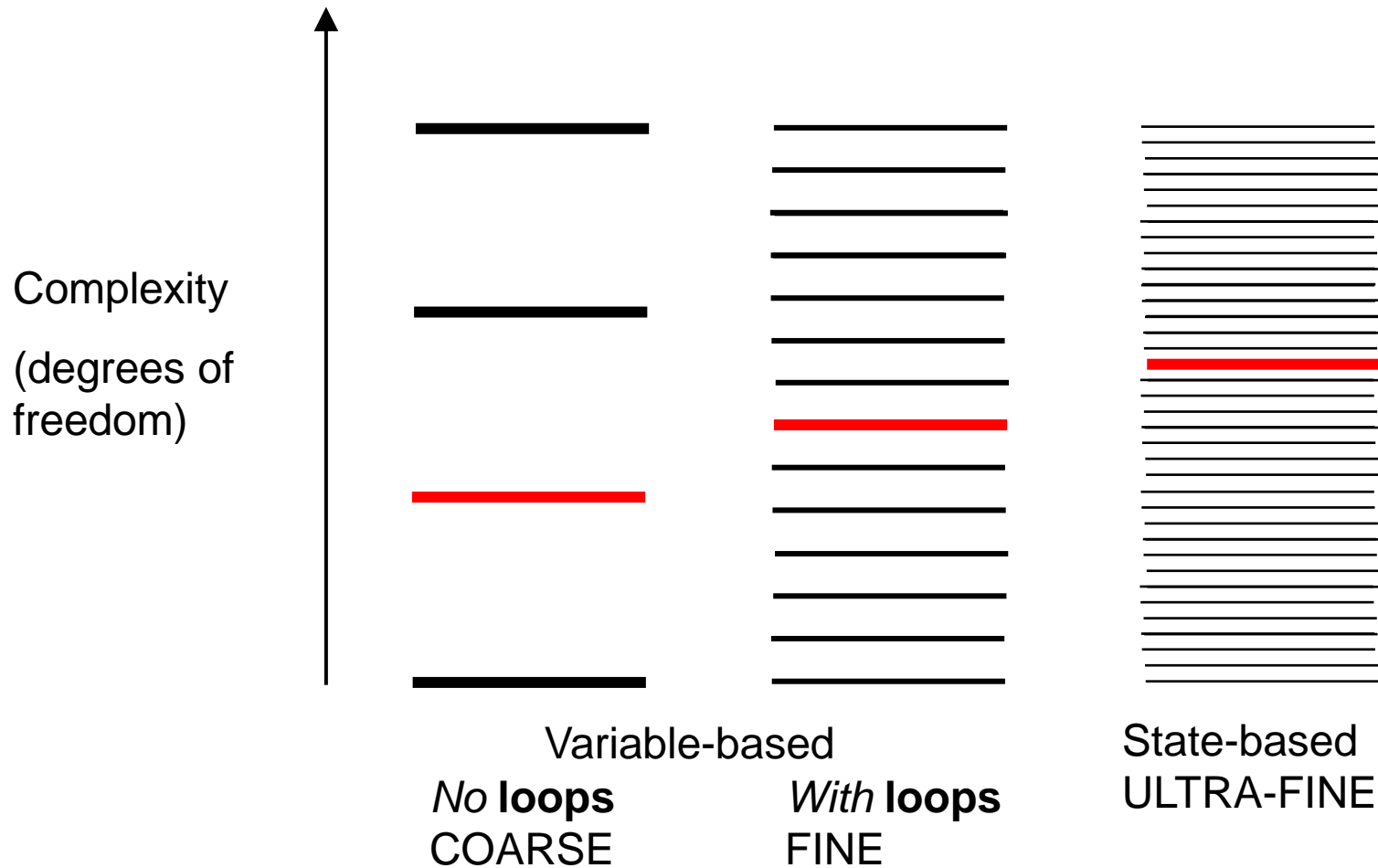
- $p(Z_1)/p(Z_0) = 1:1$, not knowing A $\rightarrow 2:1$ or 1:2, knowing A
- Reduction of uncertainty = $\Delta H(Z|A) = T(A:Z) / H(Z) = 8\%$
- 8% reduction in uncertainty (here) is *large* (unlike variance!)

Degrees of refinement of RA model search

3 degrees of search refinement (IVs: A,B,C...; DV: Z)

- *Coarse search*: variable-based models **w/o** loops, e.g., A B z
Fast, can handle *many* variables
- *Fine search*: variable-based models **with** loops, e.g., A B z : B C z
Slow, can handle 100s of variables
- *Ultra-fine search*: state-based models, e.g., A₂ B₁ z : B₀ z
Very slow, less than 10 variables

Degrees of refinement of RA model search



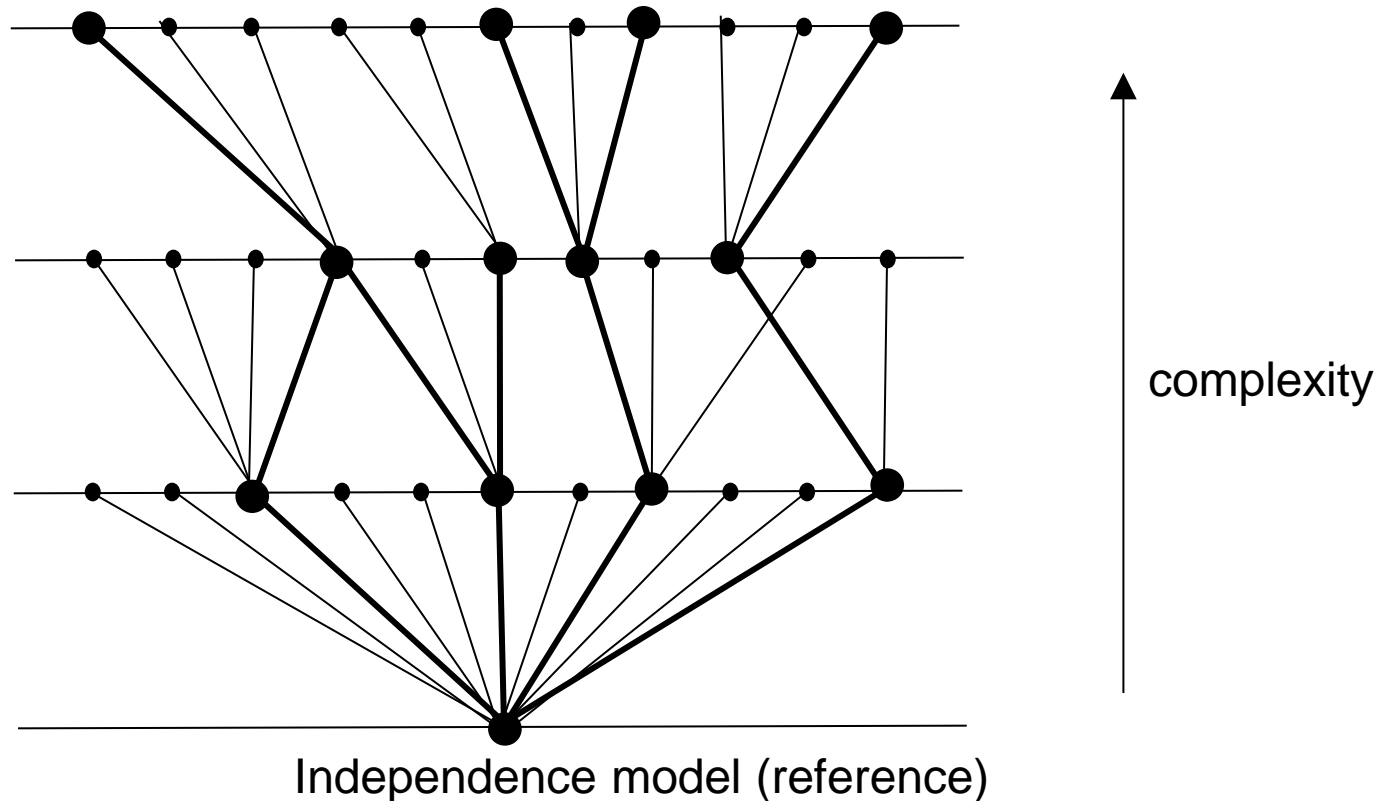
Combinatorial explosion of possible models

# variables	2	3	4	5	6	7
# neutral VB models (loops)	2	9	114	6,894	7,785,062	$2.4 \cdot 10^{12}$
For 1 DV:						
# directed VB models (loops)	2	5	19	167	7,580	$7.8 \cdot 10^6$
# directed VB models (no loops)	2	4	8	16	32	64
For binary variables:						
# neutral SB models (loops)	14	<i>even more severely exponential</i>				

NEED **INTELLIGENT HEURISTICS** TO DO **EXPLORATORY MODELING** with **52** variables (# variables in Preece data)

Can now explore a few 100 variables; if parallelized could deal with more.

Searching the space of possible models



2. Application to Preece data

- Automobile accident data: 52 variables
- Variable types
 - P = **patient** characteristics (17 variables)
 - Y = **symptoms** (25): subjective reports
 - G = **signs** (4): objective indicators
 - C = **cognitive** deficits (5)
 - N = **neurologic** deficits (1)
- N = 337; reduces to 175 or less if exclude missing data

Variables (1/3)

- Patient (**P**) variables (17)

pinjgrp,5,	pij	Injury Group: TBI patient or control (orthopedic injury)					
page,7,	pag	age					
psex,2,	psx	sex					
pyred,6,	pye	years of education					
pedlevel,8,	ped	highest level of education					
puhrsleap,5,	pul	usual # of hrs of sleep: less than or greater than normal (8 hr)					
precentill,3,	pri	recent illness 0 no 1 yes					
pmedication,3,	pmd	current medications 0 no 1 yes					
ppainkllr,3,	ppk	currently on painkillers 0 no 1 yes					
ppreheadinj,3,	pph	have they had previous head injury 0 no 1 yes					
pprecon,3,	ppc	previous concussion 0 no 1 yes					
pnumprecon,8,	pnp	how many previous concussions					
pdbqerror,13,	pqe	Driver Behavior Questionnaire self reported driving errors/violation					
pdbqviol,14,	pqv	Driver Behavior Questionnaire violations					
plitigat,4,	plg	was the case litigated					
prespacc,6,	pac	who was responsible for the accident					
pfsiq,5,	piq	full scale IQ calculated from national adult reading test					

Variables (2/3)

- Symptom (Y) variables (25)

ypainscale,5,	ypn	standard painscale used by hospitals					
yemoscale,5,	yem	sacle defining emotional state(0 no problems 1 few 2 moderate 3 many problems)					
ydassd,5,	ydd	Depression Anxiety Stress Scales: depression					
ydassa,6,	yda	Depression Anxiety Stress Scales: anxiety					
ydasss,4,	yds	Depression Anxiety Stress Scales: stress					
yheadache,6,	yhs	Rivermead headache					
ydizz,5,	ydz	Rivermead dizzy					
ynausea,5,	yna	Rivermead nausea					
ynoisensens,6,	yns	Rivermead noise sensitivity					
yslpdis,6,	ysd	Rivermead sleep disorder					
yfatigue,6,	yfa	Rivermead fatigue					
yirritable,6,	yir	Rivermead irritable					
ydepressed,5,	ydp	Rivermead depressed					
yanxious,6,	yax	Rivermead anxious					
yfrustrated,5,	yfr	Rivermead frustrated					
yforgetful,6,	yfg	Rivermead forgetful					
ypoorconc,6,	ycn	Rivermead poor concentration					
ylongthink,6,	ytk	Rivermead long time to think					
yblurredvis,6,	ybr	Rivermead blurred vision					
ylightsens,5,	yls	Rivermead light sensitivity					
ydoublevis,6,	ydv	Rivermead double vision					
yrestless,6,	yrs	Rivermead restless					
ydazed,5,	yaz	Rivermead dazed					
yrivermead,5,	ym	summation of Rivermead post concussion symptom questionnaire					
ycorrectedvis,3,	ycv	corrected vision					

Variables (3/3)

- Sign (**G**) & Deficit (**C, N**) variables (4, 5, 1)

ghrssleep,5,	ghl	number of hours of sleep, divided in less than normal normal=8hr and greater than normal							
ggcs,4,	ggc	Glasgow coma scale a measure of level of unconsciousness; lower = deeper unconsciousness							
gextcause,8,	gxc	external cause of the injury							
gpta,3,	gpt	post traumatic amnesia							
chazpt,10,	chp	hazard perception test measures how quickly potential driving hazards are predicted							
cnormsrt,6,	cnr	Spatial Reaction Time normalized for age and sex							
cspatialreac,6,	csr	Spatial Reaction Time tests how quickly patient responds to a visual stimuli							
cdgtcorrect,7,	cdg	Digit Symbol Substitution neuropsychological test							
cstarcan,4,	csc	Star Cancellation Test a test of spatial neglect							
nlogmar,4,	nlr	LogMAR Logarithm of the Minimum Angle of Resolution: a visual acuity test							

Occam input file (partial) (note missing data)

```
prece04slide - Notepad
File Edit Format View Help

:action
search

:nominal
#
subjectid,1, 0, id #1 P= patient(16) Y=symptom (25) G=sign (6) N=neurologic (1) c= cognitive (5)
study,2, 0, st #2 different format for the 2 studies
pinjgrp,5, 1, pij #3 which study the data is from (1)PAH N=55 or(2) RBWH N=282
page,7, 1, pag #4 p age
psex,2, 1, psx #5 p sex
pyred,6, 1, pye #6 p years if education
pedlevel,8, 1, ped #7 p highest level of education
ypainscale,5, 1, ypn #8 y standard painscale used by hospitals
yemoscale,5, 1, yem #9 y sacle defining emotional state(0 no problems 1 few 2 moderate 3 many problems)
ydassd,5, 1, ydd #10 y Depression Anxiety Stress Scales measure of DEPRESSION(subjective experience questionnaire)
ydassa,6, 1, yda #11 y Depression Anxiety Stress Scales measure of ANXIETY(subjective experience questionnaire)
ydasss,4, 1, yds #12 y Depression Anxiety Stress Scales measure of STRESS (subjective experience questionnaire)
ghrssleep,5, 1, ghl #13 g number of hours of sleep, divided in less than normal normal=8hr and greater than normal
puhrsleep,5, 1, pul #14 p usual number of hours of sleep, divided in less than normal normal=8hr and greater than normal
precentill,3, 1, pri #15 p recent illness 0 no 1 yes
pmedication,3, 1, pmd #16 p current medications 0 no 1 yes
ppainkiller,3, 1, ppk #17 p currently on painkillers 0 no 1 yes
ppreheadinj,3, 1, pph #18 p have they had previous head injury 0 no 1 yes
gpreloc1gth,7, 1, gpl #19 g how long unconscious
gprecon,3, 1, gpc #20 g previous concussion 0 no 1 yes
pnumprecon,8, 1, pnp #21 p how many previous concussions. N = 16 there were only 7 different values so each code defines a raw value
ggcs,4, 1, ggc #22 g glasgow coma scale a measure of the level of unconsciousness lower score = deeper level of unconsciousness
chazpt,10, 1, chp #23 c hazard perception test measures how quickly potential driving hazards are predicted
pdbqerror,13, 1, pqe #24 p Driver Behavior Questionnaire errors self reported driving errors and violations
pdbqviol,14, 1, pqv #25 p Driver Behavior Questionnaire violations

:data
# variable number, short name, number of missing
#1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
# ID st ij ag sx ye ed pn em dd da ds hl ul r1 md pk ph pl pc np gc hp ge qv xc
# 0 0 1 7 1 2 3 116 119 127 127 127 86 85 83 47 107 137 154 285 285 141 55 161 150 57
1 0 0 3 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 4 0 0 0 0 0
2 0 0 3 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 4 0 0 0 0 0
3 0 0 3 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 4 0 0 0 0 0
4 0 0 3 3 0 2 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
7 0 0 3 0 1 1 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
8 0 0 3 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
9 0 0 3 1 1 3 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
11 0 0 2 3 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
13 0 0 3 2 1 4 6 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
15 0 0 2 3 1 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
16 0 0 3 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
17 0 0 0 0 1 2 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
18 0 0 0 0 1 3 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
19 0 0 0 1 1 4 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
20 0 0 0 0 1 3 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
21 0 0 0 0 1 3 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
22 0 0 0 0 1 3 6 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
23 0 0 0 3 1 4 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
24 0 0 0 0 1 1 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
26 0 0 2 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
27 0 0 0 1 0 3 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
28 0 0 2 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
29 0 0 3 4 1 3 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
30 0 0 3 3 1 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
31
```

Directed searches

- Predicting **cognitive**, neurological deficit variables
- #bins excludes missing values

	#bins		N							
cdgtcorrect	6	Cdg	255	Digit Symbol Substitution neuropsychological test						
cnormsrt	6	Cnr	210	Spatial Reaction Time normalized for age and sex						
cspatialreac	6	csr	214	Spatial Reaction Time test: how quickly patient responds to visual stimuli						
nlogmar	3	Nlr	209	LogMAR	Log of Minimum Angle of Resolution (visual acuity)					

Cdg coarse, fine, ultra-fine searches

Predict Cdg: digit symbol substitution test (rebin |Cdg| = 2: ~ 50-50)

MODEL (IV component omitted)	Δdf	p	% ΔH	%c			
COARSE, single predictors					ΔBIC	N=240	
Pij Cdg	3	0.00	11.9	68.3	47.6	patient injury type	
Ped Cdg	7	0.00	11.7	65.0	5.9	education level	
Ggc Cdg	3	0.00	5.6	65.0	18.3	Glasgow coma scale	
Cnr Cdg	5	0.00	3.5	60.8	6.1	spatial reaction, normalized	
Pye Cdg	1	0.00	3.0	68.3	27.9	years education	
Csr Cdg	5	0.00	2.5	63.3	0.4	spatial reaction	
<i>Cdg (independence=reference)</i>	0	1.00	0.0	50.8	0		
FINE					Criterion	N=240	Cnr =6, incl missing
Pij Cdg : Pye Cdg	4	0.00	25.5	72.9	BIC		
Pij Cdg : Pye Cdg : Cnr Cdg	9	0.00	32.8	76.7	AIC		
Pij Cdg : Psx Cdg : Pye Cdg : Cnr Cdg	10	0.00	32.9	76.3	IncrP	sex	
ULTRA-FINE (state-based model)						N=175	Cnr =2, no missing
Pij₂ Cnr₁ Cdg : Pye₀ Cdg	2	0.00	13.5	68.6	BIC		
<i>Cdg (independence=reference)</i>	0	1.00	0.0	50.9			

Cdg ultra-fine (state-based) model 3/3

Model: $Pij_2 \text{ Cnr}_1 \text{ Cdg} : Pye_0 \text{ Cdg}$

Odds (high is good) = $\mathbf{Cdg}_1 / \mathbf{Cdg}_0(\text{model}) = p(\text{high digit score}) / p(\text{low score})$

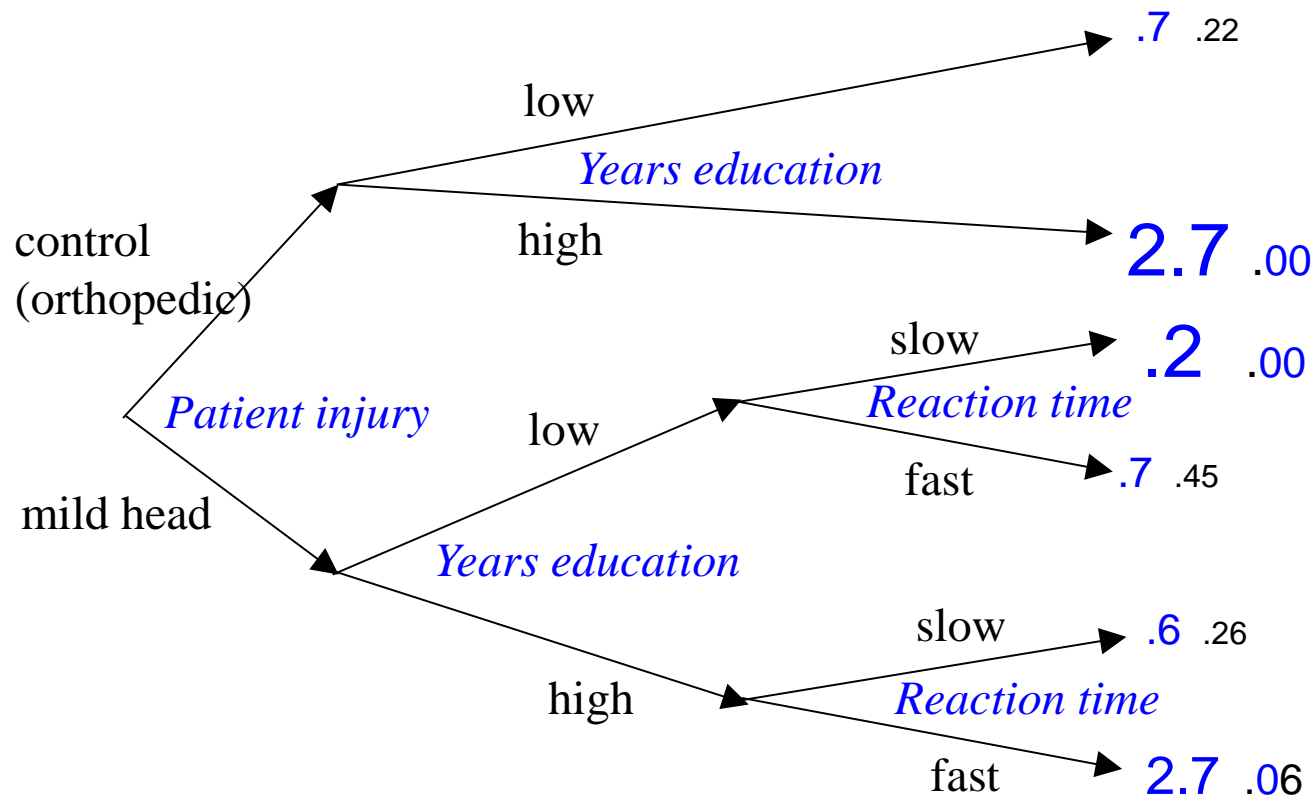
Pij_1 control (*orthopedic*), Pij_2 mild *head* injury; Pye_0 low years educ.; Cnr_0 = fast reaction

conditional probabilities of DV

IV states				data		model			
Pij	Pye	Cnr	N	Cdg ₀	Cdg ₁	Cdg ₀	Cdg ₁	Odds	p
1	0	0	18	0.50	0.50	0.59	0.41	0.7	.41
1	0	1	22	0.68	0.32	0.59	0.41	0.7	.36
1	1	0	38	0.21	0.79	0.27	0.73	2.7	.01
1	1	1	20	0.35	0.65	0.27	0.73	2.7	.05
2	0	0	15	0.53	0.47	0.59	0.41	0.7	.45
2	0	1	24	0.88	0.13	0.86	0.14	0.2	.00
2	1	0	18	0.33	0.67	0.27	0.73	2.7	.06
2	1	1	20	0.60	0.40	0.62	0.38	0.6	.26
175				0.49	0.51	0.49	0.51	1.0	

Cdg decision tree from conditional probabilities

Digit Symbol score odds (prob. high performance/ prob. low performance) & **p-values** relative to marginal prob. (odds = 1):



Cdg decision tree, verbally

- For all patients, **education predicts** performance on **digit symbol** test: more education predicts better performance.
 - Education is a **confounding** variable for digit symbol test in discriminating concussion, & must be controlled for
- For controls (orthopedic injury), **reaction time** does **not predict** digit symbol score.
- For TBI patients, fast reaction time predicts better digit symbol performance **beyond influence of education**.

Cnr coarse, fine, ultra-fine searches

Predict Cnr: reaction time, normalized by age, sex (rebin |Cnr| = 2: ~ 50-50)

MODEL	Δdf	p	% ΔH	%c		N=175		
COARSE, single component predictors								
Cdg Gpt Cnr	3	0.00	10.6	64.6	BIC, AIC	<i>Cdg = digit symbol test</i>		
Pph Cdg Gpt Cnr	7	0.00	13.1	66.9	IncrP	<i>Gpt = amnesia</i>		
<i>Cnr (independence=reference)</i>	0	1.00	0.0	50.9		<i>Pph = previous head injury</i>		
FINE								
Cdg Cnr : Gpt Cnr	2	0.00	8.8	64.6	BIC			
Pri Cnr : Pph Cnr : Cdg Gpt Cnr	6	0.00	14.7	70.3	AIC	<i>Pri = recent illness</i>		
Pye Cnr : Pph Cnr : Cdg Gpt Cnr	5	0.00	12.9	67.4	IncrP	<i>Pye = years education</i>		
ULTRA-FINE (state-based model)								
Pph₁ Cdg₁ Cnr : Cdg₀ Gpt₁ Cnr	2	0.00	12.4	64.8	BIC			
<i>Cnr (independence=reference)</i>	0	1.00	0.0	50.9				

Cnr ultra-fine model

Model: Pph₁ Cdg₁ Cnr : Cdg₀ Gpt₁ Cnr

Odds (high is good) = Cnr₀/Cnr₁(model) = p(fast = normal reaction)/p(slow)

Pph₁ previous head injury, Cdg₁ high digit score; Gpt₁ amnesia

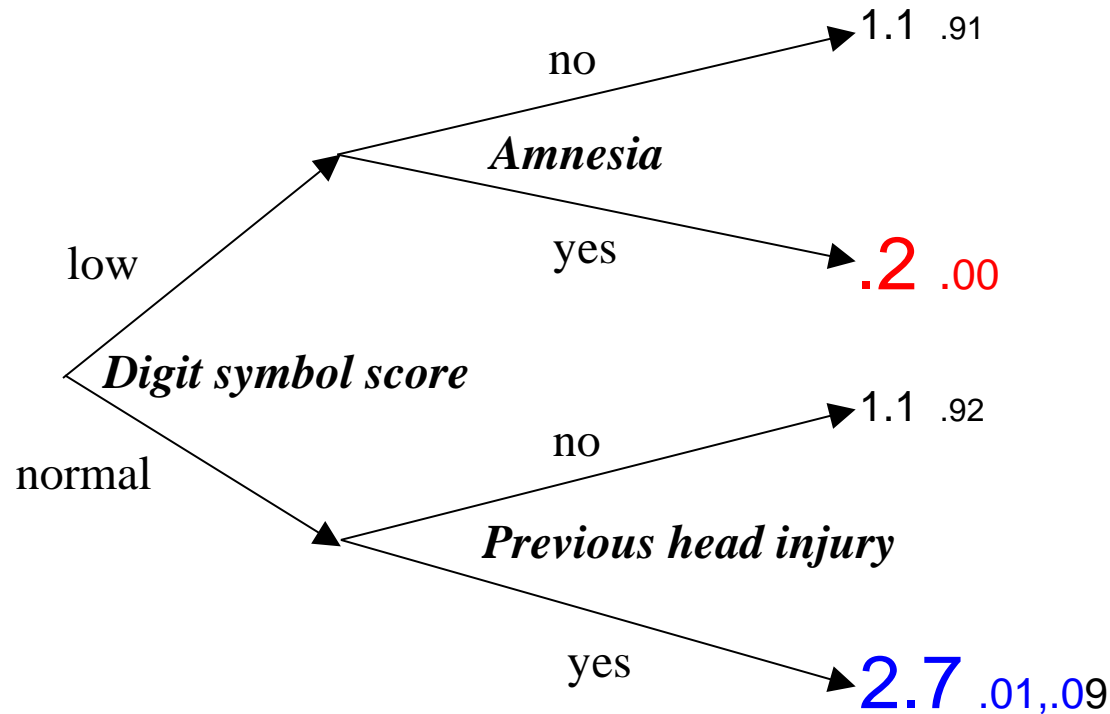
conditional probabilities of DV

IV states				data		model			
Pph	Cdg	Gpt	N	Cnr ₀	Cnr ₁	Cnr ₀	Cnr ₁	Odds	p
0	0	0	20	0.40	0.60	0.52	0.48	1.1	.92
0	0	1	19	0.16	0.84	0.16	0.84	0.2	.00
1	0	0	30	0.57	0.43	0.52	0.48	1.1	.90
1	0	1	18	0.17	0.83	0.16	0.84	0.2	.00
0	1	0	24	0.50	0.50	0.52	0.48	1.1	.91
0	1	1	13	0.61	0.39	0.52	0.48	1.1	.93
1	1	0	38	0.76	0.23	0.73	0.27	2.7	.01
1	1	1	14	0.64	0.36	0.73	0.27	2.7	.09
176				0.51	0.49	0.51	0.49	1.0	

Cnr decision tree from conditional probabilities

Reaction time **Odds** (probability **fast**/ probability **slow**)

& p-values relative to marginal prob. (odds = 1)



Cnr decision tree, verbally

- For **low** performance on **digit symbol** test, **amnesia** predicts **slow reaction time**.
- For **normal** performance on **digit symbol** test, **previous head injury** **increases** the probability of fast (**normal**) **reaction time**. THIS IS **ANOMALOUS**.
 - Need to see if it would be **replicated** in another data set.
 - Possible explanation: prior exposure to Reaction Time test introduces a **practice effect**.
 - If Reaction Time is so vulnerable to a practice effect, then it's probably **not an appropriate measure** to discriminate concussed from non-concussed patients.

Summary

- This secondary analysis yields intriguing new results.
- Since study is exploratory, these results are *tentative, needing confirmation* with other data sets.
- Study should be expanded to *additional data sets* (accident, military, sports), with *higher N, fewer missing data, new variable types* (imaging, genomic, proteomic).
- Work is *collaborative* with investigators who share data.
- Occam is *open* to researchers, web-accessible

RA (DMM) web page


<http://pdx.edu/sysc/research-discrete-multivariate-modeling>

Portland State Systems Science Graduate Program | Research: Discrete Multivariate Modeling - Mozilla Firefox

File Edit View History Bookmarks Tools Help Google

Portland State Systems Science Graduate Pr... +

www.pdx.edu/sysc/research-discrete-multivariate-modeling

Portland State
UNIVERSITY

Systems Science Graduate Program

myPSU | Contact SYSC | Quick Menu ▾

Courses | Program | Faculty | Students | Research | Resources

Search

PSU » System Science » Research » Research: Discrete Multivariate Modeling

Research: Discrete Multivariate Modeling

The methods used are also known in the systems literature as "reconstructability analysis" (RA). RA overlaps significantly with the fields of logic design and machine learning and with log-linear statistical modeling. The papers "[Wholes and Parts in General Systems Methodology](#)" and "[An Overview of Reconstructability Analysis](#)" listed below offer a concise review of RA methodology.

Artificial Life

Computational Intelligence

Discrete Multivariate Modeling

System Dynamics and Simulation

Neural Nets and Fuzzy Systems

Systems Theory and Philosophy

Projects

Theory/Methodology

OCCAM: RA software for data analysis & data mining

[Occam3](#) (web accessible; try it out)

[User manual](#) (PDF)

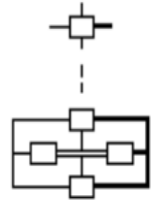
EDA: Extended Dependency Analysis

Heuristic RA search for loopless models.

[Download](#) executable, sample files, and documentation (for Windows)

RA utility programs

Below is the lattice of structures for a 4-variable *directed* system with 1 dependent variable (output).
Boxes = relations; lines = variables;
bold lines = the dependent variable.



- Thank you.