Casual Discovery: Producing a causal diagram in hopes of discovering unnoticed relationships and confounders.

Macalester College

Abstract

In this research project, we used the PLoS and Semnantic MEDLINE outputs in hopes of creating a causal diagram by building a correlation matrix through R and running it through Tetrad, a casual search program.

Vulvodynia

In this research project, we have chosen the topic of vulvodynia, using professor Devavani Chatterjea's research papers as reference.

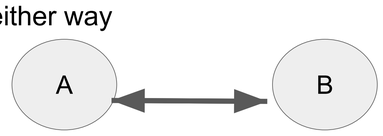
Background

There has been a great amount of research conducted on all sorts of different types of fields. There can be breakthroughs made through casual discovery, where we analyze the observational data from previous experiments and piece together information like a jigsaw puzzle.

What is a Causal Diagram?

- Huge "cause and effect" diagram
- Nodes (ideas) connected by edges (relationship)
- Since node "A" has an arrow pointed to node "B", A is a direct cause of B.

- Sometimes we are unsure and an arrow could be either way



Text Mining

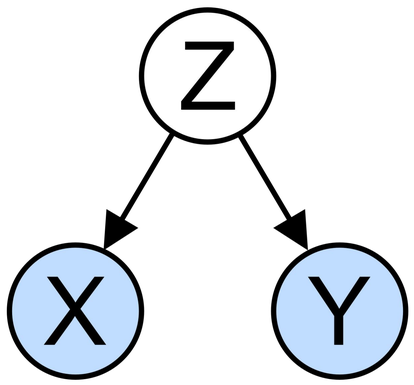
Goal: To identify Nodes and Edges

- At first we started to used a package in R titled "UDPipe" to text mine Professor Chatterjea's research paper.
- Specifically looked for nodes and edges.
- Wanted in form of noun phrases and association words.

[1] "Allergic sites were characterized by **mast cell** accumulation, **sensory hyper**-innervation and infiltration of regulatory CD4+CD25+FoxP3+ T cells as well as localized early **increases** in **transcripts** encoding Nerve Growth Factor and **nerve-mast cell synapse marker** Cell Adhesion Molecule 1"

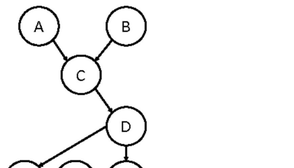
[2] "We previously provided **biological plausibility** for this association, demonstrating that single and **triple labiar skin** exposure to **hapten oxazolone** in pre-sensitized ND4 Swiss mice led to transient tactile sensitivity and an **increase in cutaneous nerve** density [9]"

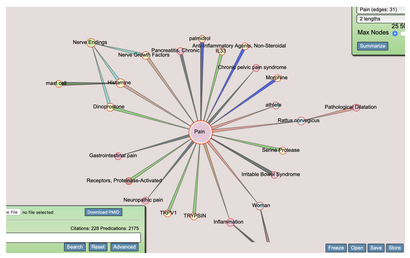
```
[[1]]

[1] "mast cell"    "sensory hyper"   "early increases" "in transcripts"  "nerve-mast cell"

[6] "synapse marker"

[[2]]

[1] "biological plausibility"    "triple labiar skin"      "hapten oxazolone"

[4] "increase in cutaneous nerve"
```



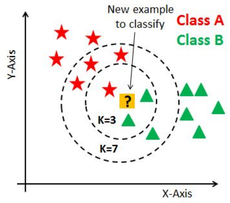
Semantic MEDLINE

Semantic MEDLINE is an application that uses natural language processing to extract predictions from research titles and abstracts from the Public Library of Science. The predictions come in the form of three words, an object, predicate and subject, and represented in a graph that has variables and links.





Using Chatterjea's keywords, we saved the diagram and, through data cleaning and reorganizing on R, we were able to work with this. There was a total of 4907 cases.



29202395	Vulvodynia	sosy	NEG_COEXISTS_WITH	skin disorder	disyn		We advocate that women with symptoms of GLV, with or without the presence of vulvar dermatoses, receive a therapeutic trial with
27129696	MAP Kinase Kinase Kinase	aapp	INTERACTS_WITH	Mitogen-Activated Protein Kinases	aapp		One of FITC targets was identified as MLK, a member of the mixed-lineage kinase family known to act as a mitogen-activated prote
25950702	Vulvodynia	sosy	PROCESS_OF	Woman	humn		Women who screened positive for PTSD had more than a two-fold increase in the prevalence of having vulvodynia (PR=2.37; 95% C
22999682	trichostatin A	antb	TREATS	Dermatitis, Allergic Contact	disyn		RESULTS: TSA treatment ameliorated skin lesion severity of DNFB-induced ACD.
23636033	Local Anesthetics	phsu	TREATS	Vulvodynia	sosy	NC	Successful therapy of vulvodynia with local anesthetics: a case report.
29854833	Interleukin-1 beta	aapp	TREATS	Mast-Cell Sarcoma	neop		Treatment of rat bone marrow MCs (BMMCs) with PAR4-AP also reduced the INOS, IL-1beta, P2X7, and tryptase protein and mRN
23635635	Enoxaparin	carb	TREATS	Vulvodynia	sosy		Enoxaparin treatment for vulvodynia: a randomized controlled trial.
20724927	Medication Management	hlca	TREATS	Overactive Bladder	disyn		SUMMARY: SNM is approved by the Food and Drug Administration for the treatment of idiopathic overactive bladder, urge urinary i
16613782	IL8RB gene	aapp	ASSOCIATED_WITH	Infiltration	patf		The present study investigated the role played by CXCR2 activity in tissue PMN infiltration and subsequent triggering of CHS.
29954457	Estrogens	horm	ASSOCIATED_WITH	Fibrosis	patf		RESULTS: The therapeutic groups treated with either UCMSCs-EVs alone or combined with estrogen exhibited a significant decrea
20059663	Muscle Hypertonia	sosy	PROCESS_OF	Woman	humn	D	Women with PVD had higher PFM tone, decreased PFM flexibility and lower PFM relaxation capacity compared with control women

We used UDPipe to annotate the sentence description to count nouns, verbs, etc. We hoped this would help train our machine learning model.

However, the immune regulatory role of TLR4 in AD remains to be defined.

token_id	token	lemma	upos	xpos
1	The	the	DET	DT
2	Role	role	PROPN	NNP
3	of	of	ADP	IN
4	Interleukin	Interleukin	PROPN	NNP
5	–	–	PUNCT	HYPH
6	19	19	NUM	CD
7	in	in	ADP	IN
8	Contact	contact	PROPN	NNP
9	Hypersensitivity	Hypersensitivity	PROPN	NNP
10	.	.	PUNCT	.

Cases from MEDLINE

Goal: Filter out "useless" cases

- When filtering useless cases, we often used the "stringr" package on r.
- We had to reorganize the data.
- Filter mostly by reducing predicates.

```
29899316|Air Pollutants|hops|PREDISPOSES|Malignant neoplasm of lung|neop|The result showed
that the emissions and       concentrations of air pollutant were higher in high-risk
regions, and the risk of       lung cancer was significantly elevated in such area.
29883874|Radon|elii|CAUSES|Malignant neoplasm of lung|neop|We systematically reviewed the
epidemiologic literature in order to assay the public's understanding about radon
and specifically, whether radon is known to cause lung cancer.
29883874|Radon|elii|CAUSES|Malignant neoplasm of lung|neop|Of those who have heard about
radon, the majority of       respondents in many studies did not know that radon causes
lung cancer.
29883874|Radon|elii|CAUSES|Malignant neoplasm of lung|neop|Specifically, these should
emphasize that radon causes lung cancer and        that household carbon monoxide
detectors do not detect it.
29737422|Radon|elii|CAUSES|Malignant neoplasm of lung|neop|Radon is the second leading
cause of lung cancer after smoking.
29869702|Malignant neoplasm of lung|neop|PROCESS_OF|Workers|humn|For lung cancer,
the relative incidence was highest among workers hired in shipbuilding between       1974
and 1984.
29743082|Malignant neoplasm of lung|neop|PROCESS_OF|Workers|humn|Significantly elevated
mortality from all causes, cardiovascular disease,       non-malignant respiratory disease
and lung cancer was observed among       silica-exposed workers, while elevated mortality
from non-malignant respiratory       disease and lung cancer was observed among smokers.
```

Assigning Coefficients

- Replace the predicates with predetermined correlation coefficients.
- Again, focused mostly on the sign of correlation (+/-)

```
"COEXISTS_WITH"] <- 0.35
"INHIBITS"] <- -0.32
"STIMULATES"] <- 0.40
"MANIFESTATION_OF"] <- 0.25
"PRODUCES"] <- 0.38
"PREVENTS"] <- -0.29
"DISRUPTS"] <- -0.32
"AUGMENTS"] <- 0.30
"OCCURS_IN"] <- 0.32
```

```
"","SUB","PRED","OB"
"1","Nifedipine","-0.29","Pain"
"2","Desipramine","-0.29","Pain"
"3","Vulvodynia","0.35","Localized pain"
"4","milnacipran","-0.29","Pain"
"5","follow-up","-0.29","Pain"
"6","Catechol O-Methyltransferase","-0.32","adrenergic system"
"7","Recombinant Macrophage Inflammatory Protein-2 Alpha","-0.32","mast cell"
"8","mast cell","0.38","CCL2 gene"
"9","KCNN4","0.3","mast cell"
"10","mast cell","0.38","Interleukin-13"
```

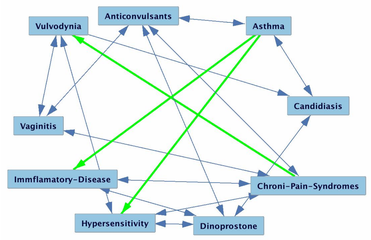
```
1000
X1 X2 X3 X4 X5 X6
1.0000
0.0312 1.0000
-0.5746 0.4168 1.0000
-0.5996 0.4261 0.9544 1.0000
0.8691 0.0414 -0.4372 -0.4487 1.0000
0.6188 0.0427 -0.1023 -0.0913 0.7172 1.0000
```

Tetrad

A casual search program. Our first attempt, we hand composed a correlation matrix.

	Anticonvulsants	Candidiasis	Dinoprostone	Hypersensitivity	Asthma	Chronic Pain S	Vulvodynia	Vaginitis	Immflamatory Di	Baclofen
Anticonvulsants	1									
Candidiasis	0	1								
Dinoprostone	0.3	0.3	1							
Hypersensitivity	0.2	0.34	-0.3	1						
Asthma	-0.24	0.31	0	0.4	1					
Chronic Pain Syndromes	-0.4	0.32	0.15	0.41	0.1	1				
Vulvodynia	-0.2	0.5	0	0.5	0	0.6	1			
Vaginitis	0.32	0.35	0.2	0.45	0	0.3	0.6	1		
Inflammatory Disease	0.2	0.31	-0.21	0.4	0.39	0.35	0.32	0.34	1	
Baclofen	0.31	0	-0.2	-0.23	-0.3	-0.2	-0.2	0.08		

Above is the input for the tetrad program.



- This chart shows many relationships among variables.
- There is still a good amount of double edges
- This process was not automated, meaning we had to hand make everything by hand.

Utilizing KNN

- Initially, we hand labeled 100 cases, PC (positively correlated), NC (negatively correlated), U (unknown) or I (incorrect), by reading the abstract of the research paper.
- We were more interested in the direction of the correlation rather than the strength of correlation (whether it was positive or negative)

RELN gene	STIMULATES	**Aromatase**	Our data provide evidence of a local increase of **aromatase** expression by **reelin**	PC
Tryptanthrine	INHIBITS	**Histamine**	Also, **TR** significantly reduced the serum levels of **histamine** and IL-4 in the AD model.	NC
histone deacetylase 3	ASSOCIATED_WITH	**Dermatitis**	We investigated the role of HDAC3 in allergic skin inflammation.	U

Predicted	Observed
PC	NC
NC	NC
PC	PC
U	U
U	NC
PC	NC
I	I
U	PC
NC	U
NC	I
PC	PC
PC	I
NC	PC
PC	NC
PC	NC
PC	I

After working with UDPIPE, we used the predicate and number of nouns to see if our model could predict whether it had a correlation or not. We tested the KNN with our 100 hand labeled cases.

- As we can see from the KNN model output, our prediction model is not so accurate.
- Overall on average it resulted in about 47% accuracy.

Improving Model

- We tried looking for patterns within consecutive nouns
- Looked into cases where we excluded nouns that were left on their own
- We would hope to compile a gigantic list of words that would give us hints in positive or negative association

Acknowledgement

- Thank you Professor Leslie Myint for the guidance
- Appreciations to Jr. Faculty-Hub Summer Research Fund for funding the project