## **Neuropathic pain diagnosis**

Source-

NeurIPS-2019-neuropathic-pain-diagnosis-simulator-for-causal-discovery-algorithm-evaluation-Paper

## **Framing the question**

### **What**

* In this work, the authors build a simulator in the neuropathic pain diagnosis setting for evaluating causal discovery algorithms.
* The simulator is based on ground-truth causal relations regarding the domain knowledge, and its parameters are estimated with a real-world dataset.
* It contains 222 nodes and 770 edges establishing complex real-world challenges. Our simulator can generate any amount of synthetic records that are indistinguishable from real-world records judged by physicians.
* The simulator can also simulate practical issues in causal discovery research such as missing data, selection bias, and unknown confounding. They can produce infinite samples without jeopardizing the privacy of the patient.
* causal relations for neuropathic pain diagnosis contains symptom diagnosis(discomfort of the patient), pattern diagnosis (identifies symptoms patterns), pathophysiological diagnosis(cause of symptoms where discoligmentous injury is most common factor in diagnosis), or radiculopathy(nerves do not work properly)

## **How**

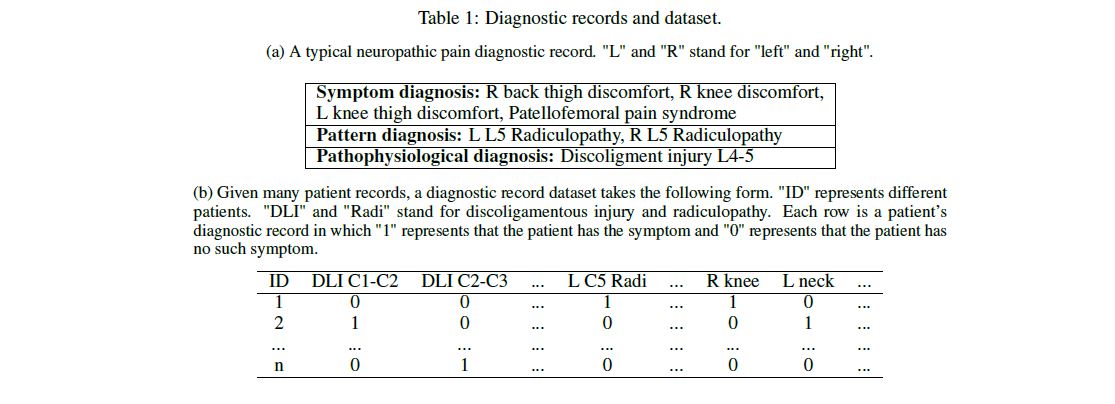
* Evaluated using major causal discovery algorithms, including PC, Fast Causal Inference (FCI), and Greedy Equivalence Search (GES) with simulated data under different settings
* This is also partnered with domain experts in disciplines such as biology and physics can provide information about well-understood causal influences in some specific scenarios. This gives us opportunities to utilize domain knowledge to reveal ground-truth causal relations and build realistic simulators. In this way, we can generate data from simulators and use such benchmark datasets for the evaluation of causal discovery algorithms

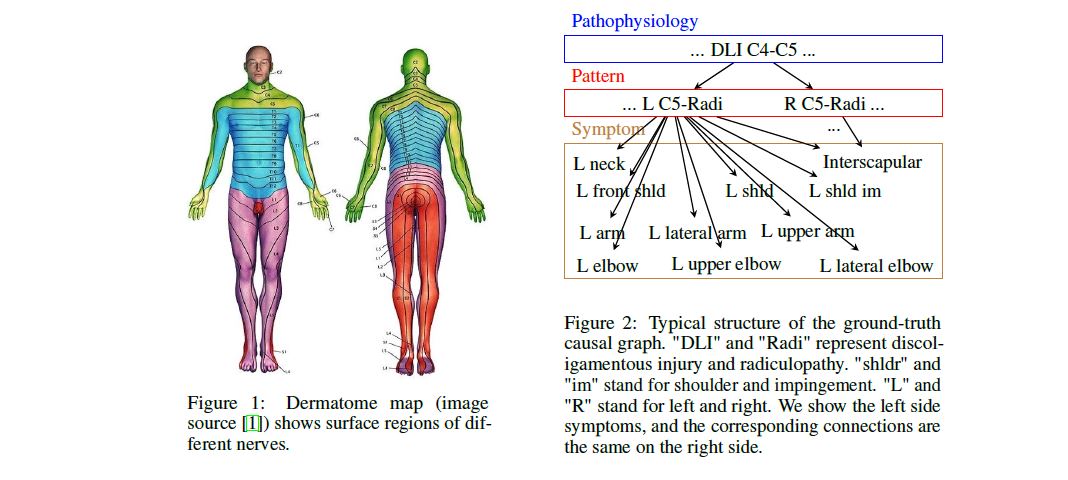
## **Why**

* Many real-life decision-making processes require an understanding of underlying causal relations. For example, understanding the cause of symptoms is essential for physicians to make correct treatment decisions; however, it is generally infeasible or even impossible to do interventions or randomized experiments to verify these causal relations. Therefore, causal discovery from observational data has attracted much attention.
* Presents a neuropathic pain diagnosis simulator for evaluating causal discovery algorithms. As one of the most important healthcare issues, neuropathic pain is well-studied in bio-medicine and has well-understood causal influences

## **Implementing the model**

### **The Neuropathic pain simulator**



****

The above is neuropathic pain diagnoses causal relations

* Neuropathic pain diagnoses mainly contain symptom diagnosis, pattern diagnosis, and pathophysiological diagnosis.
* In general, neuropathic pain symptoms in symptom diagnosis are mainly caused by radiculopathies (Radi) in the pattern diagnosis, and radiculopathy is mostly caused by discoligamentous injuries (DLI) in the pathophysiological diagnosis.
* For example, some of the causal relations are shown in Figure 2. DLI C4-C5 causes left side C5 radiculopathy and right side C5 radiculopathy. Left side C5 radiculopathy further causes symptoms at the left front shoulder, the left lateral arm, etc. We see that these locations are consistent with the dermatome map in Figure 1. Despite that there are other causes of neuropathic pain symptoms and radiculopathies such as tumors and diabetes, they rarely appear in primary care. Therefore, we focus on the causal relations among the discoligamentous injuries, radiculopathies, and neuropathic pain symptoms in this work.
* It is generally infeasible to verify causal relations as there is a lack of causal discovery algorithms
* It represents a complex real-world scenario with more than 200 variables and around 800 well-defined causal relations.  
    
   Our simulator is flexible and can be used to generate data with different practical issues, such as confounding, selection bias, and missing data

**Neuropathic Pain Diagnosis Simulator**

Real-world diagnostic records. To make our generated records close to the real-world scenario, we learn parameters from a dataset including 141 patient diagnostic records. 2 These patients’ diagnostic records are represented as a table of binary variables as shown in Table 1b. The variables in the pathophysiological diagnosis consist of the craniocervical junction injury and 26 discoligamentous injuries; the variables in the pattern diagnosis include 52 radiculopathies; the variables in the symptom diagnosis contain 143 symptoms. Similar to the real-world diagnostic records, the columns of generated records are the mentioned variables, and the rows represent the synthetic patients.

**Parameter estimation of the causal graph.**

* We estimate the Conditional Probability Distribution (CPD) of each variable given its parents in the causal graph with the real dataset. We compute the CPD of a variable X by P(X | Pa(X)) = P(X,Pa(X)).
* Both L4 and L5 radiculopathies can cause knee pain. The chance that a person with both L4 and L5 radiculopathies feels knee pain is higher or equal to the chance that a person with either one of the radiculopathies feels knee pain. In other words, P(X = 1 | Pa1(X) = 1, Pa2(X) = 1) >=P(X = 1 | Pa1(X) = 1) and P(X = 1 | Pa1(X) = 1, Pa2(X) = 1) >= P(X = 1 | Pa2(X) = 1), where Pa1(X) and Pa2(X) are L4 and L5 radiculopathies and X is knee pain. Given all the conditional probability and marginal probability distributions, we use ancestral sampling to sample neuropathic pain diagnosis data of synthetic patients

## **Testing the model**

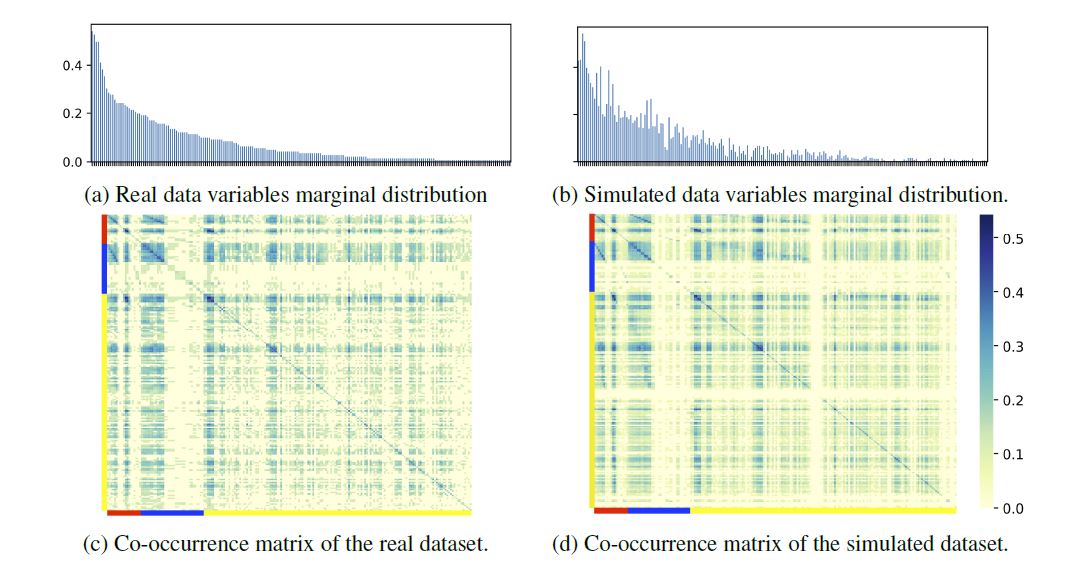
****

Figure: Comparison of the marginal distributions and the co-occurrence matrices of the real and simulated datasets. The orders of variables are the same in Panel (a) and (b). In Panel (c) and (d), the red color represents pathophysiological diagnosis, the blue color represents pattern diagnosis, and the yellow color represents symptom diagnosis

**Simulating Data with Practical Issues of Causal Discovery**

* experimental results of applying causal discovery algorithms to these simulated data reflecting different real-world problems. Unmeasured Confounding. Most causal discovery algorithms assume that all variables of concern are observed. However, in most real-life applications collected datasets may not cover all factors. If there is an unobserved common direct cause of two or more observed variables, this may produce wrong causal conclusions.
* This problem is known as unmeasured confounding, which is one of the common issues that one is faced with when applying causal discovery algorithms. More specifically, deleting the simulated data of the pathophysiology diagnosis and the pattern diagnosis variables leads to confounding in the dataset because they have at least two direct effects.
* We can also introduce external variables as confounders in the data generation process. For example, we can add patients’ occupation as a confounder which is not included in the given causal graph. With such datasets, we can evaluate how unmeasured confounding influences the results of causal discovery algorithms.
* Using our simulator, we can easily generate data with different missingness mechanisms

## **Evaluating the algorithm**

* evaluate major causal discovery algorithms with datasets generated from our simulator. We first further evaluate the simulation quality by comparing the causal discovery results of baseline methods between a real-world dataset and a simulated dataset. One advantage of the simulator is that we can generate any amount of data. Thus, we can evaluate causal discovery algorithms with different sample sizes to show the asymptotic property of causal discovery algorithms. Next, we apply causal discovery algorithms to the simulated datasets

# **Follow up question**

* The synthetic data sampled relies on graph structures of datasets that tend to oversimplify the relationship, cause-effect pairs can also be used for just linear acyclic models.
* Causal relations among multiple variables and are commonly used for the evaluation; however, few pairs of ground-true causal relations are known/labeled by domain experts and the evaluation is not systematic.
* alternative like In machine learning, there are many simulators built for other disciplines. For example, reinforcement learning benefits from the simulator’s interaction with applications. which can be used for evaluating decision making
* Creating confounding variables and missing data and other causal problems not clear to me