# Robustly Extracting Medical Knowledge from EHRs: A Case Study of Learning a Health Knowledge Graph

**Motivation**

What is the problem being solved?

* accurate understanding of causal relationship between diseases and symptoms
* assess the robustness of medical knowledge extracted from EHRs
  + for which diseases and for which patients a health knowledge graph performs poorly
  + potential confounders

Why is it important?

* improve clinical workflow and better understand diseases
* narrow sample sizes of subsets of the data -> underfitting despite the larger scale of the entire dataset.
* confounders not measured by the data -> bias the reliability of resulting models
* algorithms / findings from algorithms not generalize to entirely different populations

What previous work exists?

* LR, NB, Noise OR (“Learning a Health Knowledge Graph from Electronic Medical Records”)
* evaluated against expert physician opinion and a manually curated health knowledge graph from Google -> noisy OR model had the best precision and recall

Why is the previous work insufficient to solve the problem?

* examine the associative rather than causal link between diseases and symptoms in the clinical records
* estimate the strength of an edge between diseases and symptoms by the distance of mentions in the clinical records

**Approach**

Datasets

1. ED dataset
   * 273,174 patient de-identified records (single patient visit)
   * 140,804 patients who visited the ED between 2008 and 2013
2. Complete records (CR)
   * 140,446 of the patients from the ED dataset
   * 7,401,040 notes

Data preparation

* treat every note in a patient’s file as a separate instance -> 7,401,040 instances
* split each patient’s file into a variable number of episodes -> capture associations across temporally close notes, but avoid spurious associations from unrelated notes
* episode = a sequence of notes in which any two consecutive notes are at most 30 days apart.

⇒ 1,481,283 patient episodes (median: 3 episodes/patient and 2 notes/episode)

* consider an entire patient record for each patient as a single instance, ignoring the time component of the medical record -> 140,446 patient instances = number of unique patients in the dataset

Extraction pipeline

* Extract positive mentions of 192 possible diseases and 771 possible symptoms
* Mentions stored as binary variables to indicate the observation of a disease or symptom
* ED: extract from structured data + unstructured data
* CR: extract unstructured notes using string matching
* “sufficient support for a disease” = at least 100 positive mentions ⇒ 156
* “sufficient support for a symptom” = at least 10 positive mentions ⇒ 491
* ED admission Patient records: age, gender

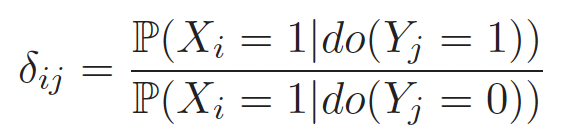
Disease predictability analysis

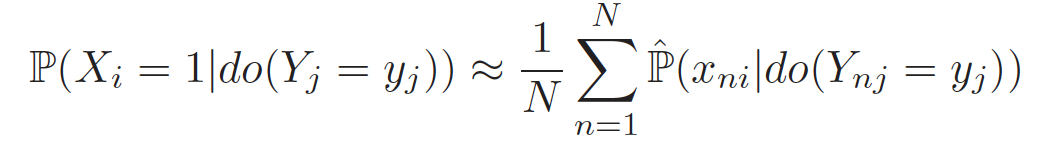
* Compare the existence of a disease-symptom edge in a proposed health knowledge graph and the same disease-symptom edge in the GHKG
  + precision
  + recall
* Sort symptoms by their importance scores -> include top 25 in graph as a binary edge between disease and symptom
* Individual diseases:
  + F1 score
    - compare top 50 diseases vs lowest 50 diseases by F1 scores
  + number of occurrences of the disease, average number of extracted diseases per patient, average number of extracted symptoms per patient, average patient age, and percentage female
    - identify abnormal disease
* Cross validation: Area under the precision-recall curve (AUPRC)

Non-linear method

Designate an edge from disease to symptom if the probability of observing a symptom is higher when we force a disease to occur than when we force a disease to not occur

Importance metric:





* Xi = symptom i
* Yj = diseases j
* x\_ni = binary features for whether patient stay n has symptom i
* y\_nj = binary features for whether patient stay n has diseases j
* do(Yj = 1) indicates intervening on diseases Yj

LR, random forest, and NB -> 3 folds CV -> AUPRC

**Results**

Noisy OR

* distribution of F1 scores of diseases span high and low values
* F1 scores decrease for diseases that have patients with many co-occurring diseases -> the multiple diseases are reducing the signal that noisy OR can learn
* diseases with lower F1 scores
  + abnormalities occur at much higher rates
  + more likely to have more co-occurring diseases, more co-occurring symptoms, more extremely low or extremely high ages, and more extremely low or extremely high female percentages
* number of occurrences not associated with high or low F1 scores: due to careful selection of diseases and symptoms in the problem construction

Demographics analysis

* age and sex does not meaningfully improve existing models
* Gender-specific graphs: reasonable AUPRC
* Graphs with groups of smaller sample size or potentially many confounders: perform poorly
  + Poor sample size -> model underfitting
  + older patients -> more diseases and additional health concerns
    - make learning the relationships between diseases and symptoms harder

Graphs learned on larger datasets

* The CR datasets decrease in granularity with more symptoms and diseases observed for each patient visit
* As granularity decreases
  + noisy OR: comparable performance -> hence most robust
  + NB: decreasing performance: violations in conditional independence assumption
  + LR: increasing performance: discern more signal from the noise
* Generalizability problem!

**Conclusions**

Data size does not always matter

* drastically larger sample size does not guarantee drastically better performance
  + lack of association between diseases and occurrence count
* limited resources of the emergency department-> focused and is biased towards more acute complaints -> chronic conditions and symptoms that are not contributory to the active problem may not be documented
* ED: adequate data set for most acute complaints, but inadequate for more chronic conditions
* CR: include notes from providers who manage more chronic conditions -> biased towards more chronic conditions and symptoms

Confounders may explain errors

* more diseases a patient has ⇒ more error
  + corresponding symptoms may be difficult to attribute to the correct disease
  + one disease may in fact increase the susceptibility to other diseases and also the resulting symptoms
* selection of diseases and symptoms themselves in the evaluation
  + original only selected diseases and symptoms relevant to the acute emergency department
  + expand to the CR dataset -> increase the number of observed diseases and symptoms
    - introduces the possibility that observed symptoms may be caused by diseases that are not in the scope of selection
    - LR improve off the higher density of information

Increased model complexity does not necessarily help

* inclusion of non-linear models
  + increase computational power and more expressive model parameterization
  + not able to compensate for the weaknesses of noisy OR
  + =/= improve AUPRC
* Continued success of noisy OR models: due to high likelihood of FN and low likelihood of FP in the dataset
* ⇒ important to understand potential sources of error to improve deficiencies in the existing models

**Contributions**

(1) Diseases with lower performance correlate with having more co-occurring diseases, more co-occurring symptoms, and fewer observations

(2) Heterogeneous impact of adding demographic data of age and gender

* demographic data improves performance for certain diseases and therefore some models.
* differences in graphs: drastically smaller sample sizes and differences in presentations

(3) Propose a method for learning a health knowledge graph from non-linear models based on principles from causal inference.

* causal methods do not yield an increase in performance
* noisy OR is robust in its performance in a wide range of datasets and settings

(4) We compare the health knowledge graph learned in the emergency department setting vs from the complete online medical records of the same patients.

* differences in coarseness -> change in ranking of performance between models
* potential sources of confounding

**Limitations**

Assessing model generalization

* should consider other potential settings including primary care or before a patient enters a hospital

Confounders and questions of causal inference from observational data

* Unmeasured confounders -> difficult to determine causal relations without additional experiments.
* Further work: include methods on improving causal inference methods or incorporating suggestions from our health knowledge graph into the clinical workflow.

Bipartite graph of binary edges between diseases and symptoms assumption

* Relax these constraints
* e.g. allowing one disease to cause another disease
* e.g. allowing for different edge strengths