# Learning a Health Knowledge Graph from Electronic Medical Records

**Motivation**

What is the problem being solved?

* develop an automated process to learn high quality knowledge bases linking diseases and network using noisy OR gates

Why is it important?

* Demand for clinical decision support systems in medicine and self-diagnostic symptom checkers
* Automatic compilation of a graph relating diseases to the symptoms that they cause -> speed up the development of diagnosis tools

What previous work exists?

* natural language processing: find relationships between diseases and symptoms from unstructured or semi-structured data
  + e.g. IBM’s WatsonPaths and the symptom checker Isabel: medical textbooks, journals, and trusted web content
  + e.g. electronic medical record (EMR)

Why is the previous work insufficient to solve the problem?

* rely on knowledge bases manually compiled through a labor-intensive process
* manual specification -> difficult to adapt to new diseases or clinical settings
* automatically derived using simple pairwise statistics
* EMR Problems:
  + text of physician and nursing notes is less formal than traditional textbooks -> difficult to consistently identify disease and symptom mentions.
  + presents real patients with all of the comorbidities, confounding factors, and nuances -> textbooks and journals often present simplified cases that relay only the most typical symptoms
  + associations between diseases and symptoms in the EMR are statistical -> easy to confuse correlation with causation.
  + Observation recorded is filtered through the decision-making process of the treating physician -> information deemed irrelevant may be omitted or not pursued -> information missing not at random

**Approach**

Concept extraction from electronic medical record

* extract positive mentions (negated concepts removed) of diseases and symptoms (concepts) from structured and unstructured data
  + Structured data: ICD-9 (International Classification of Diseases) diagnosis codes.
  + Unstructured data (de-identified free-text): chief complaint, Triage Assessment, Nursing Notes, and MD comments.
* string-matching to search for concepts
  + via common names, aliases or acronyms
    - obtained from Google health knowledge graph (GHKG) and Unified Medical Language System (UMLS) for diseases where the mapping was known
  + structured administrative data -> ICD-9 code
* Concepts, codes and concept aliases are mapped to unique IDs
* populate a co-occurrence matrix of size (Concepts) × (Patients)

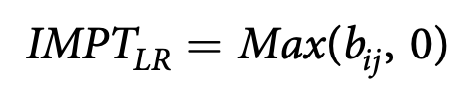
Statistical models: MLE for parameters

1. Logistic regression (LR)
   * binary classification
   * estimate parameters for each disease separately
   * L1 regularization: prevent overfitting and encourage sparsity (since most diseases only cause a small number of symptoms); hyperparameters chosen by 3-fold CV
2. Naive Bayes (NB)
   * provides a baseline of what can be inferred from simple pairwise co-occurrences
   * estimate parameters for each disease separately
   * Laplcian smoothing: prevent overfitting; hyperparameters chosen by 3-fold CV
3. Bayesian network modeling diseases and symptoms with noisy OR gates (noisy OR)
   * probabilistic model that jointly models diseases and symptoms
   * estimate parameters jointly

Constructing the knowledge graphs

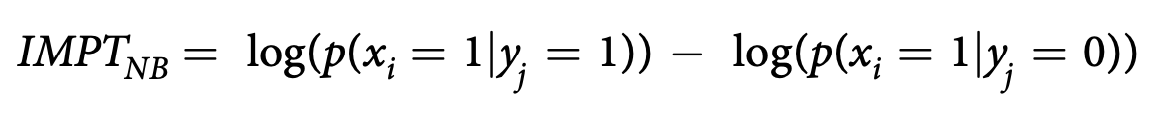
Importance measure: denote each model’s relative confidence that an edge exists between a pair of nodes -> sort symptoms for each disease by the importance measure

1. LR



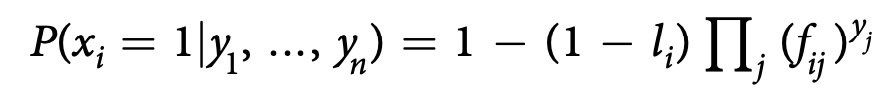
* bij = weight associated with symptom i in the logistic regression model fit to predict disease j
* if the appearance of a symptom made a disease more likely, then we believed that a corresponding edge exists in the graph
* minimum of 5 co-occurrences for any disease-symptom pair
* implicitly assume that diseases are independent

1. NB

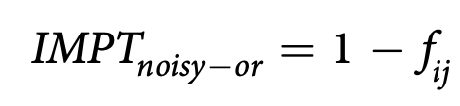


* xi = binary variable denoting the presence of symptom i
* yj = binary variable denoting the presence of disease j
* if the appearance of disease makes the observation of symptom more likely, we have higher confidence that an edge exists between the two
* multiplicative difference: relative risk idea
* minimum of 5 co-occurrences for any disease-symptom pair
* implicitly assume that diseases are independent

1. Noisy OR



* inherent noise: introducing failure and leak probabilities
* fij = probability of a disease yj that is present which fail to turn on its child symptom xi
* li = probability of a symptom being on even if all of its parent diseases are off



* the disease is likely to turn on the corresponding symptom
* no assumptions about the prior distribution of diseases

Evaluation

1. Comparison to the GHKG

* GHKG is non-exhaustive -> underestimates the precision of the models -> may denote edges as FP even when they may actually be correct and simply missing from GHKG
* only for ranking model purposes -> not true model performance measure
* assess models’ performance against a binary target (i.e. suggested edge is present or not in GHKG)

1. Evaluation by physicians

* pool top N results from each model
* rated by clinical evaluators using the 4-point scale
* ‘always’, ‘sometimes’ and ‘rarely’ -> positive; ‘never’ -> negative

1. Statistical methodology

* pooling -> Wilcoxon signed rank test
  + determine whether the differences in model precision were statistically significant
* inter-rater agreement measurement
  + one physician tagged all diseases and symptoms + second physician tagged 15 randomly selected diseases and corresponding symptoms
  + spearman rho correlation (measure inter-rater agreement)
  + calculate confidence intervals using bootstrapping

**Results**

1. LR

* not well calibrated for the task of constructing a knowledge graph
* low precision (vs GHKG)

1. Inter-rater agreement measure

* considerable agreement between evaluators
* generalizability of the results

1. Noisy OR and NB

* lower recall and higher precision in the clinical evaluation than suggested by the automatic evaluation.
* both surpass the recall of the GHKG in the clinical evaluation -> able to surface relevant symptoms that are not surfaced by the GHKG
* Noisy OR > NB
  + Wilcoxon signed rank test (p<=0.01)
  + the differences in precision were statistically significant for both evaluation frameworks (p<=0.01)

1. Differences between the edges

* additional symptoms (our graphs vs GHKG): infrequent symptoms that are not easily elicited from doctors, but are still medically and diagnostically relevant
* preciseness of language used (clinical evaluators vs GHKG)
* heightened severity of the edges (our graphs)
  + organically tailored for the emergency department data setting
  + selection bias towards higher acuity conditions and presentations
    - structural differences between our constructed graphs and the GHKG
    - provides a way of automatically adapting a knowledge graph across a range of different settings

1. Severity of symptoms

* Noisy OR model tends to rank general symptoms highly
* NB and LR wrongly suggest symptoms that are highly correlated with confounding factors and are not necessarily relevant to the parent disease
  + e.g. old age: diseases being correlated with one another
  + Noisy OR partly avoid disambiguating correlation and causation

**Contributions**

Significance

* Possible to construct a high quality health knowledge graph directly from EMR
* Noisy OR model with a precision of 0.85 for a recall of 0.6
* minimal post-processing
  + Add physicians filtering step
  + clinician reviews and rejects some of the edges suggested by the model (discard fewer than 2 out of 10 suggested edges)
  + ⇒ precision 100%
* surface ‘rare’ symptoms that are not easily elicited from doctors

How does this work compare to previous work?

* can create graphs from EMRs in any number of domains quickly and without any prior knowledge
* run regularly on current EMR data with existing knowledge graphs -> suggest new edges over time that were not previously known
* calibrate knowledge base created for one setting to an entirely different setting

**Limitations**

1. evaluation focused only on the ability of the proposed algorithms to recover known causal relations involving diseases and symptoms
   * unobserved confounding factors will affect the ability of all proposed approaches to infer correct causal relations
   * proving causality would require many additional experiments.
   * the algorithms should be construed as only providing candidate causal relations
2. Sometimes don’t have coverage for concepts
   * simplicity of the pipeline
   * 34% of the symptoms from the GHKG did not reach the required threshold of 10 positive mentions and were dropped due to insufficient support.
   * e.g. symptom ‘Bull’s Eye Rash’ for disease ‘Lyme Disease’
     + varying ways in which the symptom is recorded and punctuated (for example: “bullseye”, “bullseye rash”, “bull eye”, “bull’s eye”, etc.)
     + record it <10 times.
3. pipeline requires a base set of concepts to evaluate as potential nodes in the graph
4. Model assumptions
   * noisy OR and our baseline models don’t allow edges to exist between the symptom nodes: symptom conditional independence -> hinders complexity
   * may be reasonable to allow symptoms to cause other symptoms -> a softer classification into symptoms and diseases.
5. All models applied to the problem of knowledge graph construction are parametric -> restricted by their parametric form

**Next steps**

* more elaborate concept extraction pipeline to increase coverage and improve the subsequent graph (2)
* investigate models that do not assume symptom conditional independence in order to capture symptom dependency complexity (4)
* investigate models that are not constrained by parametric form -> have a closer match with the causal interpretation (5)