## **Supplementary Information -**

## **Learning a Health Knowledge Graph from Electronic Medical Records**

Maya Rotmensch<sup>1</sup>; Yoni Halpern, PhD<sup>2</sup>; Abdulhakim Tlimat, MD<sup>3</sup>; Steven Horng, MD MMSc<sup>3</sup>; David Sontag, PhD<sup>4‡</sup>

<sup>1</sup> Center for Data Science, New York University, New York, NY. USA

<sup>2</sup> Department of Computer Science, New York University, New York, NY. USA

<sup>3</sup> Department of Emergency Medicine, Beth Israel Deaconess Medical Center, Boston, MA. USA

<sup>4</sup> Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA. USA

# **Appendix 1: Methodology Details**

## **Model Assumptions**

Below we review some of the main assumptions that have biased our models and shaped their performance.

The defining assumption for naïve Bayes is of course its assumption of conditional independence between child nodes given the observation of the parent node. In our context this translates to symptoms being conditionally independent from one another given the parent disease. This is an over simplification, as is it possible that for a given disease, the appearance of one symptom makes another symptom more likely. For instance, for the disease Bronchitis, both congestion and headache are common symptoms. While both do not always occur when the disease is present, the presence of congestion raises the likelihood of a headache.

While Logistic Regression does not assume conditional independence between symptoms, it does rely on limited multicollinearity to perform well. Feature selection and regularization are common ways of limiting the performance issues associated with multi multicollinearity. In our paper we have opted to perform an extensive grid search of the proper regularization parameter to mitigate this problem. However, as we describe in the section below, optimizing a criteria focused on predicting disease from symptoms may be mismatched to the goal of learning a good health knowledge graph.

Additionally, for naïve Bayes and logistic regression each disease is modeled separately, resulting in an assumption of independence between diseases. One of the main advantages of the application of the noisy OR model in this setting is that it does not make any assumptions about the prior distribution of diseases  $P(y_1, ..., y_n)$ . This assumption is extremely important for our application, where many patients present with more than one disease (see Figure 3).

A central assumption made by the noisy OR model is the assumption of independence of effects. This assumption is an oversimplification, as the presentation of symptoms in a co-morbid scenario could very different than the presentation each disease produces separately. Nonetheless, a recent paper by Anand and Downs, which reviews the implication of this assumption on model performance, found that a model constructed with the noisy OR assumption "performs comparably" to a Bayesian network where no restriction were placed on the conditional probability distribution. The authors go on to claim that, at least for the setting that they consider, the minor differences that do exist between the two are "unlikely to be clinically important".

### **Modeling Details**

A threshold of at least 5 co-occurrences between disease and symptom pairs was chosen as a denoising measure for naïve Bayes and Logistic Regression, in order to allow us to compute the necessary statistic with some confidence. We chose a small number so as to not unfairly bias against rare diseases. While we did not do an exhaustive grid search, we experimented with several co-occurrence thresholds (0,2,5,10) and chose the smallest value that performed well on the automatic evaluation

### **Evaluation by physicians**

Given the resource intensive nature of the physician evaluation, we use a procedure in which the top N results from each model are pooled together and rated by clinical evaluators. For a given disease, the top two models from the automated evaluation, as well as the industry-provided graph, contributed to the pool of edges to be evaluated by the physicians. Each model suggests up to  $K+10$  symptoms, where K is the number of symptoms present in the knowledge graph for that disease. The 10 additional symptoms are introduced to allow for the possibility of the true graph being denser than the Google health knowledge graph. The figure below shows the user-interface though which the physicians tagged edges.



Figure 1: user interface for evaluation by physicians. The left column shows the list of diseases from which clinicians select a disease to be tagged. When a disease is selected, the upper panel provides a more detailed description of the selected disease. The middle column shows the symptoms that have yet to be tagged, and hover over functionality provides a detailed symptom definition. Lastly, the rightmost column shows the 4 categories into which symptoms can be sorted.

#### **Statistical Significance**

The Wilcoxon signed ranked test was used to determine the statistical significance of the differences in precision as it is non paramedic and does not assume the population is normally distributed. In the test, each pair of points represent the precision for a given condition by models a and b, where a and b represent our two best performing models. For both the automatic and clinical frameworks, model precision was calculated for the top K+10 symptoms suggested by each model.

#### **GHKG's Precision-Recall**

In the clinical evaluation we compare our resulting graphs against expert physician opinion using a precision-recall curve. We evaluate the Google knowledge health graph alongside our own models as a way to provide a realistic benchmark for performance. The GKHG does not contain continuous edge weights that are directly comparable to our generated importance measures. However, each edge in the graph is tagged with a frequency measure of 'frequent' or 'always' to indicate how often the symptom appears given the presence of the disease. To evaluate the GHKG against our model, we used the frequency tags as a measure of confidence in the existence of an edge. The star on the left represents the GHKG's performance for the 'high confidence' setting and includes only edges that were marked as 'always' in the graph. The star on the right represents the graph's performance in the setting in which it was allowed to suggest all edges that were deemed relevant, regardless of whether they occurred 'always' or 'frequently'.

#### **A causal explanation for model performance**

In reasoning broadly about the differences in model performance, we recall that the knowledge graph is, in its essence, a causal graph describing how diseases cause symptoms. In what follows we give a formal framework for characterizing what might make a good knowledge graph, and demonstrate that our results are consistent with what one would expect under this framework.

The casual query that we conjecture defines a good knowledge graph can be formalized via Pearl's 'do' operator  $2$  in which we intervene and set a disease to be either on or off, and see how this affects the likelihood of a symptom:

*IMPT*<sub>causal</sub> = 
$$
\frac{P(x_i = 1 | do(y_j = 1))}{P(x_i = 1 | do(y_j = 0))}
$$
 (5)

In other words, we believe that the weight of an edge  $i, j$  in the true graph is closely tied to the likelihood ratio of the symptom  $x_i$  manifesting, given that we *intervene* to set a disease  $y_i$  to be either on or off. By formulating the importance measure using Pearl's docalculus, we are able to disentangle correlations between a disease and a symptom that occur simply because the disease frequently occurs with other diseases that cause the symptom.

This target quantity (Equation 5) is not aligned with the objective optimized by logistic regression, which explicitly models  $P(y_j | X)$ . This misalignment may explain logistic regression's consistently poor performance.

Naive Bayes, which consistently performs better than logistic regression and worse than noisy OR, uses the logarithm of the formula shown in Equation 5 as its importance measure, but assumes that the interventional distribution  $P(x_i = 1 | do(y_i))$  is equal to the conditional distribution  $P(x_i | y_i)$ . However, one can show that these two are equal *only if the parent diseases are marginally independent*. To the extent that diseases are not independent, the estimate of the edge strengths provided by naive Bayes will be biased.

In contrast, the measure of edge importance used by noisy OR can be shown to correspond to a causal quantity closely related to the target quantity shown in Equation 5, specifically the do likelihood ratio of the symptom being *off* instead of on:

*IMPT*<sub>noisy-or<sub>ij</sub></sub> = 
$$
1 - \frac{P(x_i = 0 | do(y_j = 1))}{P(x_i = 0 | do(y_j = 0))}
$$
 (6)

Moreover, the learning method used by noisy OR, which optimizes the conditional likelihood of the observed symptoms given the diseases, makes no assumptions about the prior distribution of the diseases. In conclusion, we believe that this close alignment explains noisy OR's significantly better performance in our evaluations.

# **Appendix 2: The Constructed Graph**

Below is the full graph generated by the noisy or model with a threshold of K+10 edges, where K is the number of edges in the Google health knowledge graph for the corresponding condition. These are precisely the same edges for noisy or rated by the physicians in the clinical evaluation and displayed in Figure 3(b). The graph presented corresponds to a recall of 0.63 and a precision of 0.84.



































## **Appendix 3: Clinical Evaluation for Alternate Settings**

In the clinical evaluation presented in the main paper we took physicians' tagging of potential edges on a 4-point scale ('always happens', 'sometimes happens', 'rarely happens' or 'never happens') and binarized them, grouping results from the 'always', 'sometimes' and 'rarely' categories into the positive category and leaving the 'never' tag to be negative. Below we present the precision-recall curve of an alternate binarization according to which results from the 'always' and 'sometimes' categories are grouped into the positive category and 'rarely' and 'never' tags are grouped into the negative. The figure below shows that the relative performance of the models is consistent across binarization schemas, with noisy or outperforming naïve Bayes on both precision and recall.



**Figure 1:** Precision-recall curves for the clinical evaluation framework with alternate binarization.

## **References**

- 1. Anand, V. & Downs, S.M. Probabilistic asthma case finding: a noisy or reformulation. in *Amia* (2008).
- 2. Pearl, J. *Causality models, reasoning, and inference*, (Cambridge University Press,, Cambridge England ; New York, 2009).