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

Irene Y. Chen[†], Monica Agrawal[†], Steven Horng[*], and David Sontag[†]
[†]*Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA*
[*]*Department of Emergency Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA*
*Corresponding e-mail: iychen@mit.edu*

## Appendix A.

## 1. Datasets

### 1.1. *Demographic information*

For models that allow for continuous features, e.g. Logistic Regression, we augment the existing disease and symptom observations with the age and binarized sex values. For models that only allow binary features, we create age bracket (less than 21, 21-44, 45-64, 65-84, 85+) and include a binary indicator for each age bracket if the patient's age is within that bracket. For sex, we include one binary feature if the patient is female and one binary feature if the patient is male.

### 1.2. *Data distributions*

To better understand the datasets used, we present the distribution of diseases and symptoms for the dataset of emergency department (ED) patient visits and the three complete record (CR) datasets.

Next, to better understand the CR dataset, we examined the distributions of notes and episodes in Figure A5 - Figure A7. Figure A5 shows that there is a long tail in both the number of notes per patient, and the time span of notes available for each patient. In Figure A6, we see that the same is true for the number of created episodes per patient. Figure A7 shows that the vast majority of episodes are fewer than 5 days and 5 notes long.

## 2. Evaluation metrics

The GHKG contains the binary disease-symptom pairs for diseases that cause symptoms whereas our models for learning health knowledge graphs from EHRs return an importance metric $\delta_{ij}$ for symptom $i$ and disease $j$.

For individual disease analysis, we use the F1 score. For a given disease, find the top $E_j$ importance scores $\delta_{ij}$ where $E_j$ is the number of symptoms in the GHKG for disease $j$.

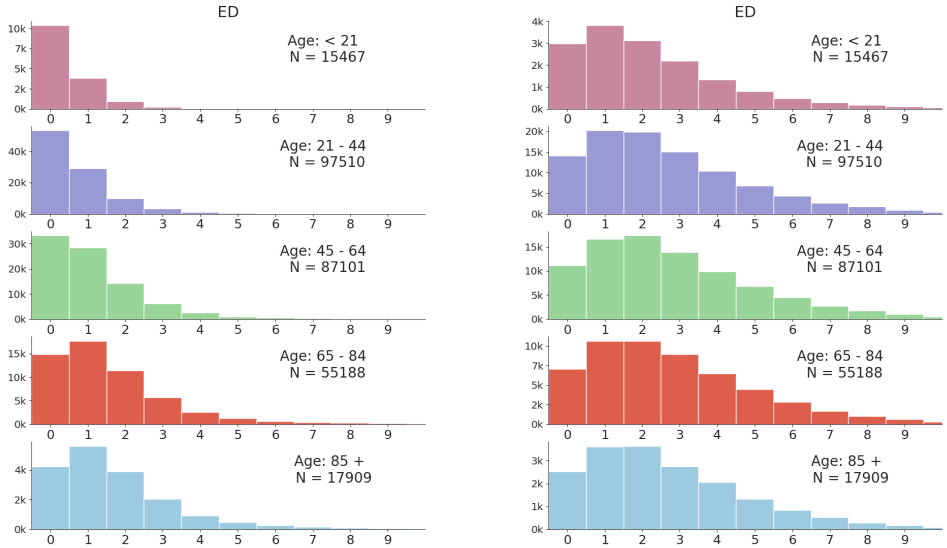


Fig. A1. Distribution of diseases (left) and symptoms (right) for emergency department dataset (ED).

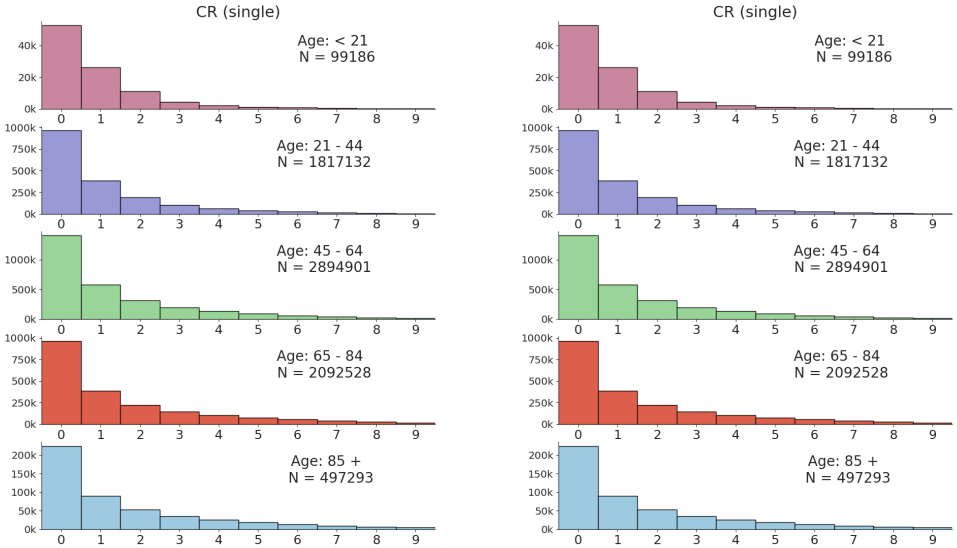


Fig. A2. Distribution of diseases (left) and symptoms (right) for complete record dataset split into single patient visits (CR, single).

For those symptoms, designate those disease-symptom pairs as selected. Then compute the F1 score $F_1$ according to $F_1 = \frac{2TP}{2TP+FP+FN}$ where: 1) $TP$ is the number of true positives or

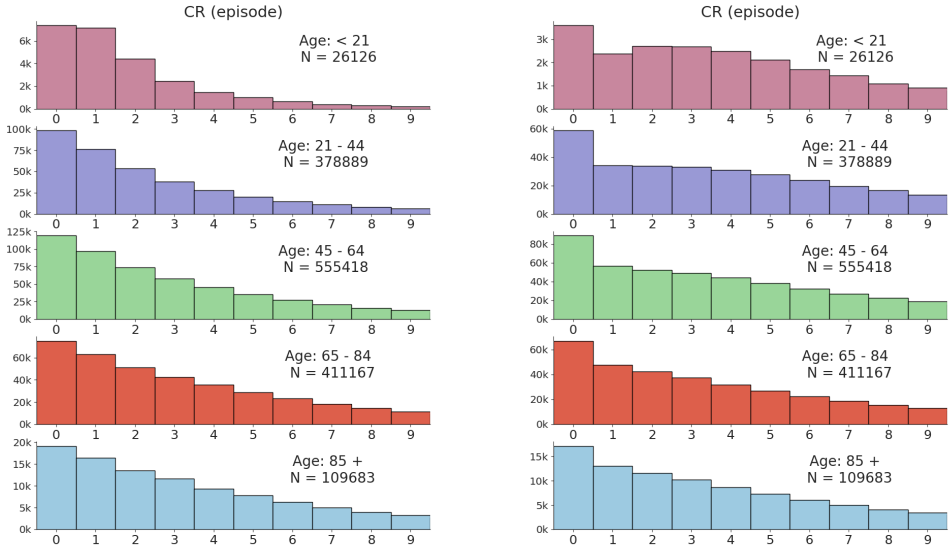


Fig. A3. Distribution of diseases (left) and symptoms (right) for complete record dataset split into episodes (CR, episode).

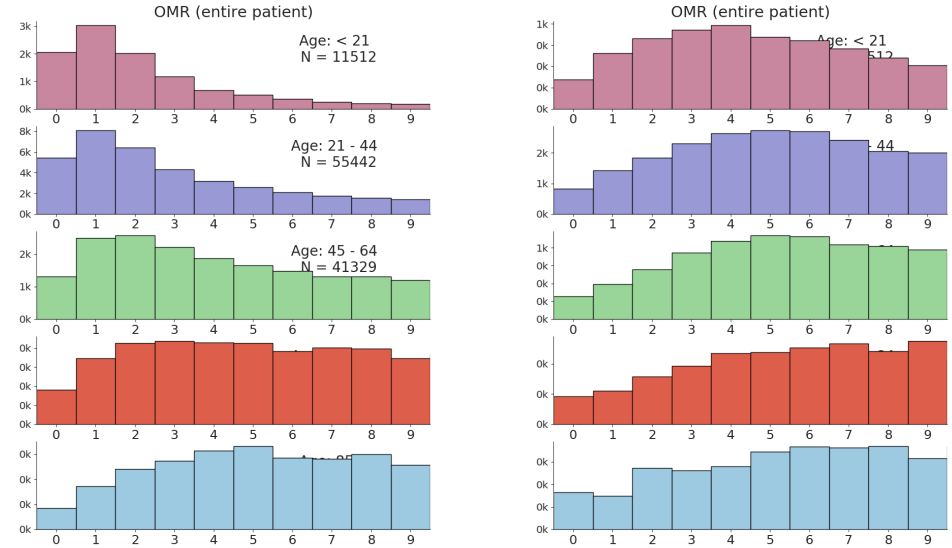


Fig. A4. Distribution of diseases (left) and symptoms (right) for complete record dataset split across entire patient histories (CR, patient).

disease-symptom edges that are selected by our model and also the GHKG, 2) $FP$ is the number of false positives or disease-symptom edges that are selected by our model but not

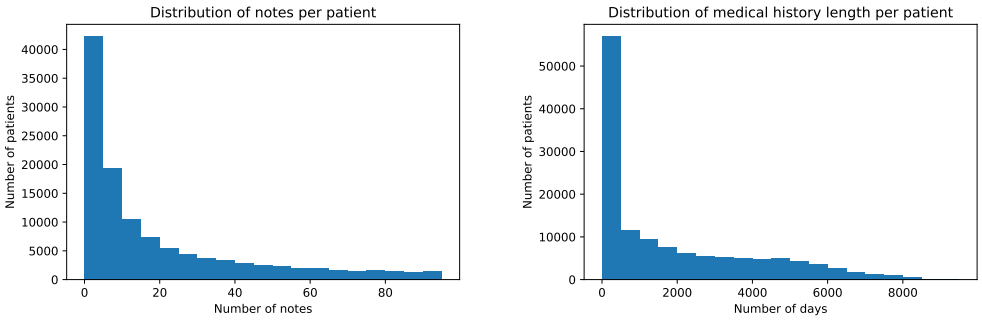


Fig. A5. Distribution of the number of notes per patient in the CR dataset (left), and the distribution of the time spanned in days for each patient by their CR record (right).

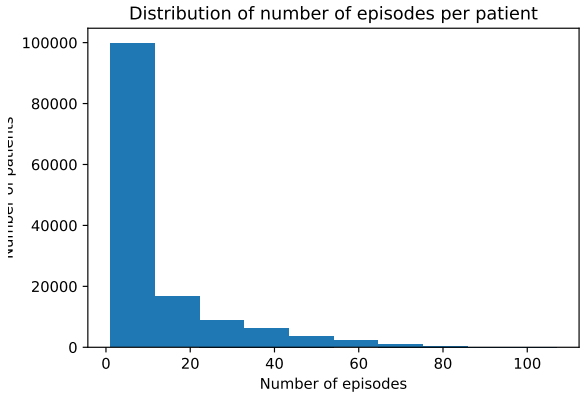


Fig. A6. Distribution of the number of episodes per patient in the CR dataset.

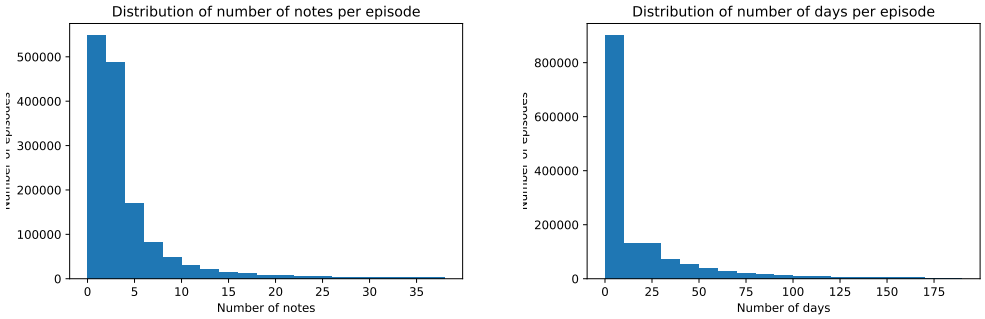


Fig. A7. The distribution of lengths of each episode in number of notes (left) and number of days (right).

selected by the GHKG, and 3) $FN$ is the number of false negatives or disease-symptom edges that are not selected by our model but are selected by the GHKG.

The area under the precision-recall curve (AUPRC) is computed as follows. For a given disease, find the top $E_j$ importance scores $\delta_{ij}$ where $E_j$ is the number of symptoms in the GHKG for disease $j$. For those symptoms, include those importance scores $\delta_{ij}$ in the precision-recall

computation and 0 otherwise. For all non-zero disease-symptom scores, compute the precision and recall according to precision $= \frac{TP}{TP+FP}$ and recall $= \frac{TP}{TP+FN}$ with differing thresholds to compute different coordinates of recall and precision. Because precision-recall curves may end at different locations,[1] we extend each curve to the point $(1, B)$ where $B$ corresponds to the precision value when recall is set to 1. In order words, when we select every edge, what is the resulting precision? $B$ then becomes the fraction of selected edges in the GHKG compared to the total possible edges. See Figure A8 for an illustration of example AUPRC curves and the extension to $(1, B)$. The area under the curve is then found with a trapezoidal approximation.

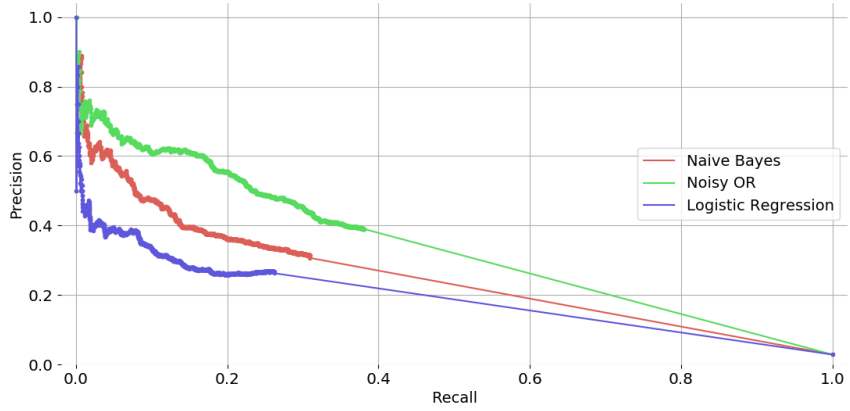


Fig. A8. Example AUPRC curves with added point to extend curves to recall=1. Area found with trapezoidal approximation.

## 3. Error Analysis

Here we give a more precise description of the disease error analysis performed.

- For every disease, count the number of patient visits where this disease was observed. We denote this value the number of occurrences. A disease is considered abnormal if the number of occurrences is then a standard deviation below the mean number of occurrences across all diseases. (`count`)
- For every disease, find all patient visits where this disease was observed. For each of these patient visits, total the number of observed diseases in each patient visit. For each disease, compute the mean number of diseases over all patient visits. We denote this value the mean number of extracted diseases. A disease is considered abnormal if mean number of extracted diseases is a standard deviation above the mean number of extracted diseases across all diseases. (`disease`)
- For every disease, find all patient visits where this disease was observed. For each of these patient visits, total the number of observed symptoms in each patient visit. For each disease, compute the mean number of symptoms over all patient visits. We denote this value the mean number of extracted symptoms. A disease is considered abnormal

if mean number of extracted symptoms is a standard deviation above the mean number of extracted symptoms across all diseases. (`symptom`)

- For every disease, find all patient visits where this disease was observed. For each of these patient visits, extract the patient age at the time of visit. For each disease, compute the mean age over all patient visits. We denote this value the mean patient age. A disease is considered abnormal if mean patient age a standard deviation either above or below the population mean. (`age`)
- For every disease, find all patient visits where this disease was observed. For each of these patient visits, extract the patient gender. For each disease, compute the percentage female over all patient visits. We denote this value the female percentage. Because the clinical records require all patients to be either male or female, we can find the inverse by subtracting female percentage from 1. A disease is considered abnormal if female percentage is a standard deviation either above or below the population female percentage. (`female`)
- any of the above abnormalities (`any`)

## 4. Disease predictability

We are interested in the predictive ability of our models to predict the disease from the symptoms. Although our main evaluation metrics use the GHKG, which has been manually curated by experts, the disease predictability can potentially shed light on whether we have extracted the correct symptoms for each disease.

We compute disease predictability using a 3-fold cross-validated area under the received operator curve (AUC),[2] searching over the same parameters as outlined in Section **??**. We report the average AUC using the best parameters. In Figure A9, we see that the relationship between logistic regression AUC and F1 score is not very strong. Additionally, in figure A10 we see a negative relationship between logistic regression AUC and average patient age. This finding points to investigating younger patients and exploring why predicting for them is more difficult than older patients. One hypothesis is that younger patients don't manifest symptoms as evidently as older patients — or that our extraction process omits symptoms that affect younger patients more.

## 5. Symptom predictability and causal method performance

In order to understand better the non-linear methods in Section **??**, we can also investigate the predictability of symptoms. That is, for observed diseases, how well can we predict the symptoms? Across the different link functions used, how much variance in symptom AUC is there? Similar to disease predictability, we use 3-fold cross validation to determine the symptom AUC. We report the average AUC found using the optimal parameters. In figure A11 we see that logistic regression and random forest AUC are correlated but can differ. Depending on the sample size and linearity of the underlying data generating function, different models may be better suited for different symptoms. One area for future research might be then to learn which predictive model to use for each symptom independently in order to build a more accurate health knowledge graph.

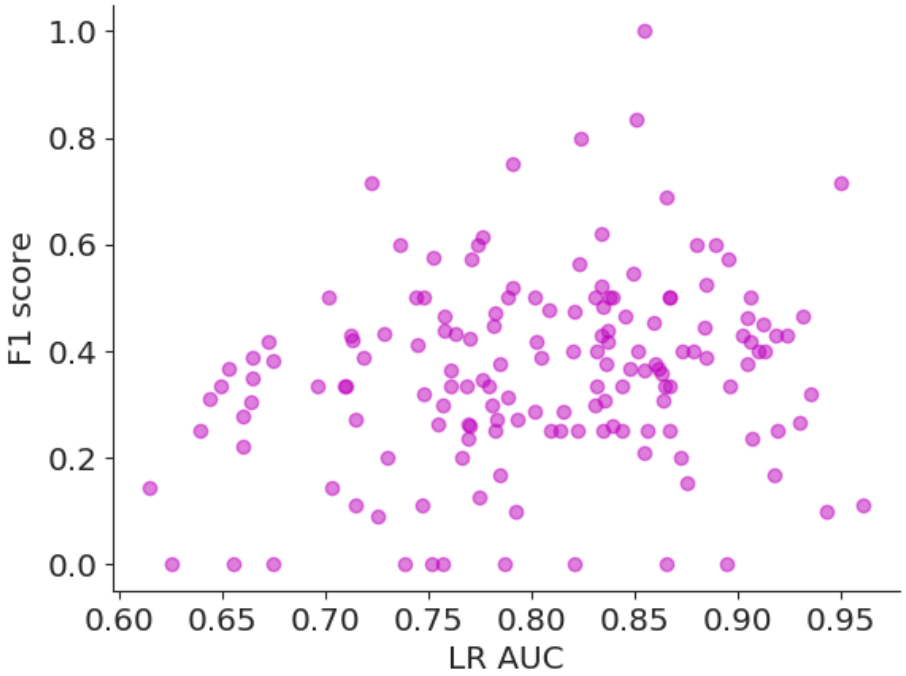


Fig. A9. Comparison of AUC and F1 score for logistic regression.

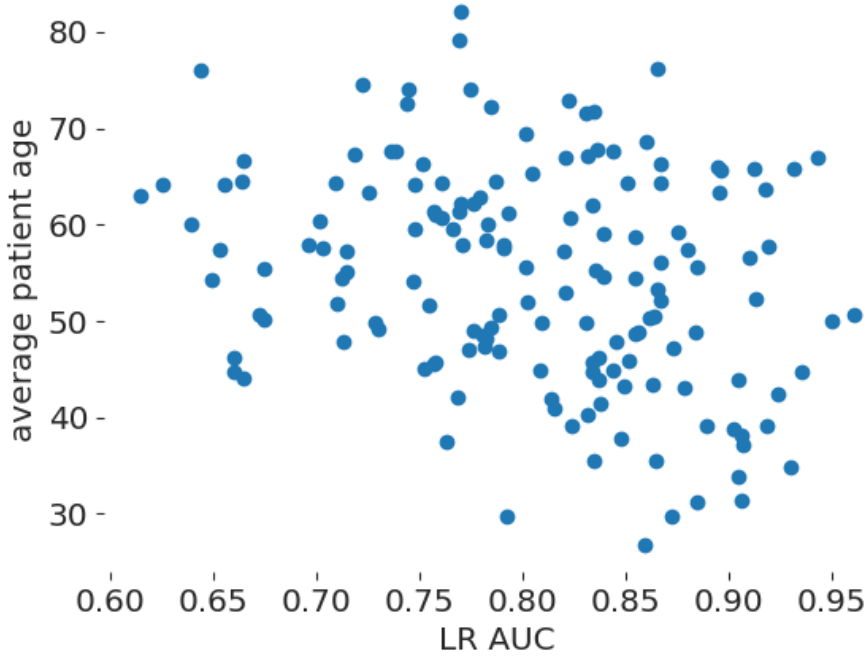


Fig. A10. Relationship between patient average age and logistic regression AUC.

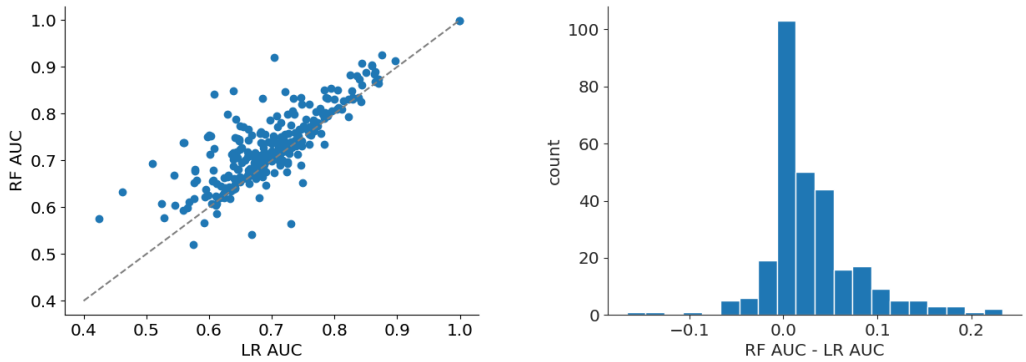


Fig. A11. **Left:** Relationship between logistic regression and random forest AUCs for the same symptom. **Right:** Histogram of differences between logistic regression and random forest AUC for the same symptom.

Additionally, we investigate if the random forest causal method ever outperforms the noisy OR method. Over the 156 possible diseases, we identify one disease where the F1 score for the random forest causal method outperforms the noisy OR method: breast cancer. For the other diseases, either the F1 scores match or the noisy OR method performs better.

## 6. CR dataset disease analysis

Here we present addition results on the CR datasets, specifically the disease analysis outlined in Section 3.

Table A1. Percentage of diseases with abnormalities learned on CR (single) data.

| | count | disease | symptom | age | female | any |
|---|---|---|---|---|---|---|
| top 50 | 16% | 12% | 0% | 10% | 4% | 38% |
| bottom 50% | 14% | 38% | 6% | 6% | 16% | 60% |

Table A2. Percentage of diseases with abnormalities learned on CR (episode) data.

| | count | disease | symptom | age | female | any |
|---|---|---|---|---|---|---|
| top 50 | 12% | 14% | 0% | 10% | 10% | 32% |
| bottom 50 | 18% | 40% | 4% | 14% | 16% | 57% |

## References

1. P. Flach and M. Kull, Precision-recall-gain curves: Pr analysis done right, in *Advances in neural information processing systems*, 2015.
2. T. Fawcett, An introduction to roc analysis, *Pattern recognition letters* **27**, 861 (2006).

Table A3. Percentage of diseases with abnormalities learned on CR (patient) data.

| | count | disease | symptom | age | female | any |
|---|---|---|---|---|---|---|
| top 50 | 10% | 14% | 2% | 14% | 10% | 36% |
| bottom 50 | 20% | 32% | 4% | 20% | 18% | 64% |