

Learning a Health Knowledge Graph from Electronic Medical Records

Ole Berg

November 19, 2024

Table of Contents

1. Motivation and earlier work
2. Goal and methods
3. Electronic medical record
4. Implementation
5. Evaluation
6. Discussion
7. Future improvements
8. Conclusion
9. Sources

1 Motivation and earlier work

- Demand for decision support systems in clinical settings
- Existing knowledge bases created manually or “*using simple pairwise statistics*” [1, p. 1]
- E. g. 15 person-years needed for *Internist-1/QMR* knowledge base
- Manually developed systems very brittle and difficult to extend
- Automatic compilation speeds up development of KBs
- *WatsonPath* by IBM and *Isabel* use NLP to find relations between diseases and symptoms in textbooks and journals

2 Goal and methods

- Utilize electronic medical record (EMR) to construct a knowledge graph
- Validation against Google health knowledge graph (GHKG)
- Three steps for knowledge graph generation:
 1. Data collection and preparation
 2. Learning of statistical models
 3. Transformation of models into knowledge graphs

3 Electronic medical record (EMR)

- EMR sometimes used interchangeably with electronic health record (EHR)
- Some authors distinguish between these terms
- EHR is patient-centric collection of EMRs [3, p. 4]
- EMR originally tool to store data *“of one or more pathological episodes concerning a patient”* [2, p. 121]
- Newer interpretation: EMR is information on a patient from one healthcare provider [3, p. 4]

3 Electronic medical record (EMR) (cont.)

- EMRs useful as data source because they represent diseases and their symptoms in a real-world environment
- Difficult data source for four reasons:
 1. Notes from physicians and nurses less formal
 2. Comorbidities, confounding factors and nuances present
 3. Associations between diseases and symptoms are statistical
 4. Pre-filtered by physicians

4 Implementation

4.1 Data collection and preparation

- Focus on positive mentions of diseases and their symptoms
- Structured data:
 - ICD-9 codes
- Unstructured data:
 - Chief complaint
 - Triage Assessment
 - Nursing Notes
 - MD comments
- Diseases (min. 100 mentions in data) and symptoms (min. 10) chosen from GHKG
- Mapping of extracted concepts to a concept ID

4 Implementation

4.2 Learning of statistical models

- Three statistical models:
 - Naive Bayes
 - Logistic regression
 - Noisy OR gates
- Parameter learning with maximum likelihood estimation
- L1 regularization used for logistic regression
- Laplacian smoothing used for naive Bayes

4 Implementation

4.3 Transformation of models into knowledge graphs

- Estimating the importance of edges (connections between diseases and symptoms)
- One importance measure for each statistical model
- Maximum of five symptoms per disease

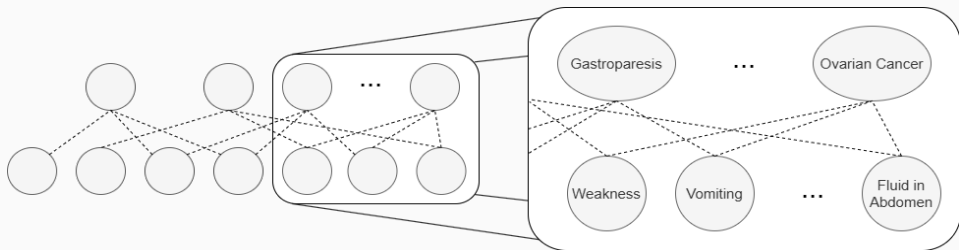


Figure 1: Resulting knowledge graph (own illustration according to [1, p. 4])

5 Evaluation

5.1 Methods

- Automatic evaluation against GHKG
- Assumption: GHKG is precise, but not complete
- Relative comparison between models, not an absolute evaluation
- Two best performing models and GHKG evaluated by physicians
- Physicians tag suggested disease-symptom edges on a 4-point scale ranging from “Always happens” to “Never”
- Binarization with “Never” as negative and other three as positive
- Precision-recall curve as evaluation measure

5.2 Results

- Naive Bayes and noisy OR perform considerably better than logistic regression
- Performance better when models compared to evaluations by physicians
- With recall of 0.5: Precision of noisy OR at 0.87, of naive Bayes at 0.8
- Conclusion: Noisy OR better than naive Bayes; statistically significant ($p = 0.01$)

6 Discussion

- Three kind of differences between edges suggested by the model and by GHKG:
 - GHKG focuses on information for web users
 - GHKG uses less precise language
 - Less severe edges in GHKG
- Naive Bayes and logistic regression suggest symptoms casued by confounding factors
- Noisy OR often suggests general symptoms
- Difficulty inferring causation from correlation
- Confounding factors difficult to recognize and eliminate

7 Future improvements

- Edges between symptoms
- Softer boundary between symptoms and diseases
- Introduce a manual filter step
- Use other, non-parametric models
- Higher coverage, more input data

8 Conclusion

Thank you for attending my presentation!

Was a knowledge graph really necessary? Is this even a knowledge graph?

- [1] M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, und D. Sontag, „Learning a Health Knowledge Graph from Electronic Medical Records“, Sci Rep, Bd. 7, Nr. 1, S. 5994, Juli 2017, doi: 10.1038/s41598-017-05778-z.
- [2] M. Fieschi, „Managing and Integrating Clinical Data: Health Records“, in Health Data Processing, Elsevier, 2018, S. 121–136. doi: 10.1016/B978-1-78548-287-8.50009-2.
- [3] S. Shafqat, S. Kishwer, R. U. Rasool, J. Qadir, T. Amjad, und H. F. Ahmad, „Big data analytics enhanced healthcare systems: a review“, J Supercomput, Bd. 76, Nr. 3, S. 1754–1799, März 2020, doi: 10.1007/s11227-017-2222-4.