# Presentation on: Learning a Health Knowledge Graph from Electronic Medical Records *by* Rotmensch et al.

Ole Berg

November 19, 2024

# Table of Contents

# 1 Metadata

- Authors: Maya Rotmensch, Yoni Halpern, Abdulhakim Tlimat, Steven Horng, David Sontag
- Joint effort of institutions at medical centers and universities
- Contains comprehensive appendix
- Published on July 20, 2017 in *Scientific Reports*
- *Scientific Reports* is multi-disciplinary open access journal [1]
- 5th-most cited journal in the world [1]
- 248 citations on CrossRef (96th percentile)

## 2 Motivation and earlier work

- Demand for decision support systems in clinical settings
- Existing knowledge bases created manually or *"using simple pairwise statistics"* [2, 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 betweens diseases and symptoms in textbooks and journals

# 3 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

# 4 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 [4, p. 4]
- EMR originally tool to store data *"of one or more pathological episodes concerning a patient"* [3, p. 121]
- Newer interpretation: EMR is information on a patient from one healthcare provider [4, p. 4]

# 4 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

# 5 Implementation

## 5.1 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

## 5.2 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

## 5.3 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

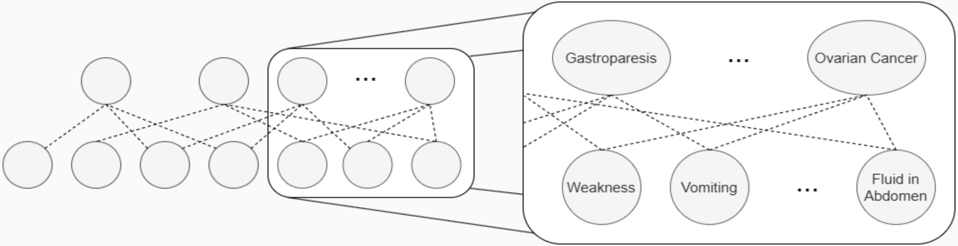


**Figure 1:** Resulting knowledge graph (own illustration according to [2, p. 4])

# 6 Evaluation

## 6.1 Methods

- Automatic evaluation against GHKG
- Assumption: GHKG is precise, but not complete
- Apt for 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 tags as positive
- Precision-recall curve as evaluation measure

## 6.2 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 ($p = 0.01$)

# 7 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 caused by confounding factors
- Noisy OR often suggests general symptoms
- Difficulty inferring causation from correlation
- Confounding factors difficult to recognize and eliminate

# 8 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

# 9 Conclusion

Thank you for attending my presentation!

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

# 9 Sources


- [1] Scientific Reports, „About Scientific Reports | Scientific Reports“. Zugegriffen: 18. November 2024. [Online]. Verfügbar unter: https://www.nature.com/srep/about
- [2] 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.
- [3] 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.
- [4] 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.