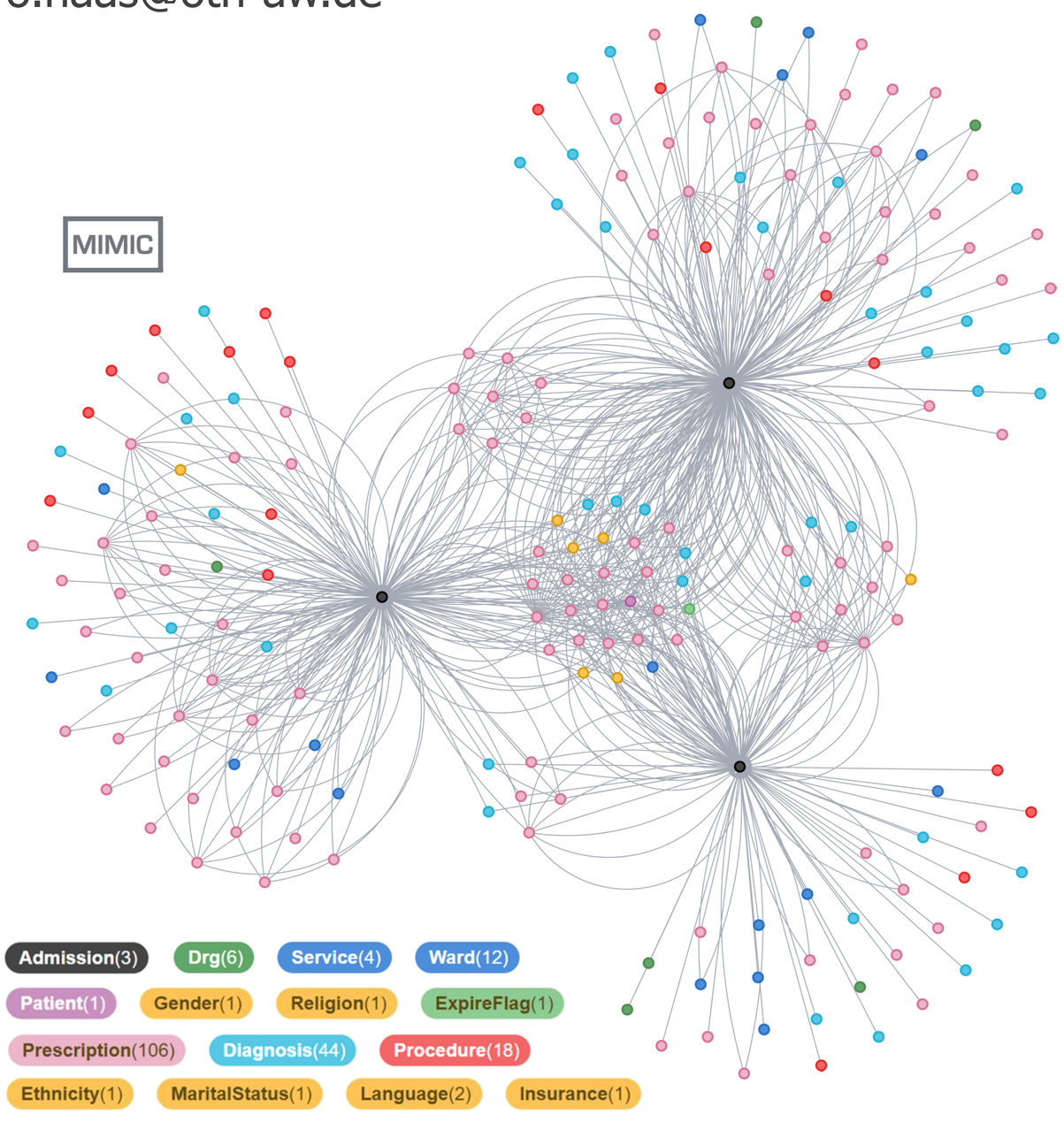
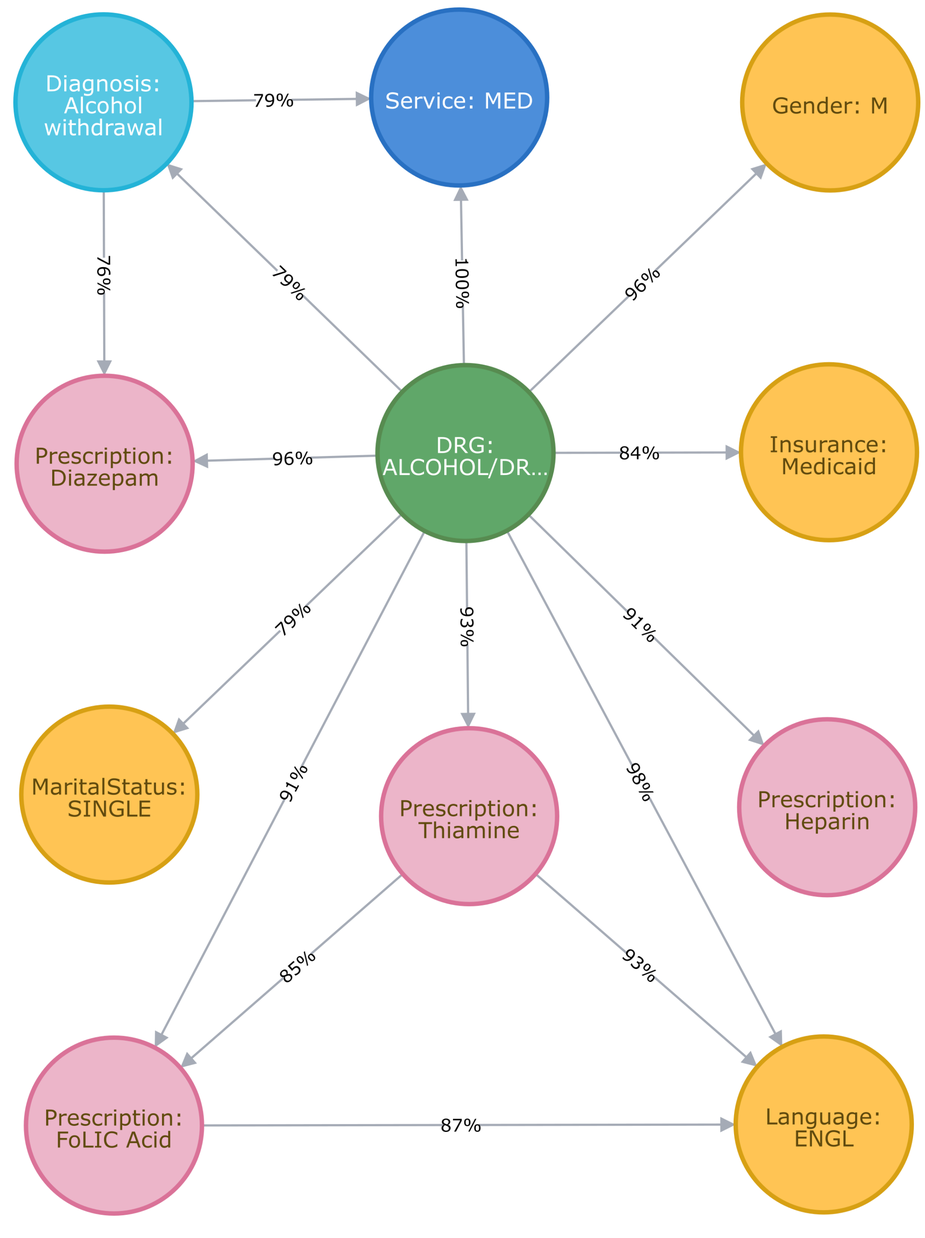
Automated learning of general association rules from large heterogeneous clinical records

Oliver Haas (OTH Amberg-Weiden, FAU Erlangen Nürnberg)
ZD.B Fellow
o.haas@oth-aw.de

Ostbayerische Technische Hochschule
Amberg-Weiden





Left: MIMIC graph, a graph version of the original MIMIC-III dataset. The highly connected black nodes are the admission nodes. All other nodes describe various aspects like diagnoses, procedures, prescriptions and many more. Nodes have been decreased in size to protect patient data. **Right**: Some of the resulting rules. These are all the rules that contain the DRG "ALCOHOL/DRUG ABUSE OR DEPENDENCE, LEFT AMA".

Introduction

Hospitals all over the world have more and more clinical records available which were gathered over the years. This data can be used to improve the treatment of future patients by enabling decision support and may be used to uncover previously unknown connections between different aspects of clinical cases. New approaches for data integration and analysis are needed. We used a publicly available database [1] consisting of clinical records. We transformed it into a graph and applied association rule mining. This creates a more condensed view of the original data, simplifying analysis.

Problem Statement and Research Question

Information from previously documented admissions from different data domains like diagnoses, procedures, prescriptions, demographic data must be integrated. By storing this information in a graph as *data nodes*, different types of data can be stored in a unified way (see left graph in figure). Every admission node is then linked to all other nodes that describe the clinical case. As an example, a male patient with sepsis creates an admission node with links to the 'male' node and to the 'sepsis' node, as well as a patient node that describes the person itself. One patient node can be linked to multiple admission nodes. This format makes it easy to explore *all* data belonging to an admission, irrespective of its source and type, as they are just the neighbors of the admission node. While this approach makes the data easier to analyze, a method which automatically detects patterns in the data is still needed. This helps clinicians and researchers understand the data, which would not be possible without computer support due to its size and heterogeneous formats.

Research Approach and Methodology

Treating admissions as baskets and the neighbor nodes as items in the basket, we can mine association rules in the data, i.e. rules of the form A → B indicating that the existence of item A in an admission likely hints at the coexistence of item B in the same admission. This is backed by statistical quality measures in the original clinical records. These rules can then be used by clinical staff and researchers to discover new relations between nodes of different types and to use the documented evidence for decision-making for new admissions by learning from similar admissions.

Using data from the MIMIC-III project [1], a graph consisting of a large part of the data has been created, making the database more approachable for the analysis of single patients and the contents of their admissions. Then, after applying association rule mining to the admissions as transactions and their neighbors as items, we created a new graph, called the *rules graph* (see right graph in the figure), by linking the items using the discovered one-to-one rules, i.e. a rule A → B becomes a link from the node A to the node B. In the figure, nodes are shown as circles with the type and the content written inside the circle. Arrows of the form A → B are rules. The confidence of the rule, i.e. the certainty that if A is connected to an admission then B is connected to the same admission as well, written on the arrow. These links between items can then be analyzed in an interactive way.

Related Work

In comparison to other work, e.g. [2], the presented approach is not restricted to certain domains: all information can be used, and rules can include any type of data and are not restricted to diagnosis → diagnosis rules. Additionally, our approach can be applied to general hospital settings - not only to intensive care data, as in other work [3]. The presented method can also produce all rules and store them in a database. This enables researchers to look at clusters of rules without the need to re-learn rules over and over again.

Preliminary Results

The learned rules could partially be validated using scientific literature. As an example, our approach automatically generated the rule that the sickle-cell trait is mostly found in the African-American population. This finding is consistent with previous research [4]. Other rules have been found that need scientific validation, mostly on the impact of social factors on the prevalence of various diseases, a field for which scientific evidence is hard to find.

Planned Next Steps

Two next steps are planned. First, we will include the temporal information to additionally discover rules of the form "two months after diagnosis A, diagnosis B occurs in 80% of the cases". Secondly, we plan to implement the proposed method in a German hospital for a usability evaluation as well as a study on the transferability of rules between hospitals, potentially even between countries.

Open Issues and Cooperation

Clinical cooperation is very welcome, as more hospitals mean more insight into differences between learned rules. Another interesting aspect would be to specifically analyze rules around certain diseases like Alzheimer's, cancer, or infectious diseases.

References

[1] Johnson, AE; Pollard, TJ; Shen, L; Lehman, L-W; Feng, M; Ghassemi, M; Moody, B; Szolovits, P; Celi, LA; Mark, RG. MIMIC-III, a freely accessible critical care database. Sci Data. 2016 May 24;3:160035.

[2] Kang'ethe, S & Wagacha, P. Extracting Diagnosis Patterns in Electronic Medical Records using Association Rule Mining. International Journal of Computer Applications. 2014; 108:19-26.

[3] Cheng CW; Chanani N, Venugopalan J, Maher K; Wang MD. icuARM-An ICU Clinical Decision Support System Using Association Rule Mining. IEEE J Transl Eng Health Med. 2013 Nov 21;1(1):122-31.

[4] Brewster B. Sickle cell anaemia: causes, signs, symptoms and treatment. Nurs Times. 2003 Jul 22-28;99(29):30-2.