
Conference or Workshop Item:
Sykes, SJ, Kingsbury, SR, Conaghan, PG et al. (Accepted: 2017) A Process Mining Approach to Discovering Cardiovascular Disease Trajectories. In: Medical Informatics Europe (MIE) 2018, 24-26 Apr 2018, Gothenburg. (Unpublished)

This is an unpublished poster presented at the Medical Informatics Europe (MIE) conference 2018.

A Process Mining Approach to Discovering Cardiovascular Disease Trajectories

S.J. Sykes [a], S.R. Kingsbury [a,b], P.G. Conaghan [a,b], M. Pujades Rodriquez [c], P.D. Baxter [d] and O.A. Johnson [e]

[a] Leeds Inst of Rheum & Musculoskeletal Medicine (LIRMM), University of Leeds, UK; [b] NIHR Leeds Biomedical Research Centre, Leeds, UK; [c] Leeds Inst of Biomedical & Clinical Sciences, University of Leeds, UK; [d] Leeds Inst of Cardiovascular & Metabolic Medicine, University of Leeds, UK; [e] School of Computing, University of Leeds, UK

Contact email: umsjs@leeds.ac.uk

Introduction

Is there an important relationship between gout and CVD? The answer is yes, according to a paper in Nature Communications, 2014 by Jensen et al. [1]. Medical researchers are interested in the way that diseases progress over time and EHRs have the potential to reveal insights into the biomechanics of disease progression or trajectories. [1] defined a disease trajectory as a *set of sequential disease associations* derived using statistical approaches. Other methods can also be used to build models from temporal data and we present our attempts to reproduce CVD trajectories using a different dataset and a variety of methods including process mining [2]. We created models from EHR data contained in a critical care database from a hospital in Boston and used a rule-based approach carefully replicating [1] to extract and transform diagnosis data, then examined the resulting trajectories using process mining software tools [3].

Method

The method developed by Jensen et al. was applied to the Danish National Patient Registry covering the population of Denmark (n=6.2 million patients) over a 14.9 year period [1]. Pairs of diagnoses where one diagnosis precedes another in the patient's record were extracted. Binomial tests were used to test directionality for statistical significance before combining the pairs into disease trajectories. Here we use MIMIC-III [4], a database containing 46,520 patients EHRs from an intensive care hospital in Boston, USA. The data covers 11 years of hospital and patient information. Extraction, transformation and load steps (figure 1) were applied to create the disease trajectories. The MIMIC-III version 1.4 dataset was installed following the procedures described in [4]. Timestamped diagnosis records for CVD patients were extracted to create a raw event log file.

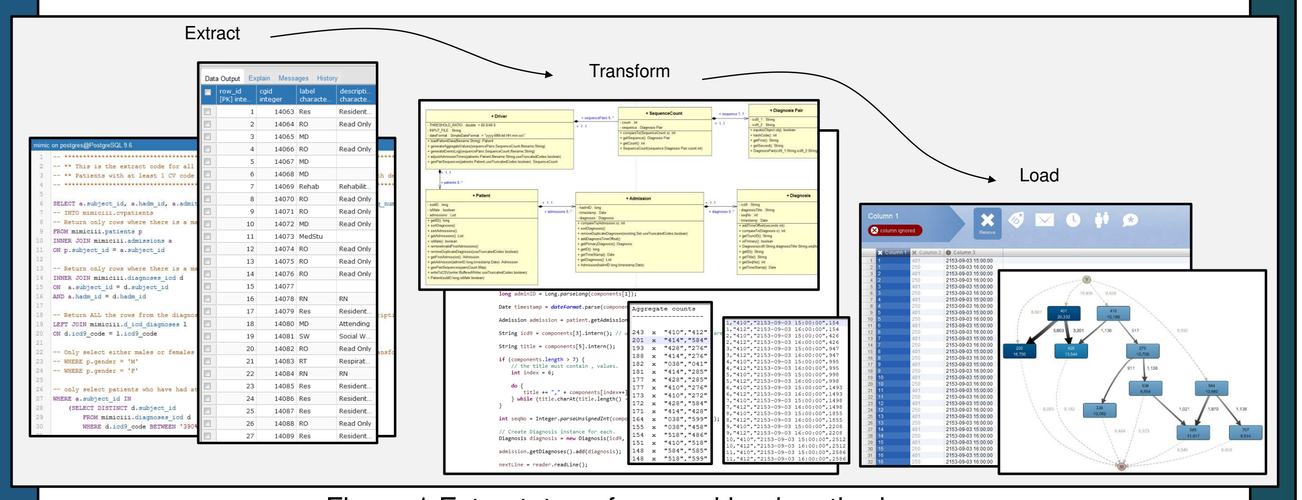

Figure 1 Extract, transform and load method.

This was transformed by using Admission time for the primary diagnosis and a synthetic timestamp was allocated to secondary diagnoses. Rather than extracting these as diagnosis pairs for statistical analysis following [1] we created an equivalent synthetic event log suitable for input into process mining tools. Counts of directional pair occurrences were taken, and an output log created containing only directional pairs with a ratio greater than 60:40. The process mining tool Disco [3] was used to model the disease trajectories, with results validated by a domain expert.

Results

Two trajectory models were produced with the first (figure 2) following closely to Jensen's rules. Results for the first model (figure 3) showed eight diagnoses in common with one pair having the same directly follows relationship and another pair having a follows relationship. Our second model, produced by refining the method by eliminating repeat diagnosis codes for a patient has eight diagnoses in common with five similar relationships.

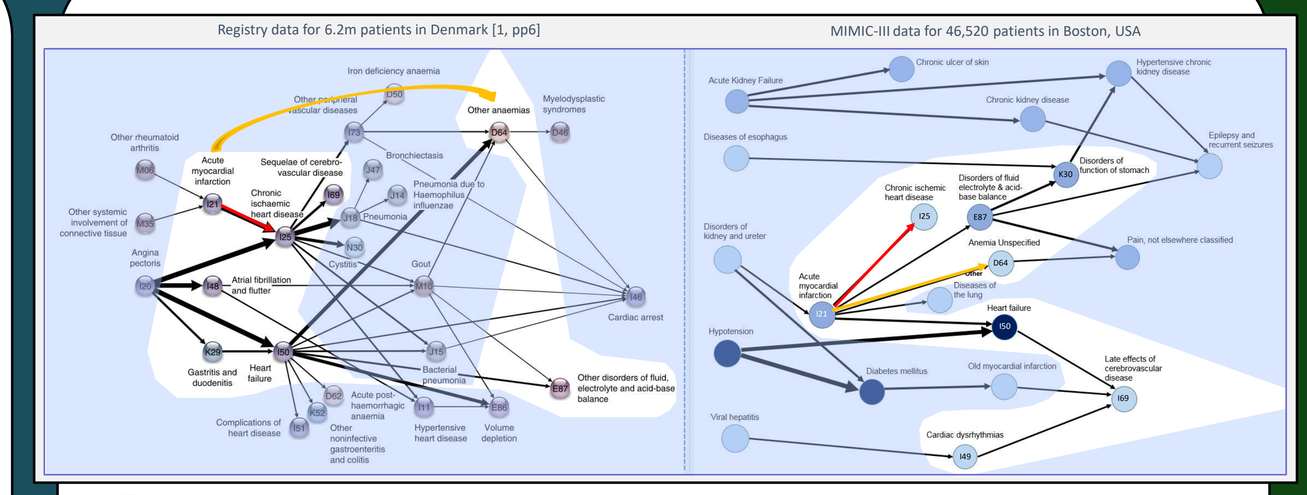

Figure 2 Comparison between results for the Jensen method and our trajectory 1 where Jensen's rules were closely followed.

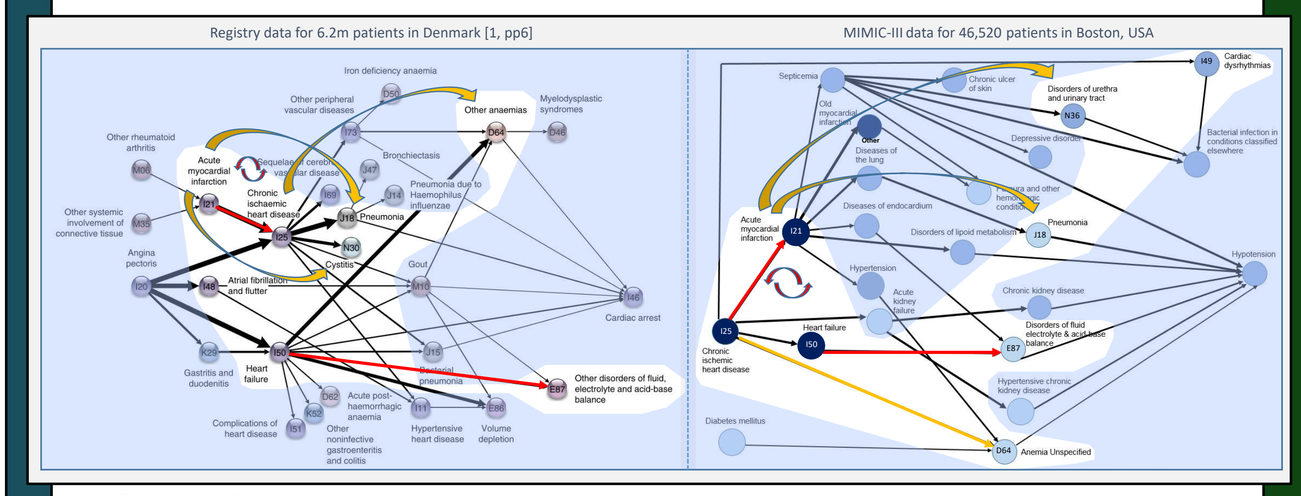

Figure 3 Comparison between results for the Jensen method and our trajectory 2 where repeat diagnosis codes were excluded.

Discussion

Our refined model excludes diagnosis codes occurring as comorbidities in the first admission and repeat diagnosis codes. This assumes that the first time a disease is seen is the first time it appears for the patient relying on all previous diagnosis codes for that patient to be recorded at the first admission. The process of excluding repeat diagnosis codes could be further improved by classifying them as chronic or acute. Only chronic diagnosis codes would be excluded from the event log, as it is likely that acute diseases would be new occurrences. Gout becomes visible in our trajectory models when the percentage of diagnosis codes displayed in Disco is increased to ~11.9%. Temporal directional pairs of diagnosis codes may be described as bigrams [5] which in computational linguistics is a sequence of two adjacent elements. The use of n-grams to create disease trajectories may be an alternative method to explore.

Conclusions

We were able to reproduce the method for modelling disease trajectories used by [1] and found gout to be, though not central, an important disease. Refinements were made to the method and new disease trajectories created. We believe the method for data mining and mapping of disease trajectories using process mining tools to be of interest to health data researchers, as it enables a more rapid modeling made possible by a simple transformation of diagnosis pairs into an event log. This work highlighted many of the complexities involved in creating disease trajectories from EHRs but demonstrated such approaches are feasible. Further enhancements by extending the methods used with n-grams and computational linguistics may be rewarding.

References

[1] A.B. Jensen, et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nat. Commun.* **5** (2014), 4022.
[2] W. van der Aalst, et al. Process mining manifesto. *Lect. Notes Bus. Inf. Proce.* **99** (2012), 169–194.
[3] C.W. Günther, A. Rozinat. Disco: Discover Your Processes. *BPM 2012;* 2012 Sep 3; Tallinn, Estonia.
[4] A.E.W Johnson, et al. MIMIC-III, a freely accessible critical care database, *Sci. data* **3** (2016), 160035.
[5] B.J. Marafino, et al, N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. *JAMIA* **21** (2014), 871–875.