Design of a framework to detect temporal clinical event trajectories from health data standardized to the OMOP CDM

Kadri Künnapuu, Solomon Ioannou, Kadri Ligi, Raivo Kolde, Sven Laur, Jaak Vilo, Peter Rijnbeek, Sulev Reisberg

Background

Identification of temporal disease sequences (trajectories) in electronic health records (EHR) could not only characterize the particular dataset but also describe progressions within the population and potentially predict future illness from the existing ones[1,2,3]. However, the number of disease trajectory studies has remained relatively small. We believe it is mostly because of two reasons - first, there is a lack of syntactic and semantic interoperability of health data which makes network studies a challenge, and second, there has not been an open-source analytical framework implementation for performing this type of analysis. While the first issue has been effectively tackled by the OHDSI community in recent years by developing the OMOP Common Data Model (CDM), and more databases are becoming available supporting this format, common principles for disease trajectory studies are also needed. We have found that the methods used in previous publications are described insufficiently for adequate replication in other datasets, making it almost impossible to verify the results or conduct a similar analysis in other settings.

In this submission, we propose a standardized framework for detecting the most prominent temporal clinical event trajectories in the observational health dataset based on the best practices of that field. We extend the previously published methods by adding new configuration options for these kinds of studies. We also introduce the implementation of the framework as open-source software that utilizes the OMOP CDM and standardized vocabularies.

Methods

The proposed framework for detecting temporal health event trajectories consists of the following steps:

1. Define a study cohort by using OHDSI tools
2. Specify study parameters (which type of events are included; additional requirements for the trajectories)
3. Identify temporal clinical event pairs by extensive statistical testing of all two-event-sequences in OMOP CDM data.
4. Build trajectory graphs from significant directional clinical event pairs
5. Align actual event sequences to the graph to identify longer trajectories

The framework described above is implemented as an open-source R package. We test the framework and package on EHR data from Estonia and the Netherlands and compare the results with previous findings in the Danish population.

Results

We ran the package on 10% of a random sample of Estonian health records (n=147K patients, 8 years)

and identified 22 directional event pairs having relative risk (RR) >2 and occurring on at least 5% of Type 2 Diabetes patients. Out of these, 5 passed the validation in Netherlands' data (IPCI database, n=2.5M) (Figure 1). Concept ID-s used in 14 pairs are not used in IPCI.

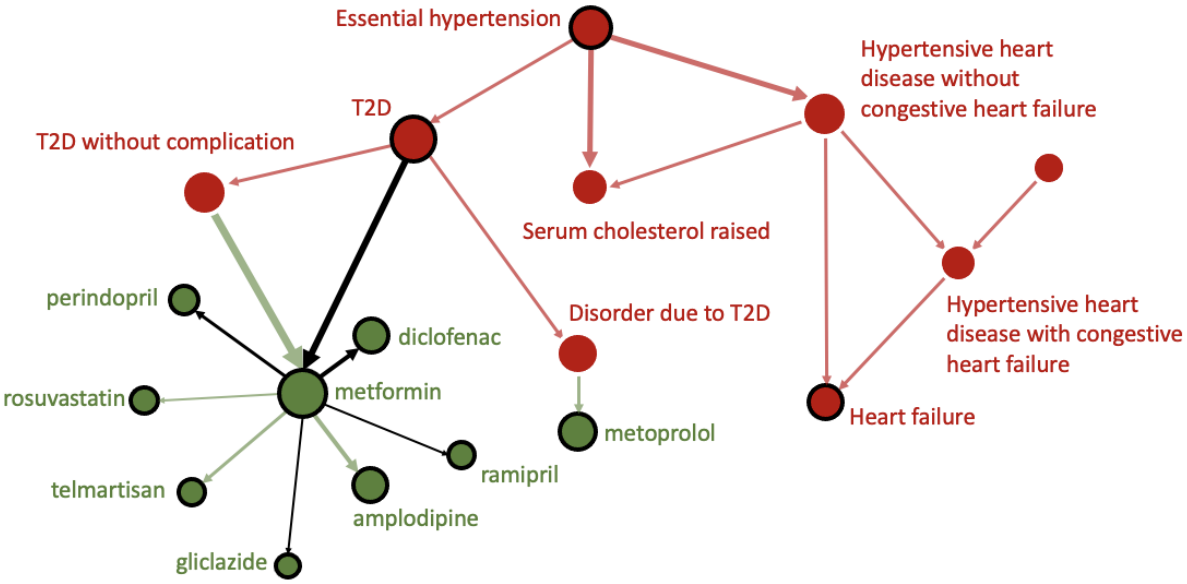

Figure 1. 20 most prevalent event sequences among Type 2 diabetes mellitus (T2D) patients having relative risk (RR) >2 in Estonian electronic health records. Five event pairs that passed validation in the IPCI database (Netherlands) are shown with black arrows. Events (Concept ID-s) with white borders are not used in IPCI.

We also validated 7733 temporal event pairs from the Danish population[2] (n=7M patients, 25 years) having prevalence >1:10000 and RR<0.8 or RR>1.2. Despite the fact that the Estonian dataset is 49x smaller in the patient count and 3x in the time range, we were able to confirm significantly altered RR and direction of 781 pairs (10.0%) (Figure 2).

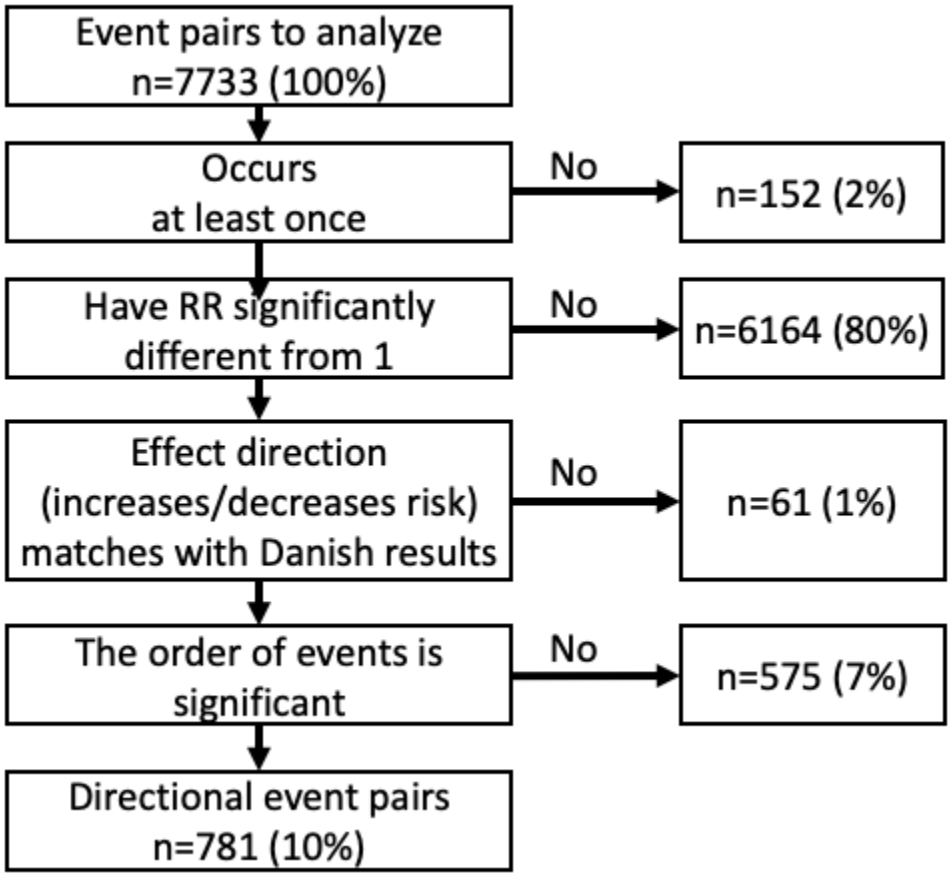

Figure 2. Attrition diagram, showing the number of event pairs after various stages in the validation analysis of the Danish results.

Conclusion

The proposed framework identifies and visualizes significant clinical event progression patterns in health data standardized to the OMOP CDM. The open-access R package, the first of its kind, allows researchers to run the same framework on their OMOP-formatted health data and compare results across databases to allow for the identification of clinical event associations. The package will be freely available on GitHub after the publication of the manuscript.

We see from the results that using different Concept ID-s for the same underlying event in different OMOP databases makes the cross-dataset comparison of event trajectories challenging. Before moving to investigate longer global trajectories, a global consensus on the simplest trajectories - pairs - need to be established first.

Funding

This work was supported by the Estonian Research Council grants (PRG1095, RITA1/02-96-11); by the European Union through the European Regional Development Fund grant EU48684; by European Social Fund via IT Academy programme. The European Health Data & Evidence Network has received funding

from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

References/Citations

1. Jensen AB, Moseley PL, Oprea TI, Ellesøe SG, Eriksson R, Schmock H, et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nature Communications. 2014 Jun 24;5(1):4022.
2. Siggaard T, Reguant R, Jørgensen IF, Haue AD, Lademann M, Aguayo-Orozco A, et al. Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7.2 million Danish patients. Nature Communications. 2020 Oct 2;11(1):4952.
3. Hu JX, Helleberg M, Jensen AB, Brunak S, Lundgren J. A Large-Cohort, Longitudinal Study Determines Precancer Disease Routes across Different Cancer Types. Cancer Res. 2019 Feb 15;79(4):864–72.