Patient Journey: Temporal Disease Pattern Identification Using Machine Learning and Data Mining

Yahui Tian[1], Nan Shao[1], Zheng Zhu[1], Ziwei Yin[1], Linwei Li[1] │ [1]Boehringer Ingelheim

Abstract:

Patient journey is a collection of topics including understanding patient experience in the healthcare ecosystem in order to improve patient care, as well as understanding the disease development paths and their relation to clinical outcomes to potentially facilitate clinical decision making. Electronic health records (EHR) and health insurance administrative databases have provided rich and inexpensive sources of information for patient journey research. We have looked into temporal disease pattern identification in such databases for a rare disease to improve disease understanding, and we have applied sequential pattern mining and network analysis in the exploratory phase. Many results are in alignment with clinical understanding of the disease and we hope it can also facilitate hypothesis generation for further research. We will discuss methods and limitations and share experience for analyzing such large-scale health databases.

BACKGROUND & STUDY OBJECTIVES

- Machine learning and Data mining in healthcare
  - Advantage: discover unknown features through data driven process
  - Applications: natural disease history, comorbidity, cohort identification, risk prediction, biomarker discovery, etc.
- Topics in patient journey research includes
  - Understanding patient experience in the healthcare ecosystem in order to improve patient care
  - **Understanding the disease development paths and their relation to clinical outcomes to potentially facilitate clinical decision making**
- Potential rich and inexpensive source of information for such research:
  - Electronic health records (EHR)
  - Health insurance administrative databases
  - Characteristics
    - ➢ Primary goal: document patients' care; reimbursement
    - ➢ Data format: structured, semi-structured and unstructured
    - ➢ Advantages: large sample size, long follow-up, cost-effective and time-saving source of research
    - ➢ Challenges: censoring, irregular time series data, completeness, correctness and confounding effects
- **Objective**: identify temporal disease pattern leading to the initial diagnosis of a rare disease using claims and EHR data to improve disease understanding.

METHODS

Workflow:

- Initial disease cohort selection
  - Based on disease definition and corresponding ICD codes, patient age and study period
- Analysis cohort
  - Further selection of cohort based on length of information coverage in the database, e.g. length of continuous enrollment in health insurance program
  - A trade-off between length of history and # of patients included in the analysis
- Data pre-processing
  - Remove codes which are too general or not relevant to the disease of interest (e.g. codes correspond to general office visit)
- Apply two methods
  - Sequential pattern mining (Figure 1)
  - Network analysis (Figure 2)

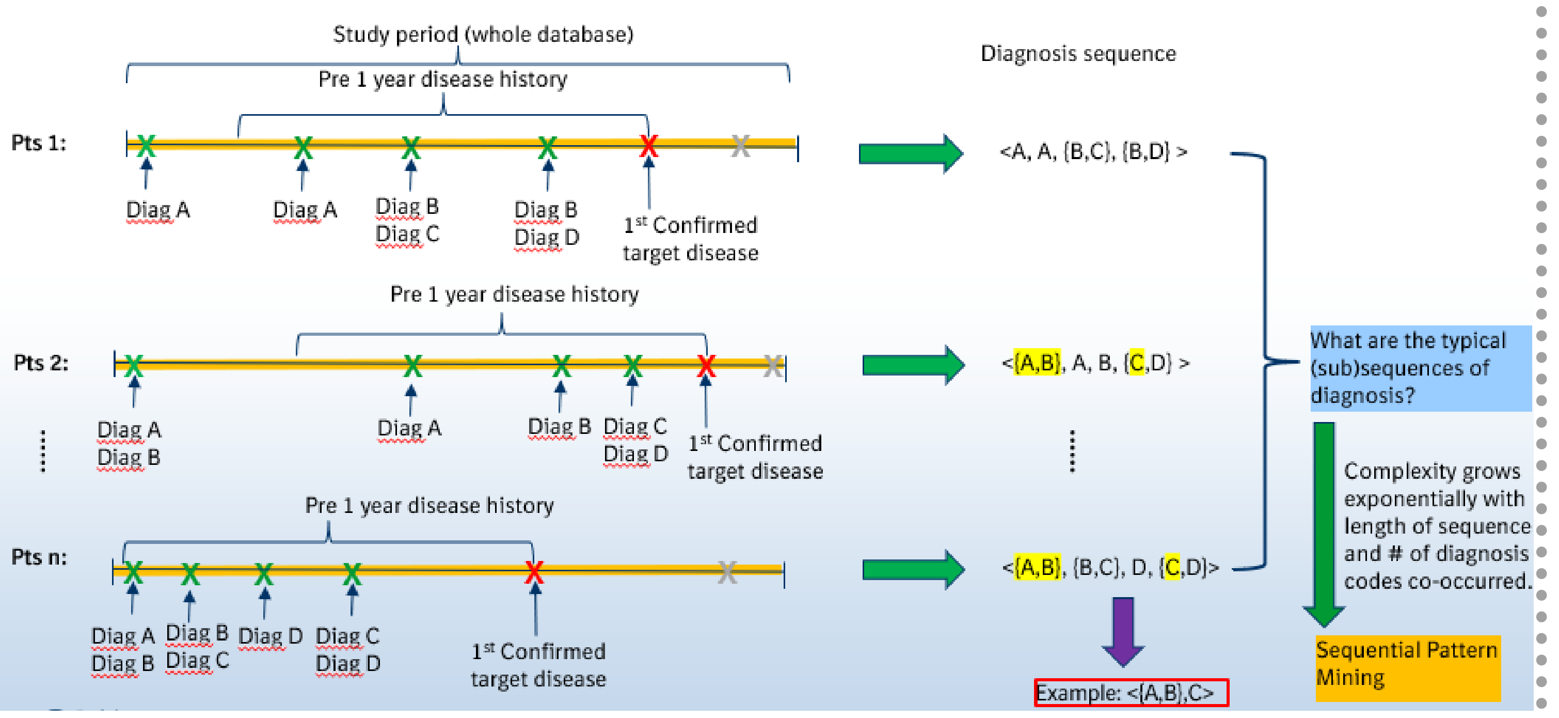
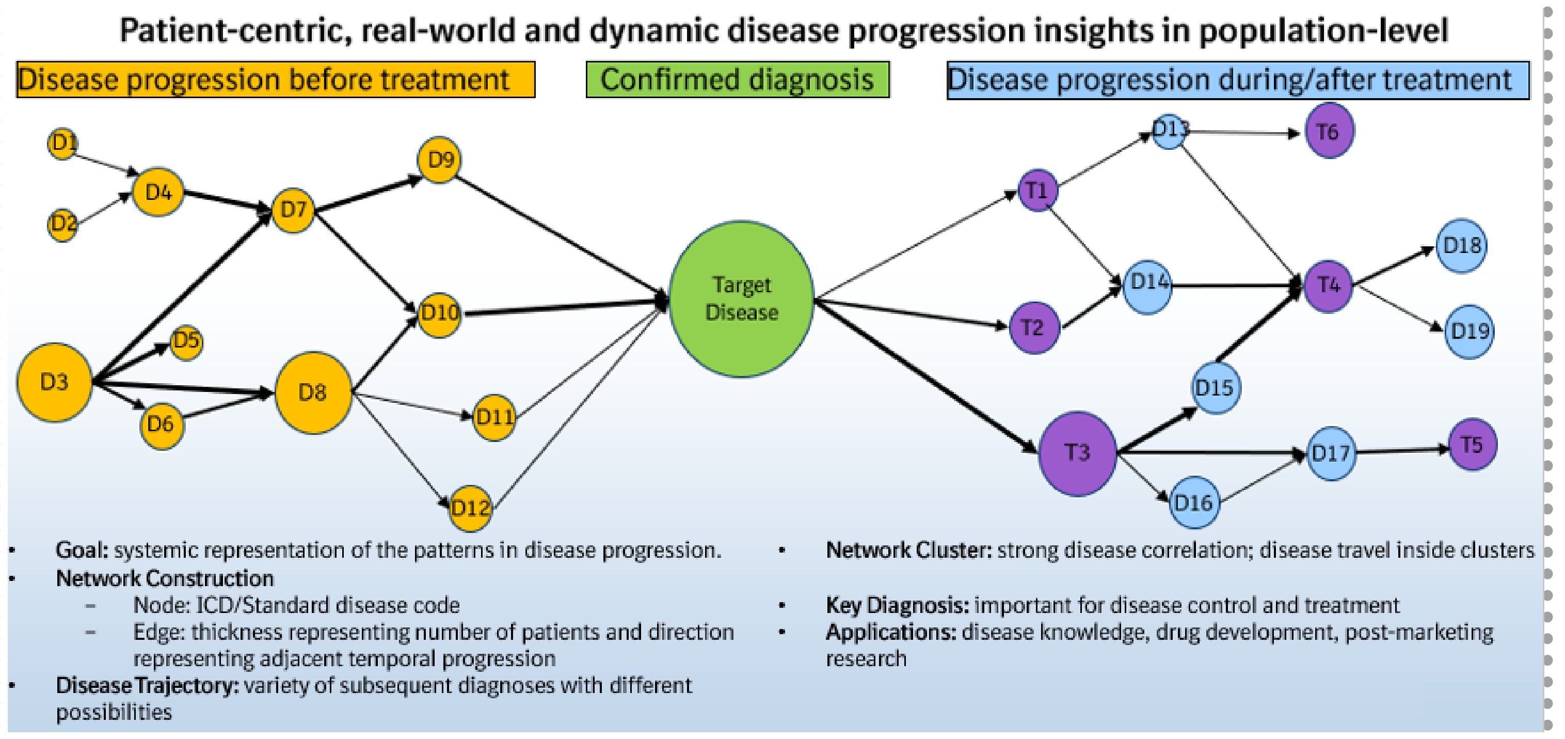

Study period (whole database)
Pre 1 year disease history
Diagnosis sequence
Pts 1:
Diag A
Diag A
Diag B
Diag C
Diag B
Diag D
1st Confirmed target disease
<A, A, {B,C}, {B,D} >
Pre 1 year disease history
Pts 2:
Diag A
Diag B
Diag A
Diag B
Diag C
Diag D
1st Confirmed target disease
<{A,B}, A, B, {C,D} >
Pre 1 year disease history
Pts n:
Diag A
Diag B
Diag B
Diag C
Diag D
Diag C
Diag D
1st Confirmed target disease
<{A,B}, {B,C}, D, {C,D}>
What are the typical (sub)sequences of diagnosis?
Complexity grows exponentially with length of sequence and # of diagnosis codes co-occurred.
Sequential Pattern Mining
Example: <{A,B},C>

Figure 1: Illustration of Sequential Pattern Mining


Patient-centric, real-world and dynamic disease progression insights in population-level
Disease progression before treatment
Confirmed diagnosis
Disease progression during/after treatment
D1
D2
D3
D4
D5
D6
D7
D8
D9
D10
D11
D12
Target Disease
T1
T2
T3
T4
T5
T6
D13
D14
D15
D16
D17
D18
D19

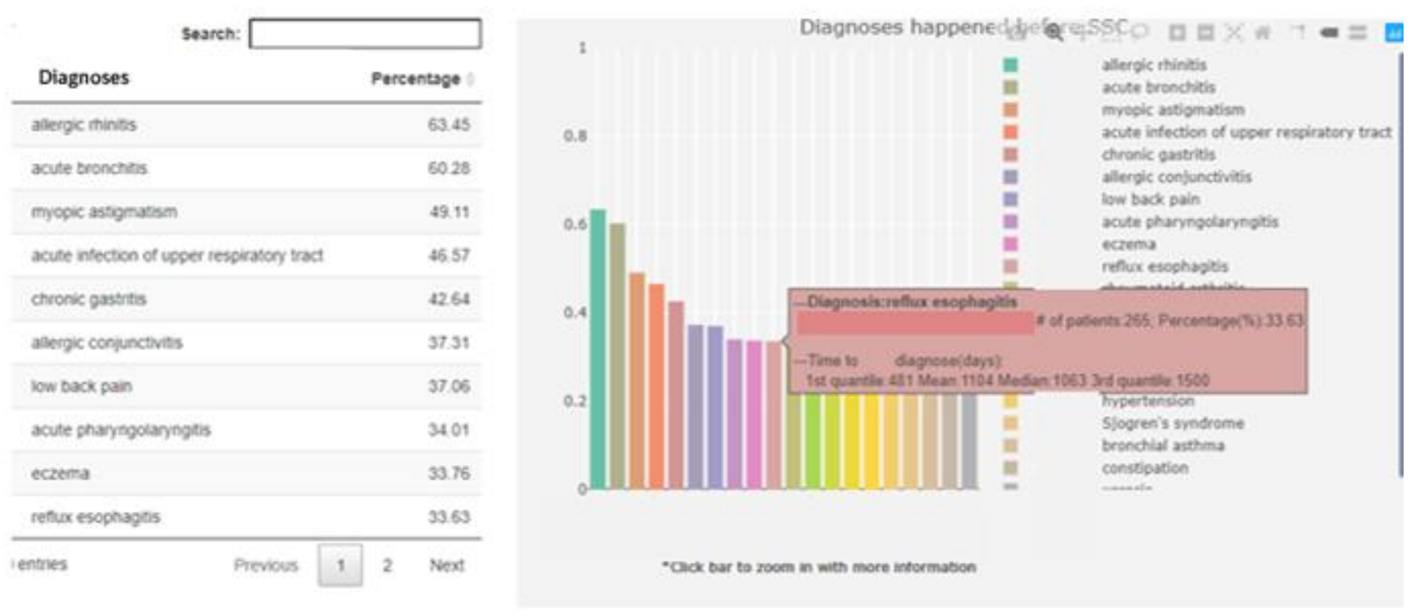
- **Goal:** systemic representation of the patterns in disease progression.
- **Network Construction**
  - Node: ICD/Standard disease code
  - Edge: thickness representing number of patients and direction representing adjacent temporal progression
- **Disease Trajectory:** variety of subsequent diagnoses with different possibilities
- **Network Cluster:** strong disease correlation; disease travel inside clusters
- **Key Diagnosis:** important for disease control and treatment
- **Applications:** disease knowledge, drug development, post-marketing research

Figure 2: Illustration of Network Analysis

RESULTS

Sequential Pattern Mining

- Representative diagnoses before target disease



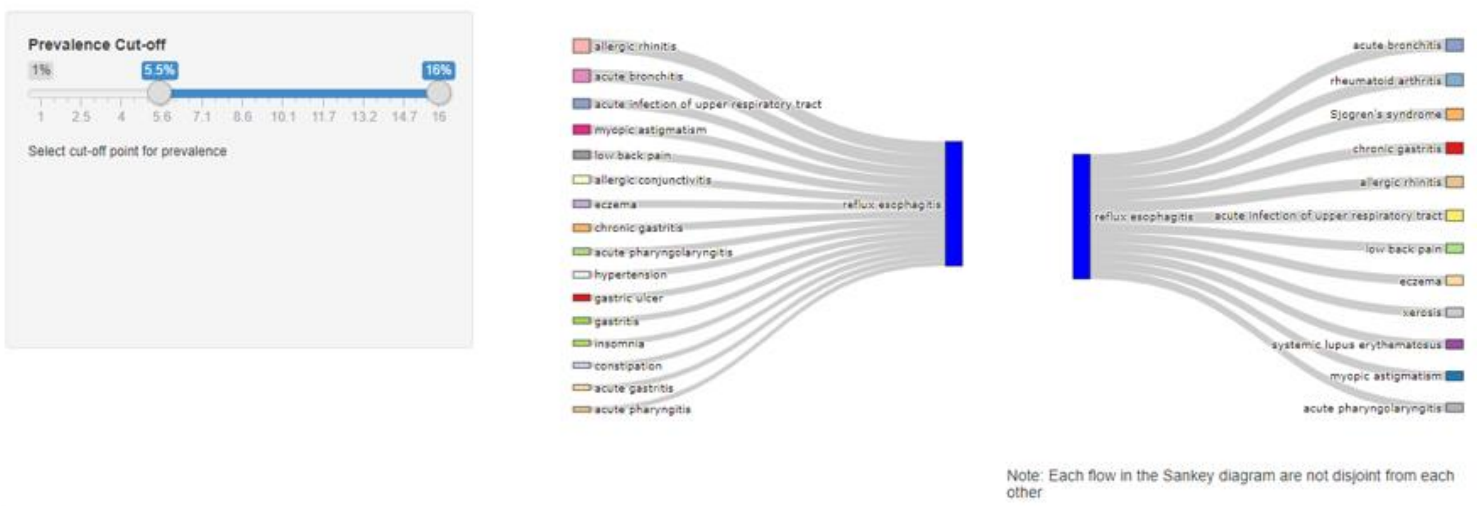
Diagnoses	Percentage
allergic rhinitis	63.45
acute bronchitis	60.28
myopic astigmatism	49.11
acute infection of upper respiratory tract	46.57
chronic gastritis	42.64
allergic conjunctivitis	37.31
low back pain	37.06
acute pharyngolaryngitis	34.01
eczema	33.76
reflux esophagitis	33.63

Diagnosis:reflux esophagitis
# of patients:265; Percentage(%):33.63
Time to diagnose(days):
1st quantile:481 Mean:1104 Median:1063 3rd quantile:1500

*Click bar to zoom in with more information

Figure 3: Interactive digital platform of representative diagnoses. Table on the left shows percentage. Histogram on the right shows detailed information same as Figure 3.

- Zoom in Reflux Esophagitis



Note: Each flow in the Sankey diagram are not disjoint from each other

Figure 4: Interactive digital platform of diagnosis patterns. Adjustment plate on the left enables change of prevalence cut-off to show more or less patterns. Sankey diagram on the right shows diagnoses which happen before and after reflux esophagitis.

- Further Zoom in allergic conjunctivitis to reflux esophagitis

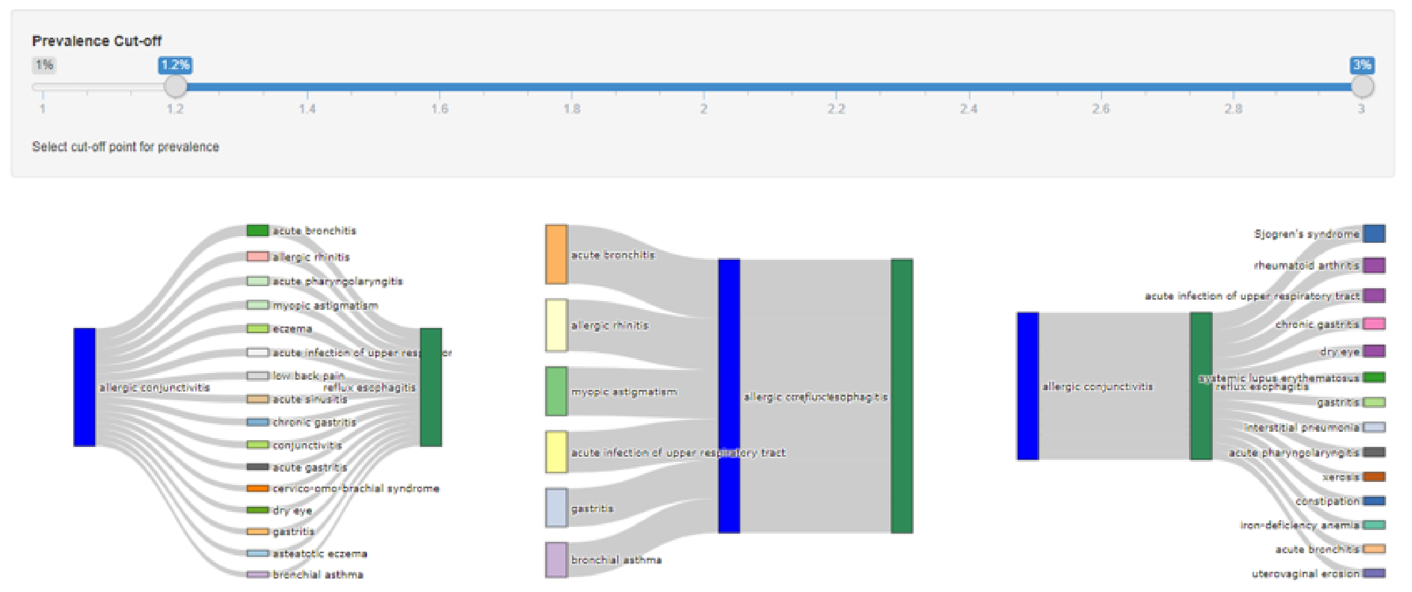
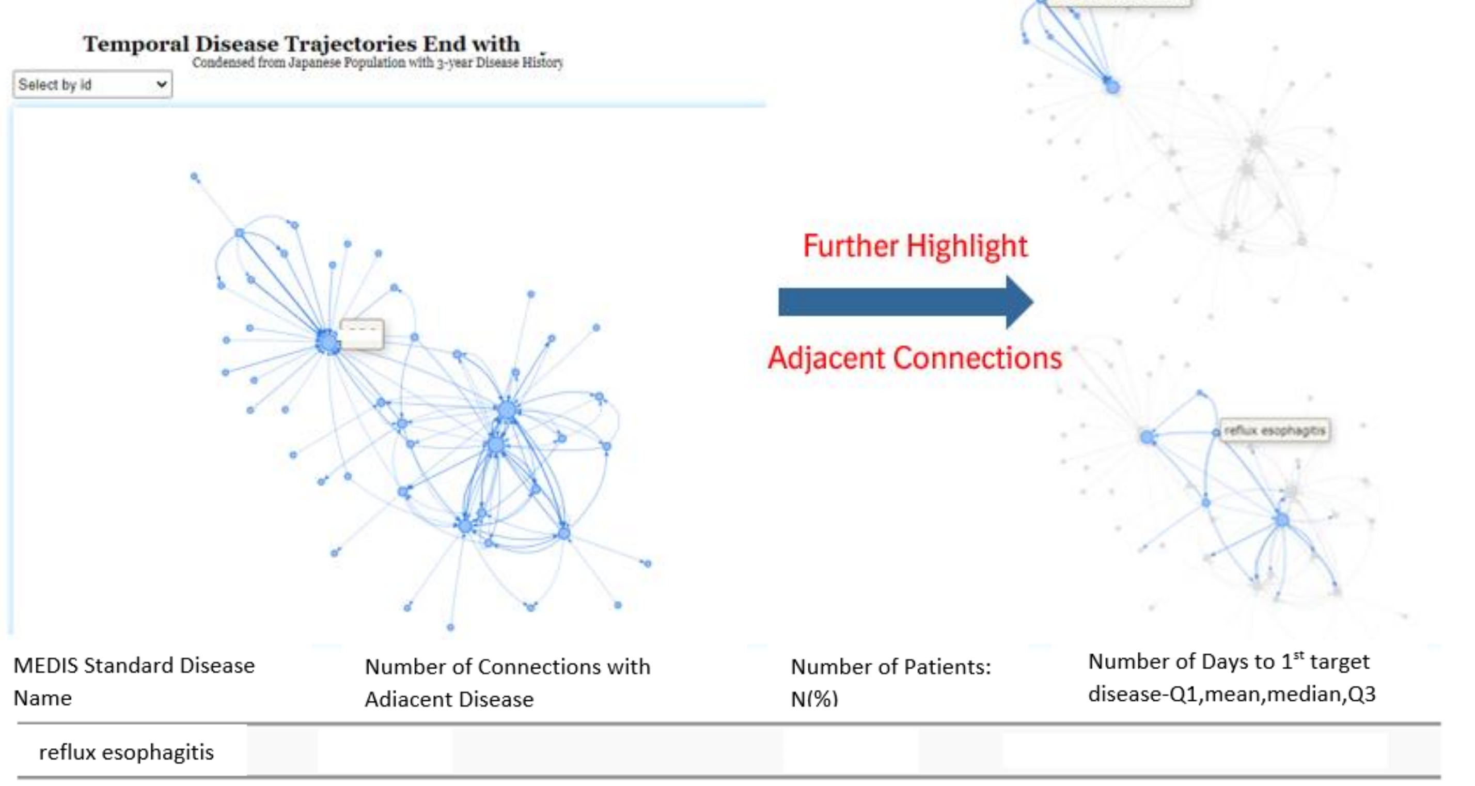


Figure 5: Further zoom in includes more details but also fewer patients. Adjustment plate on the top enables change of prevalence cut-off to show more or less patterns. Sankey diagram on the bottom shows diagnoses which happen before, in-between, and after allergic conjunctivitis and reflux esophagitis.

Network Analysis

- Diagnoses network before target disease


Temporal Disease Trajectories End with
Condensed from Japanese Population with 3-year Disease History
Select by id
Further Highlight
Adjacent Connections
rheumatoid arthritis
reflux esophagitis

MEDIS Standard Disease Name	Number of Connections with Adjacent Disease	Number of Patients: N(%)	Number of Days to 1st target disease-Q1,mean,median,Q3
reflux esophagitis			

Figure 6: Interactive digital platform of diagnosis network. Table on the bottom shows detailed information of selected diagnosis, including standard disease name, number of neighbor diagnoses, prevalence and progression time needed from selected diagnosis to target disease.

DISCUSSIONS

- The effort is exploratory
  - The results largely confirm existing clinical knowledge of the disease, e.g. common conditions leading to disease diagnosis
  - Identify patterns which may not be well understood and facilitate hypothesis generation for further research
- Data limitation needs to be well understood and communicated
  - Not all conditions are captured in the ICD codes
  - Claims data may capture all encounters during enrollment period, but EHR (especially single EHR provider data) may not
  - Lab results are usually not captured in claims data but can be found in EHR
  - Validation across different databases or different time periods may be desired

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Fei Wang from Cornell University, Dr. Thomas Leonard (retired) and Dr. Donald Zoz from Boehringer Ingelheim for their support and discussions during the project.

REFERENCES

1. Perer, Adam, and Fei Wang. "Frequence: Interactive mining and visualization of temporal frequent event sequences." Proceedings of the 19th international conference on Intelligent User Interfaces. 2014.
2. Perer, Adam, Fei Wang, and Jianying Hu. "Mining and exploring care pathways from electronic medical records with visual analytics." Journal of biomedical informatics 56 (2015): 369-378.
3. Beck, Mette K., et al. "Diagnosis trajectories of prior multi-morbidity predict sepsis mortality." Scientific reports 6.1 (2016): 1-9.

Presented at: The ASA Biopharmaceutical Section Regulatory-Industry Statistics Workshop, September 24, 2021