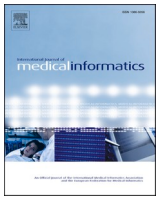
International Journal of Medical Informatics 175 (2023) 105068

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Review article

Predicting onset of disease progression using temporal disease occurrence networks

G.I. Choudhary *, P. Fränti

School of Computing, University of Eastern Finland

ARTICLE INFO

Keywords:
Chronic Disease
Data Mining
Disease Progression Network
Disease Future Risk Prediction
Health Informatics
Network Theory

ABSTRACT

Objective: Early recognition and prevention are crucial for reducing the risk of disease progression. This study aimed to develop a novel technique based on a *temporal disease occurrence network* to analyze and predict disease progression.

Methods: This study used a total of 3.9 million patient records. Patient health records were transformed into temporal disease occurrence networks, and a *supervised depth first search* was used to find *frequent disease sequences* to predict the onset of disease progression. The diseases represented nodes in the network and paths between nodes represented edges that co-occurred in a patient cohort with temporal order. The node and edge level attributes contained *meta*-information about patients' gender, age group, and identity as labels where the disease occurred. The node and edge level attributes guided the depth first search to identify *frequent disease occurrences* in specific genders and age groups. The patient history was used to match the most frequent disease occurrences and then the obtained sequences were merged together to generate a ranked list of diseases with their *conditional probability* and *relative risk*.

Results: The study found that the proposed method had improved performance compared to other methods. Specifically, when predicting a single disease, the method achieved an *area under the receiver operating characteristic curve* (AUC) of 0.65 and an F1-score of 0.11. When predicting a set of diseases relative to ground truth, the method achieved an AUC of 0.68 and an F1-score of 0.13.

Conclusion: The ranked list generated by the proposed method, which includes the probability of occurrence and relative risk score, can provide physicians with valuable information about the sequential development of diseases in patients. This information can help physicians to take preventive measures in a timely manner, based on the best available information.

1. Introduction and background

The *International Classification of Diseases*, tenth revision (ICD-10) is a coding system used to record patient diagnoses, maintain procedure histories, and facilitate cost reimbursement [1,2]. The *World Health Organization* (WHO) recognizes these codes,[1] and they are used as a standard across the globe. In recent years, researchers have used codes to model disease co-occurrence and progression for disease prediction purposes [3–10]. Davis *et al.* [11–13] introduced the concept of collaborative filtering for disease prediction using patient histories based on ICD-9-CM codes. The recommendation engine relies on the behavior of similar patients to produce a recommendation for a given patient. Steinhaeuser and Chawal [14] constructed disease networks and studied their structural properties to better understand disease relationships and behavior over time. They trained a generalized predictive model that takes patient history as input, extracts patient networks based on the concept of nearest neighbors, and generates a ranked list of medical conditions.

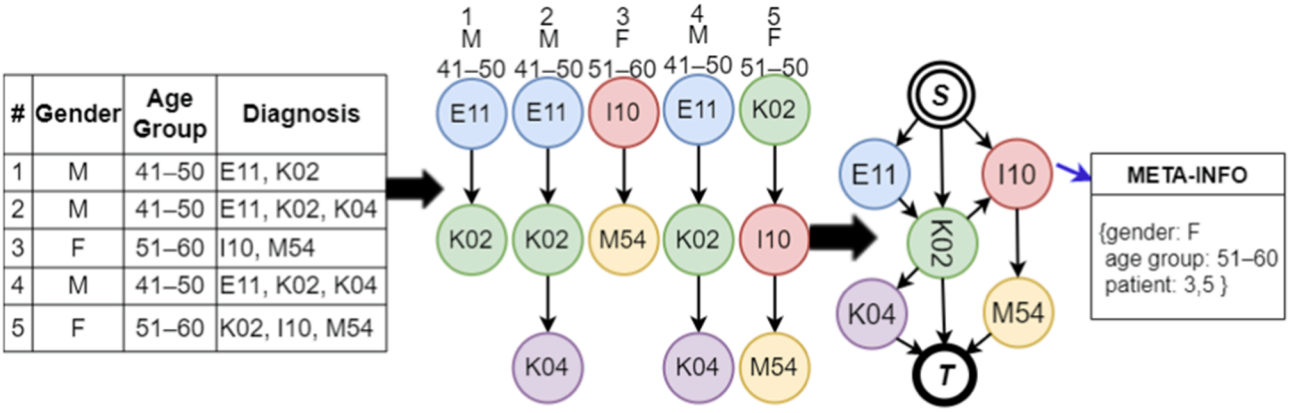
Jensen *et al.* [15] used a sequential approach to identify pairs of diagnoses with temporal directions that are statistically significant. These pairs were combined to form longer disease trajectories. Folino and Pizzuti [16,17] proposed a recommendation engine using association rules mining with and without the Markov model. The results presented in the study showed that combining the association rule with the Markov model improved prediction performance.

Another study developed a phenotypic comorbidity network to

* Corresponding author.
E-mail addresses: gulraiz@cs.uef.fi, ch.gulraiz@gmail.com (G.I. Choudhary), franti@cs.uef.fi (P. Fränti).

[1] https://icd.who.int/browsehttps://doi.org/10/2019/en.

https://doi.org/10.1016/j.ijmedinf.2023.105068
Received 4 September 2022; Received in revised form 27 March 2023; Accepted 5 April 2023
Available online 11 April 2023



#	Gender	Age Group	Diagnosis
1	M	41–50	E11, K02
2	M	41–50	E11, K02, K04
3	F	51–60	I10, M54
4	M	41–50	E11, K02, K04
5	F	51–60	K02, I10, M54

Fig. 1. Process to construct temporal disease occurrence network from disease sequences. Transform a patient record into an individual temporal disease occurrence network. Combine individual temporal disease occurrence networks into a combined temporal disease occurrence network. *S* represents the start of the disease progression and *T* represents the terminal disease.

investigate the structural properties of diseases network and a prediction model to predict disease comorbidity [18]. Ding *et al.* [18] combined association rule mining with clustering and collaborative filtering to predict the future conditions of patients from health insurance data. The method merged patient disease codes together which resulted in a loss of information. The prediction results showed that 71 % of acute and 82 % of chronic conditions are predictable.

Khan *et al.* [9] proposed a comorbidity network to study the progression of type 2 diabetes. However, the study was limited to selective patient cohorts having type 2 diabetes or heart disease. Lu *et al.* [5] proposed a bipartite graph to map patient and disease interactions, which was later transformed into an undirected graph. The graph convolution network was applied to the features obtained from the disease network and patient information to achieve prediction accuracy. However, the diseases and comorbidity relations relied on missing links that were predicted using the bipartite graph which may have introduced bias. Thygesen *et al.* [19] used the approach developed by Siggard and colleagues [4] to study COVID-19 event trajectory networks to chronologically sort COVID-19 events.

Previous research has also used link prediction algorithms [20–25], artificial neural networks [10,26], Markov chains [17], and clustering algorithms [11,27] to study disease risks by analyzing the frequency of different disease conditions that may occur simultaneously among patient cohorts. Although neural network-based algorithms and matrix factorization-based algorithms showed good performance, they cannot generate easily understandable rules due to the complexity of their internal decision-making process. Thus, the results generated by these models are not interpretable for physicians and patients.

The aforementioned studies were mostly successful, however, there are a few fundamental elements that need to be addressed. First, these studies introduced disease co-occurrence networks [16], comorbidity networks [5–7,9,16,28], and disease progression [4,15,29] but they did not take into account that by retaining patient information, a generalized network can be developed that combines the characteristics of different networks. These networks transform patient information into an intermediate format that results in the loss of key information such as temporal information, disease progression (transition from one disease to another), prevalent cohorts, the disease starting point, and disease mortality. Second, while they have proposed different networks, extracting useful information from these networks is still an open research problem. Finally, these studies are limited to immediate descendants and were not able to extend beyond direct immediate descendants (i.e. $A|B$ and $B|C$ can model disease progression, but not $A|B$, C). Therefore, these models are not capable of modeling dependencies that have higher disease sequence lengths.

In this paper, we propose a new method for creating a temporal disease occurrence network that is capable of retaining patient information along with temporal information. This method can be used to transform networks into sub-networks to study the morbidities, disease progression, and risk prediction in specific patient cohorts. This approach uses a patient's medical record as input and generates a ranked list of future disease progression along with *conditional probabilities* and *relative risks*. The supervised depth first search approach is used to find the *k*-frequent diseases from the current state of disease progression. A ranked list containing diseases and their probability of occurrence in the future is generated in the next step. The higher the probability score the more likely the disease will occur in the future, whereas higher relative risk values indicate a higher association between two diseases. The proposed model can help physicians with diagnosis and prognosis by providing an understandable score.

To evaluate the performance of the proposed method, we used an *area under the receiver operating characteristic curve* (AUC) and an F1-score to quantify the prediction accuracy. Experimental results show that the proposed approach is suitable for studying patient cohorts with certain morbidities and disease progression by taking into account the patient's recent past illnesses.

2. Method

2.1. Temporal disease occurrence network

This section describes a network-based approach that captures the patterns of morbidities and comorbidities as they occur over time, allowing for the prediction of a patient's future conditions. The temporal disease occurrence network represents disease progression based on the availability of overall health sequences of a population in terms of disease progression over time. This network incorporates both the properties of a disease occurrence network and the temporal aspect of diseases.

First, we constructed the health trajectory of an individual patient that revealed the patient's transition from one disease to another during the patient's healthcare visits over time. Second, an overall disease progression network was generated by aggregating individual disease networks and iteratively updating the node and edge level attributes (shown in Fig. 1). The node and edge level attributes maintain the information of associated diseases that have occurred in patients, as well as the basic information (e.g. patient labels, gender, and age group) of patients who exhibit similar progression between two diseases.

2.2. Frequent disease occurrences

Depth first search is a graph traversal algorithm used to search graph

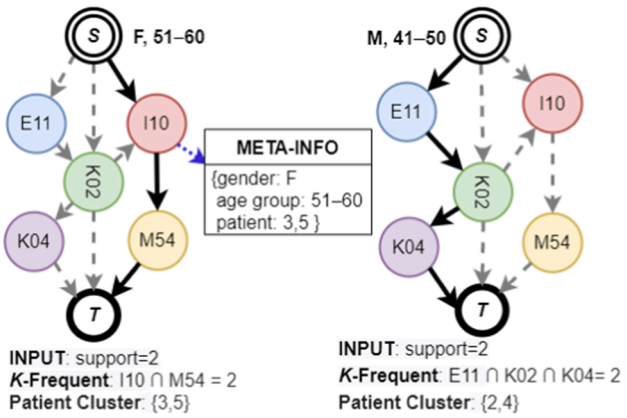


Fig. 2. *k*-Frequent disease sequences using node and edge level attributes to guide depth first search. The support threshold 2 is used. *k*-Frequent disease sequences for a female of age group 51–60 and *k*-frequent disease sequence for a male of age group 41–50 and corresponding patient clusters.

data structures. The algorithm explores all the nodes as far as possible before backtracking. The supervised depth first search strategy explores the branches based on patients' gender and age groups and stops exploring depth when a relative threshold is reached. The supervised depth first search traversal algorithm was used to extract the frequent disease sequences which occurred in the selected patient cohorts and temporal disease occurrence network. The algorithm uses patients' gender, age group, and relative *support count* threshold in its application.

The support threshold determines how often a group of diseases appear together in patients of a particular gender and age group. A simple intersection between the patient labels stored on every node of d_i and d_j determines whether the sequence is frequent based on the relative support threshold. The *meta*-information, such as patient labels under gender and age group, guides the depth first search to explore the related frequent sequences. It stops exploring the depth when a given threshold is reached and then moves to unexplored paths. All the paths visited during the traversing are labeled as *frequent disease sequences* that co-occurred (shown in Fig. 2). However, if a group of diseases appears too infrequently, it is likely just by chance and it will not be considered significant. Thus, a frequent disease sequence must have a support greater than a threshold value. The frequency of a diagnosis from a single diagnosis to its successor and variations in patient disease sequences is shown in Fig. 3.

2.3. Prediction model

After obtaining the *k*-frequent disease sequences for each gender and age group, the next step was to select the most prevalent disease sequences based on patient history. To achieve this, we selected the disease sequences that had a higher length and showed progression from the diseases that the patients in the corresponding group had. We sorted the disease sequences based on length and merged them one by one until we had the desired number of diagnoses. To make the results easier to understand, we ranked the list based on conditional probabilities and relative risk scores. We defined the conditional probability for an event d_i given d_j and relative risk for a sequence d_i and d_j as:

$$CP(d_i|d_j) = \frac{p(d_i d_j)}{p(d_i)} \tag{1}$$

where d_i occurred before d_j. The higher value indicates that d_i and d_j will occur together in a temporal order. The relative risk is defined as:

$$\mathrm{RR}_{d_i,j} = \frac{\text{observed}}{\text{expected}} = \frac{p(d_i, d_j)}{p(d_i)p(d_j)} \tag{2}$$

where $p(d_i)$ and $p(d_j)$ are the probabilities that a patient has the disease d_i and d_j, respectively, and $p(d_i d_j)$ is the probability that a randomly chosen patient has both diseases in a temporal order. An RR value > 1.0 indicates that the two diseases are highly associated. An example is shown in Fig. 4.

The proposed prediction method utilizes the patient's gender and age group along with the diagnosis history as a base to find the frequent sequences of pre-existing diagnoses, merging the most commonly occurring diagnoses together based on their frequencies, conditional probabilities, and relative risk scores. These factors are used to rank the likelihood of future disease progression.

2.4. Evaluation criteria

In order to perform a fair evaluation of the prediction algorithm we divided the dataset into a training set and a testing set using the *k*-fold cross validation method with $k = 20$. The sequence from the test set is further split into two distinct parts in a temporal order. The first part of the sequence is used as patient history and the latter as ground truth to validate the result.

The algorithm assigns a higher score to a diagnosis that has a higher probability to occur in the temporal disease occurrence network. The score quantifies the likelihood of a disease co-occurrence with existing diseases. If a conditional probability score shows that diagnosis has a higher chance to occur, then the occurrence of the disease is confirmed and considered likely to occur in the near future. The AUC and F1-score

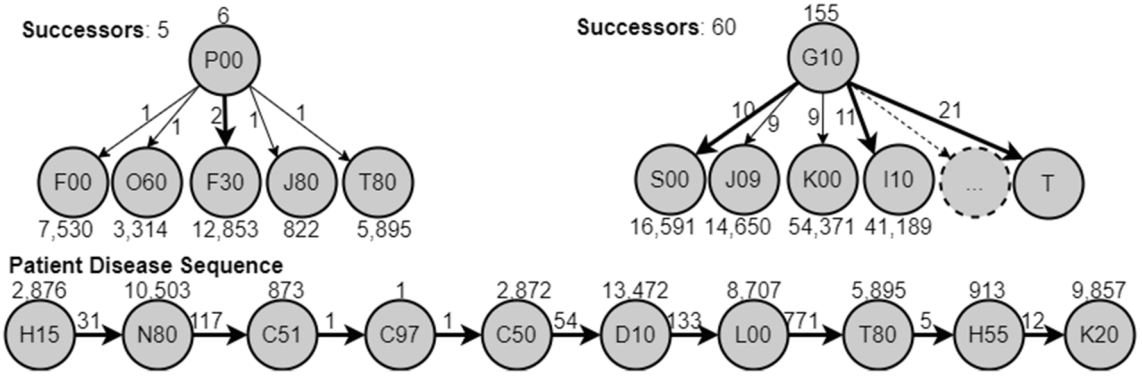


Fig. 3. Examples of infrequent diseases in temporal disease occurrence network. A rare disease (P00: fetus and newborn affected by maternal factors and by complications of pregnancy, labor, and delivery) with lower support count and slower progression. Huntington's disease (G10) with higher progression rate than 60 different diseases and has high probability to occur as the final disease in a sequence. A real patient sequence containing malignant neoplasms at multiple sites (C97) as a disease with lowest support and sequences with higher support values.

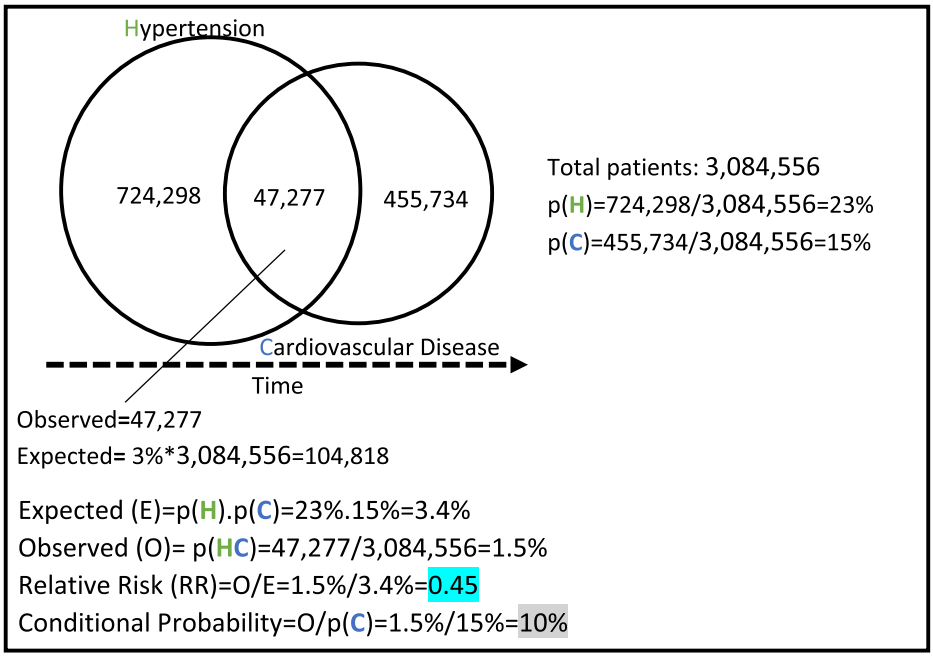


Fig. 4. Measuring the likelihood of future progression (I10->I30) using relative risk and conditional probability. Here hypertension and heart disease show low correlation but higher likelihood to occur in the future. If they were independent of each other, the probability of a person having both should be p(A) p(B) = 3.4 % while their observed co-occurrence is 10 %.

Table 1
Criteria for diagnosis selection.

ICD-10	3-Digits	Block Codes
All	1,383	250
Serious Diagnosis	1,232	196
Minus {R, U, V, W, Y, Z}	1,152	196

are used as evaluation criteria to assess the performance of the algorithm on the selected diagnosis. AUC values indicate the probability that a randomly chosen existing diagnosis is given a higher score than a randomly chosen non-existent diagnosis. The AUC is obtained by comparing the score of existing (true positive) and non-existing (false positive) diagnoses. From n independent observations, let n_e observation result in a higher score for existing diagnosis and n_{ne} observation have resulted in the same score, then the AUC is calculated as follows [30],

$$AUC = \frac{n_e + 0.5n_{ne}}{n} \tag{3}$$

A good diagnosis prediction algorithm should have an AUC value close to 1. We also use precision and recall defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives. Based on the precision and recall, we calculate the F1-score as their harmonic mean:

$$\mathrm{F1} - score = 2^{*}\frac{Precision^{*}Recall}{Precision + Recall} \tag{6}$$

3. Experimental setup

3.1. Dataset

We use a patient dataset that was used in previous studies [13]; supplementary information about the dataset can be found online.[2] The patient data was acquired from the national administrative Care Register for the years 2015–2018**.** The Health Care Registers include patients' basic information (age, gender, municipality of residence) and service events such as the type of contact (visit, phone call, inpatient admission), and the reason for the visit (recorded using ICD-10 or ICPC-2 codes). The database contains information about 4.3 million patients that are over 18 years old along with patient identity code, gender, age, date of visit, and disease code. Patients diagnosed with a disease are marked in the register during their visits to healthcare centers. Patients received 1.6 diagnoses on average during a single visit to healthcare centers.

Let n represent the total number of patients obtained from the clinical data set of the patient histories. The data preparation module transforms the raw patient data into a new data set $R = \{r_1, r_2, r_3 \ldots, r_m\}$, where R is a patient health center visit record, and summarizes the patient medical histories related to m number of diseases. Running analyses on m number of codes individually makes it difficult to study, comprehend, and visualize the results. As the ICD-10 system has a hierarchical structure, we used two different structures to round diagnosis codes which are aligned with WHO's recommended block codes. These codes comprise initial 3-digit codes (e.g. E11: E11.9) and block codes with the first 3-digits (e.g. A00: A00–A09, A15: A15–A19).

Extending the diagnosis selection process further excludes duplicate diagnosis codes that were used for general symptoms and signs, administrative purposes, and external causes (shown in Table 1). The process is repeated for both the 3-digit and block codes. The data is transformed into a sequence of medical disease history as shown in Table 2, which consists of a unique set of diagnoses against every

[2] https://cs.uef.fi/ml/impro/prediction/.

Table 2
Sample patient records.

Patient	Gender	Age Group	Diagnosis
1	M	41–50	E11 K02
2	M	41–50	E11 K02 K04
3	F	51–60	I10 M54
4	M	41–50	E11 K02 K04
5	F	51–60	K02 I10 M54

Table 3
Temporal disease occurrence statistics after filtering sequences with < 5 diseases.

	3-Digits	Block Codes
Average Sequence Size	8	5
Patients	3,987,382	3,084,556
Nodes	1,383	196
Edges	567,881	31,800

patient. The first occurrence of diagnosis is considered as evidence of disease progression. Then we filter out the patients with<5 diagnosis codes in any sequence. After filtering, there are 3,987,382 patient records when diagnoses are rounded to 3-digit codes and 3,084,556 patient records for block codes. The selected data contains 351,652 patients, of which 11 % of the patient population has been diagnosed with diabetes (E10), 1,542,285 patients (45 %) with dental caries (*K*00), and there are 196 distinct ICD-10 block codes. The number of diagnoses for each patient ranges from 5 to 72.

3.2. Compared methods

We compare the performance of the proposed algorithm with the following set of link prediction algorithms.

Common Neighbor: Common Neighbor [20] is a simple approach used for link prediction. Two diseases from a patient's recent history are likely to form a link if they have *common neighbors* (CN) in the temporal disease occurrence network. Index $CN(d_i, d_j)$ for the common neighbor method is computed as:

$$CN(d_i, d_j) = |\Gamma(d_i) \cap \Gamma(d_j)| \tag{7}$$

where d_i and d_j are two diseases and $\Gamma(d_i)$ and $\Gamma(d_j)$ denote the set of neighbor diseases of disease d_i and d_j.

Aggregate Common Co-occurrence (ACC): This algorithm computes the pointwise mutual information to determine the degree of association between given diseases and their common co-occurred diseases [20]. The frequency of co-occurrences is counted for every single disease that co-occurred with another disease. ACC is defined as:

$$ACC(d_i, d_j) = |\Gamma(d_i) \cap \Gamma(d_j)| \tag{8}$$

where $\Gamma(d_i)$ and $\Gamma(d_j)$ are sets of neighbors of nodes d_i and d_j, respectively, and the aggregate common co-occurrence of diseases is calculated by summing the number of co-occurrences of common diseases.

Jaccard Index (JI): The JI [21] considers only the common co-occurrence between the diseases and is defined as follows:

$$JI(d_i, d_j) = \frac{|\Gamma(d_i) \cap \Gamma(d_j)|}{|\Gamma(d_i) \cup \Gamma(d_j)|} \tag{9}$$

Preferential Attachment (PA): The PA algorithm depends on the growth of diseases co-occurred with diseases d_i and d_j [22]. Note that k_i represents the degree of disease d_i and k_i represents the degree of disease d_j. PA is defined as follows:

$$PA(d_i, d_j) = k_i \cdot k_j \tag{10}$$

Common Neighbor and Centrality-Based Parameterized Algorithm (CCPA): CCPA uses the common neighbor and centrality to identify the potential future connection between nodes [23].

$$CCPA(d_i, d_j, \alpha) = \alpha \cdot (|\Gamma(d_i) \cap \Gamma(d_j)|) + (1 - \alpha) \cdot C(d_i, d_j) \tag{11}$$

Supervised Random Walk (SRW): SRW is a Markov chain that visits a sequence of nodes using a random walker [24]. This process involves state transition, where a random walk is started from a given node, and at each step, the walker decides the next node using the transition probability. The walk ends when a termination node is reached [25].

4. Results and discussions

To discuss and analyze the temporal disease occurrence network and prediction, we divide this section into two parts: network analysis and prediction model.

4.1. Disease sequence analysis

We select patients that have>5 diagnoses from a total of 4.3 million patients' health records. The selected patient records include 1,383 distinct diseases as shown in Table 3. We transformed these records into a temporal disease occurrence network, where every node in the network represents a unique disease. The path between diagnoses represents the temporal occurrence of diseases in patients, whereas node and edge level attributes contain patient *meta*-information such as gender and age group. Constructing temporal disease occurrence network from unique sequences avoid the self-loop in temporal disease occurrence networks. The temporal disease occurrence network is shown in Fig. 5. This network allows us to study different characteristics that show the highly probable diseases with a higher risk of morbidity progression. We applied community detection algorith with resolution 0.5 and found 17 diagnosis communities having similar structural structural characteristics [31].

Our analysis first focused on the frequencies of individual diseases within the temporal disease occurrence network, which was constructed using ICD-10 block codes (196 nodes). We observed that diseases of the oral cavity, salivary glands, and jaws (K00–K14 = 1,542,285) occur at a higher rate in the Finnish population, followed by hypertension (I10 = 724,298). The results indicate that disease groups such as oral cavity, salivary glands, and jaws (K00), hypertension (I10), acute upper respiratory infections (J00), deforming dorsopathies (M40), disorders of the muscles (M60), and acute pericarditis (I30) are prominent and highly likely to occur in a large patient cohort with a higher risk of progression.

To investigate sequential disease progression, we analyzed the successors of all nodes and identified 23,994 pairs of directional diagnoses. The directional strength of a diagnosis d1->d2 or d2->d1 is determined by RR > 1 or CP > 1 %. The RR value reveals the correlation strength between the diagnosis in a direction whereas the conditional probability measures the likelihood of one disease occurring based on the presence of another. During this process, we identified a significant population-wide disease progression (shown in Table 4) and their link statistics. We identified 932 pairs of diagnoses with high correlation (RR > 1) and likelihood (CP > 1 %) to occur in the near future. Among these, 89 pairs of diagnoses had a higher occurrence rate in terms of their unconditional probability but lower correlation (RR < 1), and 111 of them appeared to have a low occurrence rate in terms of their unconditional probability and high correlation and progression. We found that a proportion of the Finnish patient population (39 %) were diagnosed with periodontal or gum disease and is at significant risk of metabolic syndrome (710,707), which is consistent with evidence from a recent study that showed periodontal disease and metabolic syndrome were linked [32]. We also observed a strong correlation (RR > 1) between neurotic stress-related disorders (F40) and mood disorders in 53,750 patients in both directions.

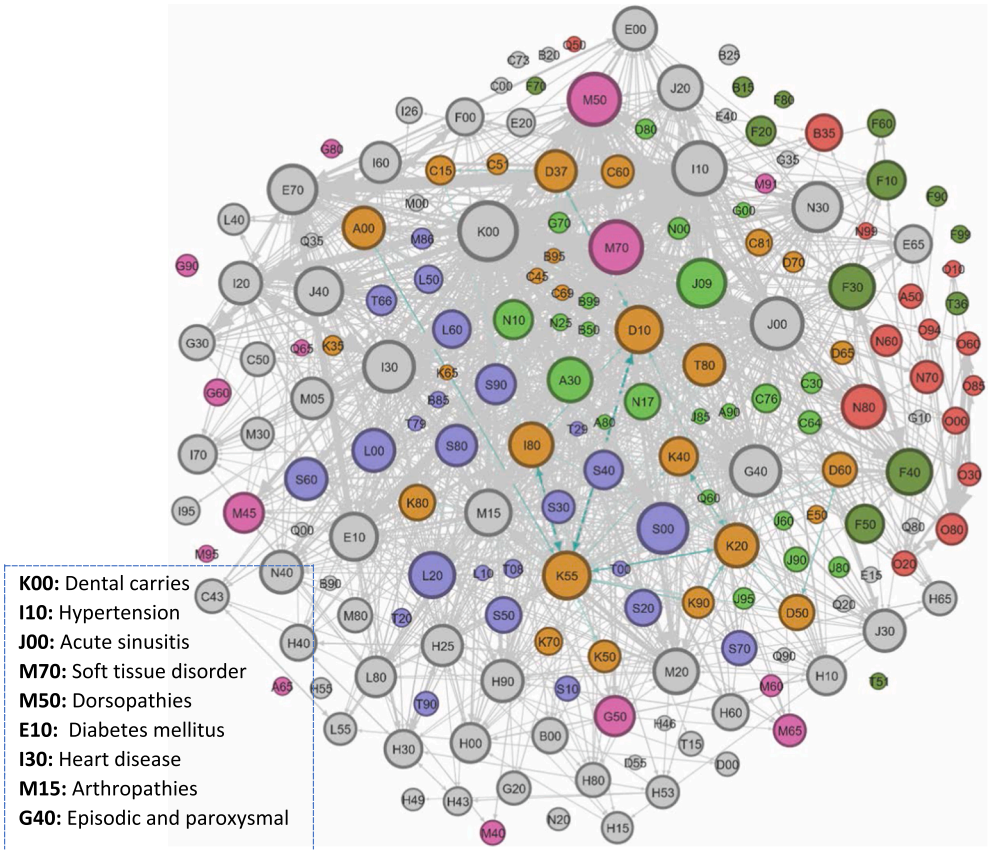


Fig. 5. Visualization of temporal disease occurrence network and diagnoses communities. The node size shown is proportional to disease occurrence and nodes with same color represent communities associated with the same characteristics. A large node has higher occurrence, higher connections, and is very central in the temporal disease occurrence network.

Table 4
Comparison of disease occurrences using individual, forward, and backward link frequencies, their conditional probabilities, and relative risk values. RR > 1 shows a higher risk of disease progression.

ICD	Description	Patients	Forward			Backward		
			f	CP	RR	*f*	CP	RR
F00	Mental disorders	114,873	27,715	24 %	7.19	11,760	11 %	3.05
G30	Degenerative disease of the nervous system	103,427						
F40	Neurotic stress-related disorders	74,895	20,793	7.59 %	0.92	32,957	13 %	1.46
F30	Mood disorders	254,328						
H25	Disorders of lens	188,262	11,564	6 %	1.69	8,493	8 %	1.24
H30	Disorders of choroid and retina	112,153						
F70	Mental retardation	11,816	327	3 %	8.17	142	1 %	3.56
G80	Cerebral Palsy	10,453						
F70	Mental retardation	11,816	34	2 %	4.87	26	0 %	3.72
Q00	Congenital malformation	1,824						
I100	Acute rheumatic fever	114	5	0 %	3.51	5	4 %	3.51
L50	Urticaria and erythema	38,518						
G35	Demyelinating diseases	11,477	147	3 %	7.76	194	2 %	10.24
H46	Disorders of the optic nerve	5,091						
A50	Predominantly sexual transmission	35,289	46	11 %	9.64	35	0 %	7.34
A70	Disease caused by chlamydia	417						
K00	Oral cavity	1,542,285	21,462	6 %	0.12	16,877	1 %	0.09
E10	Diabetes mellitus	351,652						
E10	Diabetes mellitus	351,652	35,440	10 %	0.86	10,409	3 %	0.25
E70	Metabolic disorders	359,055						
J00	Acute upper respiratory infections	569,295	22,485	4 %	0.19	22,529	4 %	0.19
M40	Deforming dorsopathies	626,253						
J00	Acute upper respiratory infections	569,295	23,514	4 %	0.19	22,931	4 %	0.18
M60	Disorders of muscles	679,583						
I10	Hypertensive diseases	724,298	47,277	10 %	0.44	36,219	5 %	0.34
I30	Heart disease	455,734						
I30	Heart disease	455,734	23,321	10 %	0.66	27,720	6 %	0.78
I20	Ischaemic heart diseases	239,119						

** Forward: F00–>G30 Backward: G30–>F00.

Table 5
Selected frequent and rare sub-sequences from temporal disease occurrence network constructed using ICD-10 block codes.

Frequent Patterns	Patients		
	Forward	Backward	Both
K00–>M15–>M70	4,160	3,852	8,012
I10–>K00–>M15–>M70	143	107	250
H30–>E10–>I10	459	243	702
I10–>E10–>K00	905	3,320	4,225
I10–>E10–>I20	589	1,142	1,731
F40–>F30–>I30	155	200	355
F30–>F40–>I30	234	180	414
F40–>I30–>F30	81	71	152
J20–>J40–>I20	115	106	221

Forward: K00–>M15–>M70 Backward: M70–>M15–>K00.

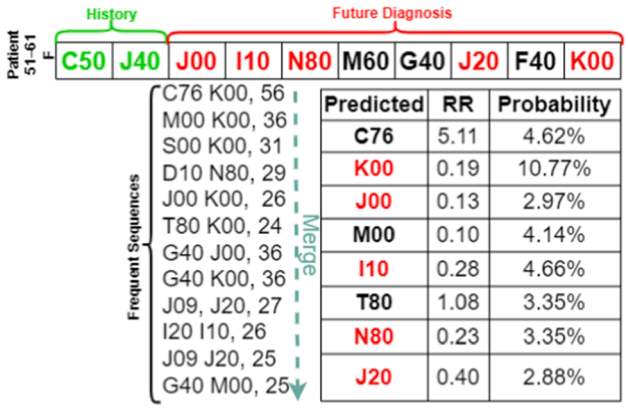


Fig. 6. Prediction on a real patient record from the test set. The gender of the patient is female, and the age group is 51–60. The test set is divided into history and future diagnosis (on top). *k*-Frequent sequences occurrence (on right), and *N*-ranked list of diagnoses that are likely to occur in the future (on left).

To study larger sequences (shown in Table 5), we use depth first search to find the longer sequences from the temporal disease occurrence network. The process finds the sequences of sizes 2 to 6. To maintain a unique list of sequences, we filtered out all the sequences of smaller size that co-exist within another sequence of higher size. We found a total number of 23,517 unique sequences and their frequencies {1,760, 8,946, 9,427, 3,111, 273}, ranging from 2 to 6. This highlights the complexity of disease heterogeneity and disease progression. We applied this method to patients with diagnoses of oral cavity, hypertension, mental disorder, and epilepsy (ICD-10 codes K00, I10, F00, and G40 respectively) on two different cohorts. The number of resulting subsequent underlying conditions are 6,667 for K00, 2,637 for I10, and 2,388 for F00 or G40. The results show that these sequences are complex in nature and have a high multimorbidity spectrum that necessitates protocols and strategies for studying multiple diseases and multiple therapies jointly [4,33]. Note that the underlying data reflects that patients received 1.6 diagnoses per visit, which may not accurately represent the true temporal relationships between diseases.

Our approach of using a temporal disease occurrence network and supervised depth first search is a better fit for identifying directional pairs of diagnosis and longer disease sequences occurring in patient cohorts of different genders and age groups. In comparison, using association rule mining algorithms to find temporal pairs of diagnosis and longer disease trajectories requires every pair of diseases to be iteratively added and the number of times the trajectory occurred in the patient population counted [4,15,19]. That method requires a series of scans to extend the directed pair of diagnosis into a longer disease trajectory and takes $O(2^T)$ time to obtain the longer disease trajectory of size T. Therefore, the number of scans grow exponentially, resulting in higher time complexity which is not suitable for larger datasets with many longer disease trajectories. In contrast, our approach requires a single scan to construct the temporal disease occurrence network in $O(n)$ time and a single supervised depth first search to obtain a temporal disease sequences $O(v + e + \log(p))$ time, where v represents the diagnosis on each node, e represents the directed edge between diagnoses, and p is the number of patients on each edge. In addition, our method also retrieves the prevalent disease cohorts of a given gender and age

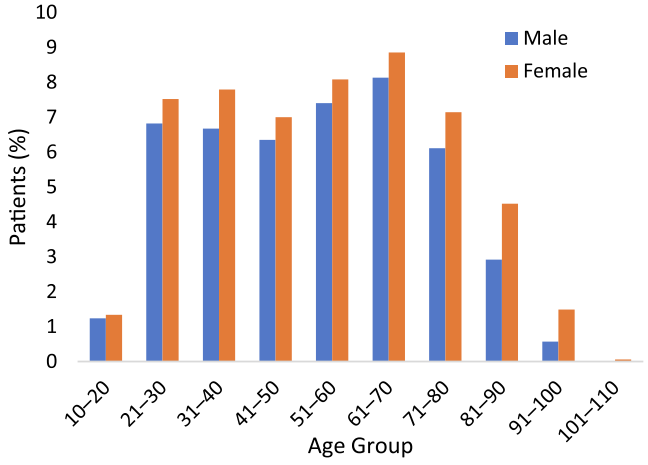


Fig. A3. Patients by gender and age group.

Table 6
Comparison of algorithms' accuracy quantified by AUC and F1-score. Each algorithm applied on temporal disease occurrence network constructed using ICD-10 3-digit disease codes and block codes. The obtained ranked list of size 1 and relative to ground truth is used to compare the results with selected ground truth.

List Size	3-Digit Codes				Block Codes			
	Next Disease		Relative		Next Disease		Relative	
	AUC	F1-score	AUC	F1-score	AUC	F1-score	AUC	F1-score
*FDO (Proposed)	**0.65 ± 0.21**	**0.11**	**0.58 ± 0.09**	**0.16**	**0.68 ± 0.20**	**0.13**	**0.62 ± 0.09**	**0.23**
**FDO (Proposed)	**0.66 ± 0.19**	**0.12**	**0.60 ± 0.08**	**0.17**	**0.72 ± 0.13**	**0.16**	**0.64 ± 0.06**	**0.25**
***FDO (Proposed)	**0.66 ± 0.19**	**0.12**	**0.60 ± 0.08**	**0.17**	**0.72 ± 0.13**	**0.16**	**0.64 ± 0.06**	**0.25**
CN [20]	0.63 ± 0.22	0.09	0.58 ± 0.09	0.16	0.66 ± 0.23	0.12	0.61 ± 0.09	0.22
SRW [24]	0.60 ± 0.15	0.10	0.54 ± 0.07	0.05	0.62 ± 0.22	0.12	0.56 ± 0.06	0.15
ACC [20]	0.53 ± 0.11	0.02	0.56 ± 0.07	0.11	0.53 ± 0.11	0.02	0.58 ± 0.09	0.16
JI [21]	0.55 ± 0.15	0.04	0.54 ± 0.07	0.07	0.55 ± 0.16	0.04	0.55 ± 0.08	0.14
PA [22]	0.60 ± 0.20	0.07	0.54 ± 0.03	0.12	0.63 ± 0.22	0.10	0.55 ± 0.03	0.16
CCPA [23]	0.50 ± 0.01	–	0.56 ± 0.07	0.12	–	–	0.55 ± 0.03	0.16

* **Evaluation set $k = 20$.**
** **Evaluation set $k = 10$.**
*** **Training set $k = 10$.**

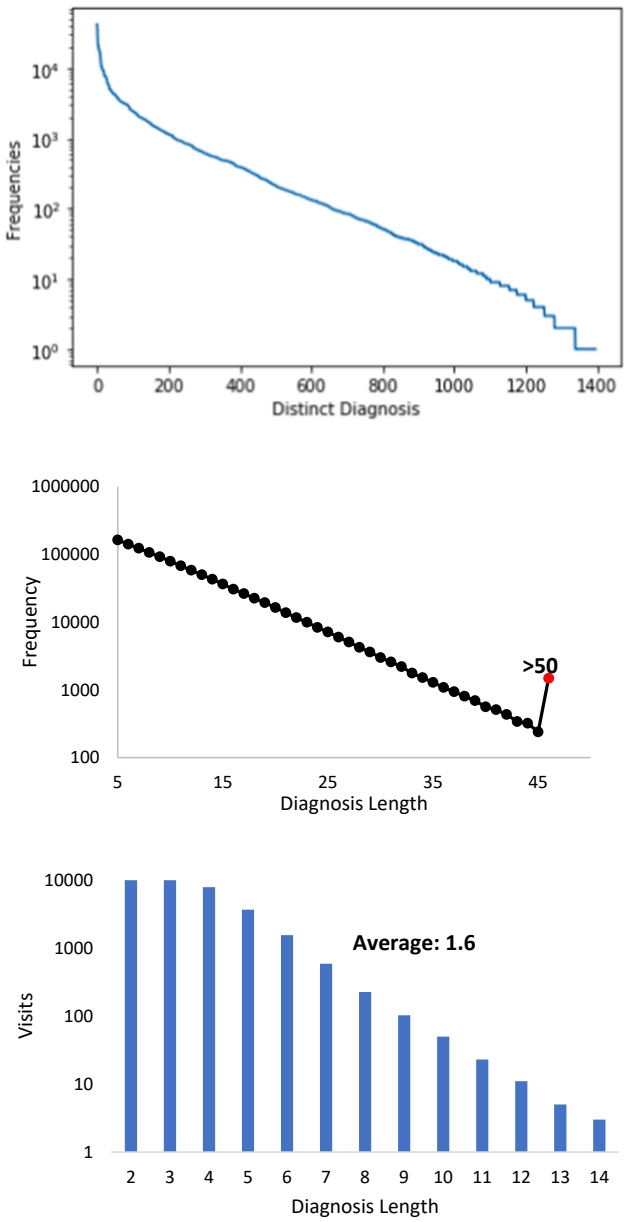


Fig. A4. Frequency distribution of diseases (on top) shows the prevalent and rare ICD-10 disease codes in the dataset, different lengths of disease sequences in patients (middle), and average diagnosis assigned per visit during a patient visit to a health unit (bottom).

and updates to patient diagnoses are easy and changes reflect in real-time.

4.2. Prediction

We used disease sequences to predict future disease progression and tested the proposed prediction method in the following way. We randomly divided the disease sequences into two sets, with 80 % being used for training and 20 % for testing. The sequences selected for training were used to construct the temporal disease occurrence network and generate frequent disease sequences using the supervised depth first strategy. The supervised depth first strategy approach reveals comorbidity relations and generates predictions about which diseases a patient may incur based on their gender, age group, and disease history.

Each sequence of diagnosis in the testing set t is divided into two parts based on past and future diagnoses. The first part of diagnosis, called t_{head} along with patient gender and age group, is used for generating predictions, while the remaining diagnosis, referred as t_{tail}, is used to evaluate the predictions generated. The length of t_{head} is closely related to the maximum window size allowable for the experiments. Thus, given a window size $w \leq |t|$, we select the first w diseases for generating the predictions and the remaining for $|t|-w$ for testing the predictions. With a fixed t_{head} and a method specific threshold T, we produce the prediction set $P(t_{\text{head}},T)$ containing all the predictions whose score is $\geq T$. Then we compared the predicted set with t_{tail}.

For example, we used a real patient record from test set t = {51–61, F, C50, J40, J00, I10, N80, M70, G40, J20, F40, K00} that is divided into two parts t_{head} = {51–61, F, C50, J40} and t_{tail} = {J00, I10, N80, M70, G40, J20, F40, K00}. The prediction set relative to ground truth produced $T = 8$ and is shown in Fig. 6. The result suggests that 5 out of 8 diagnoses exist in the ranked list whereas 3 diagnoses are non-existent in the ranked list. The probability score and relative risk score show that non-existent diagnoses are highly likely to occur as the disease progresses further. The probability score along with the ranked list is easier to understand for patients, physicians, and decision-makers [16,34]. Both physicians and decision-makers can combine their domain-specific knowledge with probability-based generated future conditions with our proposed algorithm in clinical decision-making and planning. Patients can understand the risk and follow the guidelines provided by physicians to prevent or delay disease progression.

We measured the accuracy of the predictions using the AUC and an F1-score on the temporal disease occurrence network constructed based on 3-digit ICD-10 and block codes (shown in Table 6). We used two different ranked list sizes to calculate the F1-score and AUC. In general, the *frequent disease occurrence* (FDO) method outperforms other methods. The FDO method had higher accuracy for AUC and F1-score compared to other methods, except for the PA method, for all networks constructed on different diagnosis selection criteria. The improvements in the temporal disease occurrence network constructed on ICD-10 block codes were particularly significant. Our analysis showed that CN-based algorithms assign equal weight to the common neighbors

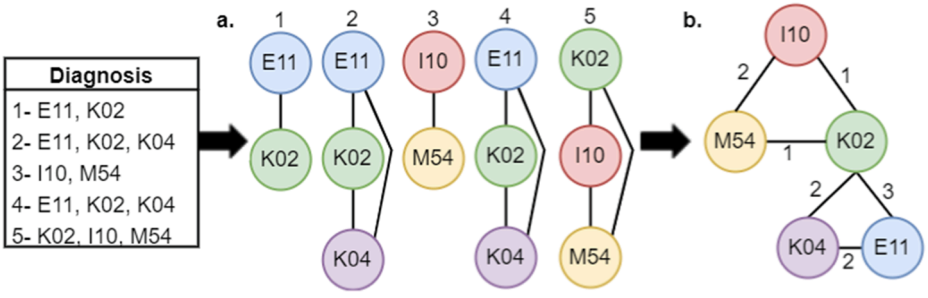
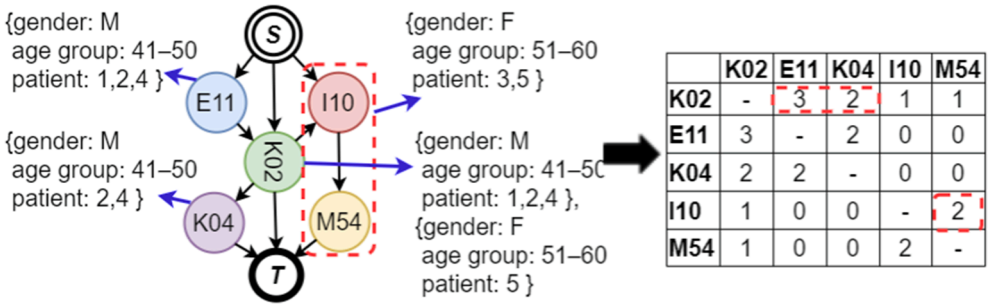


Fig. A1. Process to construct disease co-occurrence network from disease sequences. Transform a patient record into disease co-occurrence network. Combine individual disease co-occurrence networks into a combined disease co-occurrence network.



	K02	E11	K04	I10	M54
K02	-	3	2	1	1
E11	3	-	2	0	0
K04	2	2	-	0	0
I10	1	0	0	-	2
M54	1	0	0	2	-

Fig. A2. Process to transform temporal disease occurrence network into co-occurrence matrix.

whereas the FDO effectively utilizes the frequent disease sets and the conditional probability of all possible diagnoses that can co-occur with a given diagnosis.

For validating the consistency across different folds of data we performed additional tests. First, we divided the dataset into a training set and a testing set using the k-fold cross validation method with $k = 10$. The performance of the prediction algorithm slightly improves and a slight decrease in STD values is observed. However, STD remains relatively high due to the nature of data distribution in temporal disease occurrence network. Secondly, we randomly obtained $k = 10$ from the training set to separately validate the performance of prediction method on training dataset. The results obtained are align with the earlier observations which indicates that performance of proposed method is consistent across different folds of the data.

One of the major challenges with machine learning, deep learning, and depth first search is that they are often considered "black box" approaches because the internal processes are not easily explainable to stakeholders. This lack of transparency can be problematic in critical applications where stakeholders such as physicians, decision makers, and patients need to understand the reasoning behind the decisions made by the model. In this regard, the depth first strategy reveals more information for the stakeholder with little effort. It is also important for stakeholders to be aware of the limitations and assumptions of the models being used and approach them with a critical eye.

5. Conclusion

Our study introduces a new approach for predicting the onset of diseases using a temporal disease occurrence network and supervised depth first search strategy to obtain frequent disease sequences. The rank list of disease sequences is then ranked based on conditional probability to predict disease onset. The advantage of this method is that it can determine the role of various diagnoses that are performed and assigns them different weights based on the computed probabilities. The proposed method is compared to existing neighbors-based algorithms such as aggregated common occurrence, SRW, JI, PA, and CCPA. We found that our approach outperformed these algorithms in terms of accuracy. In addition, our approach provides a rank list with probability and relative risk scores, which can help physicians to understand the sequence of disease events in patients.

6. Future work

Future work should aim to incorporate additional patient health information such as lab history and physician observations to better understand disease progression and improve disease prediction accuracy. Future work should also address inconsistencies in the dataset, such as patients receiving multiple diagnoses during a single visit (in the current dataset the average number of diagnoses received per visit is 1.6), which can make the temporal disease occurrence network sparse and reduce the accuracy of disease prediction algorithms. Furthermore, information on polypharmacy (increases harmful drug effects) [35], physical and physiological activities [36], and other laboratory tests should be included as they can help to improve prediction accuracy. To improve the awareness and early warning system, deceased patients should be marked on a termination node and a temporary termination node should be used for patients who are still alive and have a disease in progression. These future improvements will help us to develop more effective disease prediction models and provide better insights into disease progression. (See Fig. A3 and A4 for data statistics and inconsistencies in the dataset).

7. Summary points

- This study reports a novel prediction approach based on a temporal disease occurrence network and supervised depth first search algorithm to predict disease progression.
- The frequent disease occurrence algorithm creatively combines the frequent disease sequences based on relative risk and likelihood to occur in the future.
- The ranked list with conditional probability and relative risk score provides useful insights to understand disease progression.
- The method has reliable performance assessed by an AUC and F1-score.

8. Summary Table

What is already known on this topic?

- Phenotypic comorbidity networks are developed to investigate the connections between diseases.
- Co-occurrence networks and data mining approaches are used to predict disease co-occurrences.
- Pairwise disease progression is studied to analyze the morbidity progression.

What did this study add to our knowledge?

- To study highly probable diseases with a higher risk of morbidity progression, we can extract forward and backward disease progression, as well as *meta*-information, from the temporal disease occurrence network.
- Conditional probability determines the possibility of a disease occurring in the future and relative risk can be used to measure the correlation between diseases.
- Supervised depth first search strategy is a useful approach to obtain prevalent and longer sequences, identify patient clusters with prevalent disease progression and eliminate diseases that have a lower chance to occur in a specific gender and age group.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The project was funded by the Strategic Research Council (SRC) at the Academy of Finland (grant numbers 312706 and 336325).

Appendix A

Disease Co-Occurrence Network.

A *disease co-occurrence network* [16,27] represents the diseases that co-occurred in many patients. The disease co-occurrence network has also been called the phenotypic disease network [37], the comorbidity network [16], and the disease comorbidities network [38]. It has been used in previous studies [13,16,28,37–39]. In the disease co-occurrence network, we represent disease d as a node in the network and a connection between nodes represents the co-occurrence. First, we construct an individual patient disease co-occurrence network by connecting every d_i with d_j that co-occurred in an individual patient. Second, these individual co-occurrence networks are merged to form a combined disease co-occurrence network. The node-level attribute list maintains the unique identity list of the patient and the number of times a disease is recorded in the patient's electronic record. An edge exists between d_i and d_j if both diagnoses co-occurred in an individual patient. The disease co-occurrence network is shown in Fig. A1.

Disease Co-Occurrence Matrix.

The temporal disease occurrence network can be transformed into a *disease co-occurrence matrix* (shown in Fig. A2) that is used to construct a disease co-occurrence network. Two diseases d_i to d_j co-occurred if there is a path between d_i to d_j in the temporal disease occurrence network and can be defined as follows:

$$CM(d_i, d_j) = \begin{cases} |\Gamma(d_i) \cap \Gamma(d_j)| & \text{if path}(d_i, d_j) \\ 0 & \end{cases} \quad (1)$$

Appendix B. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2023.105068.

References

[1] S.G. Nadathur, Maximising the value of hospital administrative datasets, Aust. Health Rev. 34 (2) (2010) 216–223.
[2] A. Bottle, P. Aylin, Intelligent information: a national system for monitoring clinical performance, Health Serv. Res. vol. 43 (1p1) (2008) 10–31.
[3] J.F. Ludvigsson, C. Almqvist, et al., Registers of the Swedish total population and their use in medical research, Eur. J. Epidemiol. 31 (2) (2016) 125–136.
[4] T. Siggaard, R. Reguant, et al., Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7.2 million Danish patients, Nat. Commun. 11 (1) (2020) 4952.
[5] H. Lu, S. Uddin, A disease network-based recommender system framework for predictive risk modelling of chronic diseases and their comorbidities, Appl. Intell. 52 (9) (2022) 10330–10340.
[6] H. Lu, S. Uddin, F. Hajati, M.A. Moni, M. Khushi, A patient network-based machine learning model for disease prediction: the case of type 2 diabetes mellitus, Appl. Intell. 52 (3) (2022) 2411–2422.
[7] M.E. Hossain, S. Uddin, A. Khan, Network analytics and machine learning for predictive risk modelling of cardiovascular disease in patients with type 2 diabetes, Expert Syst. Appl. 164 (2021), 113918.
[8] A. Khan, S. Uddin, U. Srinivasan, Chronic disease prediction using administrative data and graph theory: the case of type 2 diabetes, Expert Syst. Appl. 136 (2019) 230–241.
[9] A. Khan, S. Uddin, U. Srinivasan, Comorbidity network for chronic disease: a novel approach to understand type 2 diabetes progression, Int. J. Med. Inf. 115 (2018) 1–9.
[10] H. Lu, S. Uddin, A weighted patient network-based framework for predicting chronic diseases using graph neural networks, Sci. Rep. 11 (1) (2021) 22607.
[11] D.A. Davis, N.V. Chawla, N.A. Christakis, A.-L. Barabási, Time to CARE: a collaborative engine for practical disease prediction, Data Min. Knowl. Disc. 20 (3) (2010) 388–415.
[12] D. A. Davis, N. V Chawla, N. Blumm, N. Christakis, and A.-L. Barabasi, "Predicting individual disease risk based on medical history," in *Proceedings of the 17th ACM conference on Information and knowledge management*, 2008, pp. 769–778.
[13] S. Sieranoja, P. Fränti, Adapting k-means for graph clustering, Knowl. Inf. Syst. 64 (1) (2022) 115–142.
[14] K. Steinhaeuser, N.V. Chawla, "A network-based approach to understanding and predicting diseases", in *Social computing and behavioral modeling*, Springer (2009) 1–8.
[15] A.B. Jensen, P.L. Moseley, et al., Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients, Nat. Commun. 5 (2014) 4022.
[16] F. Folino, C. Pizzuti, and M. Ventura, "A comorbidity network approach to predict disease risk," in *International Conference on Information Technology in Bio-and Medical Informatics*, 2010, pp. 102–109.
[17] F. Folino and C. Pizzuti, "Combining Markov models and association analysis for disease prediction," in *International Conference on Information Technology in Bio-and Medical Informatics*, 2011, pp. 39–52.
[18] R. Ding, F. Jiang, J. Xie, Y. Yu, Algorithmic prediction of individual diseases, Int. J. Prod. Res. 55 (3) (2017) 750–768.
[19] J.H. Thygesen, C. Tomlinson, et al., COVID-19 trajectories among 57 million adults in England: a cohort study using electronic health records, The Lancet Digital Health 4 (7) (2022) e542–e557.
[20] F. Lorrain, H.C. White, Structural equivalence of individuals in social networks, J. mathematical social. 1 (1) (1971) 49–80.
[21] P. Jaccard, Bulletin de la société vaudoise des sciences naturelles, Etude comparative de la distribution florale dans une portion des Alpes et des Jura 37 (1901) 547–579.
[22] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.
[23] I. Ahmad, M.U. Akhtar, S. Noor, A. Shahnaz, Missing link prediction using common neighbor and centrality based parameterized algorithm, Sci. Rep. 10 (1) (2020) 1–9.
[24] S. Brin, L. Page, The anatomy of a large-scale hypertextual Web search engine, Computer Networks and ISDN Systems 30 (1) (1998) 107–117.
[25] L. Backstrom and J. Leskovec, "Supervised random walks: predicting and recommending links in social networks," in *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011, pp. 635–644.
[26] J. Zhang, J. Gong, and L. Barnes, "HCNN: Heterogeneous convolutional neural networks for comorbid risk prediction with electronic health records," in *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, 2017, pp. 214–221.
[27] F. Folino and C. Pizzuti, "A comorbidity-based recommendation engine for disease prediction," in *2010 IEEE 23rd International Symposium on Computer-Based Medical Systems (CBMS)*, 2010, pp. 6–12.
[28] H. Tanushi, H. Dalianis, and G. Nilsson, "Calculating prevalence of comorbidity and comorbidity combinations with diabetes in hospital care in Sweden using a health care record database," 2011.
[29] X. Ji, S. A. Chun, J. Geller, and V. Oria, "Collaborative and trajectory prediction models of medical conditions by mining patients' Social Data," in *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2015, pp. 695–700.
[30] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1) (1982) 29–36.
[31] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech: Theory Exp. 2008 (10) (2008) P10008.

[32] K. Watanabe, S. Katagiri, et al., Porphyromonas gingivalis impairs glucose uptake in skeletal muscle associated with altering gut microbiota, FASEB J. (2020).
[33] J. Woodcock, L.M. LaVange, Master protocols to study multiple therapies, multiple diseases, or both, N. Engl. J. Med. 377 (1) (2017) 62–70.
[34] G.K. Lighthall, C. Vazquez-Guillamet, Understanding decision making in critical care, Clin. Med. Res. 13 (3–4) (2015) 156–168.
[35] E.R. Hajjar, A.C. Cafiero, J.T. Hanlon, Polypharmacy in elderly patients, Am. J. Geriatr. Pharmacother. 5 (4) (2007) 345–351.
[36] Z. Zajkowska, A. Walsh, et al., A systematic review of the association between biological markers and environmental stress risk factors for adolescent depression, J. Psychiatr. Res. 138 (2021) 163–175.
[37] C.A. Hidalgo, N. Blumm, A.-L. Barabási, N.A. Christakis, A dynamic network approach for the study of human phenotypes, PLoS Comput. Biol. 5 (4) (2009) e1000353.
[38] M.J. Divo, C. Casanova, et al., COPD comorbidities network, Eur. Respir. J. 46 (3) (2015) 640–650.
[39] K. Srinivasan, F. Currim, S. Ram, Predicting high-cost patients at point of admission using network science, IEEE J. Biomed. Health Inform. 22 (6) (2017) 1970–1977.