Original Paper

Frequent Pattern Mining for Early Risk Signal Detection in Population Health Data: Nationwide Retrospective Cohort Study

Sazid Nur Ratul[1], BSc, MSc; Jimi Patrik Tuononen[1], BSc, MSc; Tiina Laatikainen[2], MD, PhD; Pasi Fränti[1], MSc, PhD

[1]School of Computing, Faculty of Science, Forestry and Technology, University of Eastern Finland, Joensuu, North Karelia, Finland
[2]School of Medicine, Faculty of Health Sciences, University of Eastern Finland, Kuopio, North Savo, Finland

Corresponding Author:

Sazid Nur Ratul, BSc, MSc
School of Computing, Faculty of Science, Forestry and Technology
University of Eastern Finland
Yliopistokatu 2
Joensuu, North Karelia 80100
Finland
Phone: 358 294451111
Email: contact@sazidnur.com

Abstract

Background: Many diseases often develop through sequences of related conditions over time. Identifying how diagnoses occur over time may help detect early risk signals before severe outcomes arise. Clinically significant patterns are limited because many population-level studies focus on disease co-occurrence rather than the temporal order of diagnoses.

Objective: The study aimed to identify and validate temporal disease associations using frequent pattern mining and statistical validation techniques in a nationwide patient-record database.

Methods: We analyzed health records of 3,987,382 Finnish patients. The records were transformed into patient timelines by using the date on which the first record was entered into the database. To identify the most common patterns for each unique disease, we applied the frequent pattern growth algorithm using a minimum support of 5% of patients with the target disease and at least 5 patients. To validate each pattern, we applied a combination of relative risk (RR), a 95% CI, and relative width to measure precision.

Results: We identified several clinically interpretable temporal disease connections. For example, acute kidney failure was most strongly preceded by unspecified kidney failure (RR 15.13, 95% CI 14.24-16.07), chronic kidney disease (RR 11.02, 95% CI 10.53-11.54), and sepsis (RR 9.76, 95% CI 9.27-10.27). It was also grouped with heart failure–related combinations (RR 10.76, 95% CI 10.27-11.27). Diabetic foot ulcer with type 1 or type 2 diabetes was strongly associated with osteomyelitis, with an RR of 157.02 (95% CI 137.03-179.93) for type 1 and 84.84 (95% CI 77.07-93.39) for type 2. At the block level, cerebrovascular diseases were linked to hypertension (RR 2.47; 95% CI 2.44-2.49), atherosclerosis (RR 2.52; 95% CI 2.47-2.58), and dementia (RR 2.96; 95% CI 2.90-3.01). Drug poisoning patterns were also connected to psychiatric diagnoses, including mood disorders (RR 7.24, 95% CI 7.03-7.46) and combinations of alcohol and mood disorders (RR 18.25, 95% CI 17.51-19.01). The generated patterns and statistical measures are publicly available in a web interface for research purposes.

Conclusions: Frequent pattern mining, when integrated with RR/CI and precision filtering, produces clinically interpretable temporal connections that can support hypothesis generation. External validation with other datasets and cohorts is essential.

JMIR Med Inform 2026;14:e94284; doi: 10.2196/94284

Keywords: disease progression; disease pattern; medical data analysis; frequent pattern growth; FP-growth; relative risk; CI; relative width; *International Statistical Classification of Diseases, Tenth Revision*; *ICD-10* codes

Introduction

Chronic diseases are the leading cause of death globally, with an estimated 41 million fatalities annually [1]. A significant percentage of this burden affects working-age individuals, causing substantial social and economic limitations on health systems [2]. The World Health Organization (WHO) reports that 82% of premature deaths are caused by noncommunicable diseases, such as cardiovascular disease, cancer, chronic respiratory disease, and diabetes, especially in many low- and middle-income countries [3]. Understanding how diseases develop over time and which conditions tend to appear first may help identify people at higher risk earlier. We analyzed health records for 3.9 million individuals in Finland from the Health Care Register administered by the Finnish Institute for Health and Welfare. We used these periodic records to analyze the diagnoses that usually precede specific target diseases and assessed relationships that may require clinical consideration.

We applied a frequent pattern mining algorithm to identify diagnosis patterns that commonly precede a given target disease. For each target disease, we constructed patient timelines from all diagnoses recorded before the first occurrence of that disease. This method ensures that we only find diseases that happened before the patient was diagnosed with the target disease. We calculated the relative risk (RR) to assess how much more likely patients with a certain pattern were to develop the target disease than those without the pattern [4].

To estimate the reliability of RR for each disease pattern, we also estimated the 95% CI for each RR using the Wald method [5], and we assessed the precision of each CI with the relative width (RW) [6]; only patterns with RR greater than 1, CIs greater than 1, and RW values indicating high to moderate reliability were selected for further clinical review. We examined the frequency of patterns at 2 levels of *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) codes: at a single level (representing individual diseases, eg, E10) and at a higher level of aggregation using larger code blocks (grouping related conditions, eg, E10-E14). The use of both methods allowed us to compare highly specific disease associations and more general disease groupings.

We identified multiple patterns involving chronic kidney disease, sepsis, and heart failure that demonstrated high RRs with relatively narrow CIs among patients diagnosed with acute kidney failure. We also observed that codes representing diabetes-related conditions demonstrated very strong associations among patients diagnosed with retinal disorders. Similarly, we identified that patterns involving diabetic foot ulcers and peripheral arterial disease were associated with osteomyelitis. The findings we identified are consistent with existing clinical pathways, provide evidence-based confirmation of established clinical knowledge, and may guide hypotheses for further research.

Methods

We followed a step-by-step approach, from collecting patient data to generating patterns and calculating statistical measures.

Data Collection

We collected data from the electronic health records of 3,987,382 adults from whole Finland, covering all recorded care contacts from 2015 to 2018 [7]. The total number of records in the dataset is 60,276,010. Each record includes an *ICD-10* code describing the diagnosis and a sequence time (in minutes) calculated from the original visit time stamps. The sequence represents when the diagnosis first appears for a patient, measured from the patient's first recorded visit in the dataset. This time shows when a diagnosis was first recorded between 2015 and 2018, not when the disease actually began. If a patient already had the condition before 2015, the first record simply reflects their first visit in this period. The dataset contains 2 levels of coding: level 1 specifies a particular diagnosis (eg, E10), whereas level 2 represents a larger group of diagnoses based on the *ICD-10* code block (eg, E10-E14 for all diabetes mellitus types).

Data Preprocessing

All records with *ICD-10* codes starting with V, W, X, Y, Z, R, or U were eliminated from the analysis, as these codes describe external causes of injury, health status factors, or symptoms rather than actual diagnoses. A chronologically ordered list of diseases diagnosed for each patient was created using both level 1 and level 2 *ICD-10* codes.

To conduct data mining for a specific target disease, the subset of patients who had been diagnosed with the target disease and whose disease history precedes the first occurrence of the target disease was isolated, and all diseases appearing prior to the first occurrence of the target disease were also extracted. This method preserves the temporal relationship between preceding diagnoses and the target disease and thus allows for the detection of preceding disease patterns rather than simply detecting diseases that occur simultaneously. After preprocessing, we found 1473 unique level 1 diseases and 206 unique level 2 disease blocks.

Pattern Mining

To look for patterns associated with disease, we used the frequent pattern growth (FP-Growth) algorithm [8], which uses a compact tree-based data structure called a frequent pattern tree (FP-Tree) to find the most frequently occurring patterns. First, the algorithm finds the frequency of occurrence of items in the dataset to establish a global ordering of the items. Second, it creates the FP-Tree by adding transactions to the tree based on the global ordering of the items. Third, the FP-Growth algorithm searches for the patterns recursively through the FP-Tree in a depth-first manner to locate all patterns that exceed the established support threshold.

All patterns found by the algorithm were ordered from highest to lowest frequency. We set the minimum support to 5% and required at least 5 patients with the target condition. We balanced the quality of the mined patterns and computational efficiency using this threshold. This mining process produced single and multidisease patterns, which have one-to-one or complex multidisease relationships.

Apriori and Eclat (equivalence class clustering and bottom-up lattice traversal) are other common algorithms for mining unordered itemsets, and they find the same frequent patterns as FP-Growth. Apriori requires repeated database scans for its candidate generation method, which is computationally expensive for datasets like ours. We repeated the mining with Eclat, which produced identical patterns and support counts. Therefore, the choice between FP-Growth and Eclat does not affect our findings. We do not compare with sequential pattern mining methods, which also recover the order of the diagnoses within a pattern. Our patterns are only sets of diagnoses recorded before the target disease, and we do not report any order within them.

Statistical Analysis

To evaluate the association of disease patterns and the target conditions, we used RR [4,6]. RR is a measure of the ratio of the incidence of the target disease among individuals with a particular disease pattern to the incidence of the target disease among those without the disease pattern. We chose RR because it is easy to interpret as a direct comparison of risk between groups, and it does not overstate impact when outcomes are common, unlike the odds ratio. It is calculated using a standard 2×2 contingency table.

$$\mathrm{RR} = [a/(a+b)]/[c/(c+d)] \quad (1)$$

Here, a refers to patients with *both* the disease pattern and the target disease, b refers to patients with the *disease pattern* but without the target disease, c refers to patients without the disease pattern but with the *target disease*, and d refers to patients *without* either the disease pattern or the target disease.

Thus, $a + b$ represents the total number of patients with the disease pattern, and $c + d$ represents the total number of patients without the disease pattern.

We focused only on RRs that are greater than 1, since we are searching for how much the risk is elevated. We also calculated the 95% CI for each RR using the Wald method [5,9,10] to measure the uncertainty around each RR estimate.

$$\mathrm{CI} = \exp\left[\ln(\mathrm{RR}) \pm z \times \mathrm{SE}(\ln(\mathrm{RR}))\right] \quad (2)$$

$$\mathrm{SE}(\ln(\mathrm{RR})) = \sqrt{1/a - 1/(a+b) + 1/c - 1/(c+d)} \quad (3)$$

Here, exp refers to the exponential function (with base e, the Euler number), ln refers to the natural logarithm, and z=1.96 represents the z score for a 95% confidence level.

The 95% CI gives the range of RR values that are compatible with the data. When the lower bound is greater than 1, the RR is statistically significant at the 5% level.

To evaluate the precision of the RR estimate, we calculated the RW of each 95% CI by dividing the width of the CI by the RR [11,12]. We used RW to identify patterns with wide, imprecise CIs and to prioritize patterns with more stable RR estimates. RW was calculated as follows:

$$\mathrm{RW} = (\text{upper CI} - \text{lower CI})/\mathrm{RR} \quad (4)$$

Here, the upper CI refers to the upper bound of the 95% CI, the lower CI refers to the lower bound of the 95% CI, and RR refers to the relative risk value.

We next classify RW-values into 3 groups based on how reliable they were: *very reliable* (RW ≤0.5), *moderately reliable* (0.5<RW≤1.0), and *not reliable* (RW>1.0). These categories allow us to differentiate between patterns with strong and weak evidence of correlation.

The combination of RR, CI, and RW provides a complete framework for evaluating the associations of disease patterns and target conditions. For example, if RR is 2.5, the 95% CI ranges from 2.3 to 2.8, and RW of 0.2, the disease pattern is characterized as high risk, statistically significant, and highly reliable. Conversely, a combination of RR is 2.5, the 95% CI ranges from 1.2 to 5.7, and RW is 1.8 implies high risk with high uncertainty and should be interpreted with caution.

Analysis Scope

We have conducted pattern mining and all statistical analyses at both the level of *ICD-10* code (single level) and block levels as 2 separate batches. The dual-level of analysis permitted us to investigate both detailed associations between individual diagnoses and larger patterns across related disease groups.

Result Presentation

The results were stored in an SQLite (Hipp, Wyrick & Company, Inc) database and then shown as an interactive web interface built with HTML, CSS, JavaScript, and PHP. The pattern for every disease was visualized with patient counts, support percentages, RR values, CIs, and RWs in tabular form. To improve understanding, we also added color-coding based on the conditions of the cells, as follows: RR values greater than 1 were colored red (higher values are darker). Cells displaying a CI in which both bounds were greater than 1 were colored green (statistically significant). The RW cells were colored using a green-yellow-red gradient showing high, moderate, and low reliability levels. Complete results for all analyzed diseases are publicly available [13].

Clinical Validation

The study validated the identified patterns based on the existing scientific research to confirm whether they are biologically plausible. This was done to help determine which identified patterns may be indicative of real disease progression vs those that may simply represent the association by

way of coding, timing of diagnosis, or other confounders. The study used a combination of how much clinical support was found for each pattern in the literature to determine where to focus the analysis for each condition; if a pattern lacked a clear clinical basis, it was noted as requiring additional research.

Ethical Considerations

When only deidentified, register-based data are used and study participants are not contacted, ethical approval and informed consent from individual patients are generally not required under the Finnish Data Protection Act (1050/2018) [14] and the Act on the Secondary Use of Health and Social Data (552/2019) [15]. All data were deidentified, and the analyses were conducted in accordance with the relevant guidelines and regulations, including the Declaration of Helsinki.

Results

Overview

Next, we provide comprehensive results for selected diseases at both single-code and block levels. We analyzed patterns that showed strong associations (high RR), narrow CIs, and low RWs, which are statistically consistent with expectations based on current clinical practice. A summary of these patterns is provided in Table 1.

Table 1. Overview of a few analyzed diseases and their key characteristics.

Code	Disease	Patients	RR[a] range	Primary associated conditions
H36	Retinal disorders	21,123	2.78-48.38	Diabetes and kidney disease
N17	Acute kidney failure	19,111	4.05-15.13	Sepsis, heart failure, and chronic kidney disease
M86	Osteomyelitis	3733	4.80-157.02	Diabetic foot ulcers and peripheral arterial disease
E85	Amyloidosis	1408	2.10-1492.12	Kidney, cardiac, and neurological disorders
D51	Vitamin B12 deficiency	10,339	2.21-5.88	Anemia, gastritis, and thyroid disease
I95	Hypotension	23,672	2.06-10.05	Parkinson disease, dementia, and cardiac disorders
T36	Antibiotic poisoning	16,689	2.60-27.23	Mental health disorders and substance use
J60-J69	Lung diseases (external agents)	11,105	2.25-10.18	Pleural disease, aspiration, and occupational exposure
I60-I69	Cerebrovascular diseases	137,544	1.79-2.96	Hypertension, cardiac disease, and dementia
N17-N19	Renal failure (block)	64,807	2.25-16.90	Kidney disease spectrum, cardiac, and diabetes
T36-T50	Drug poisoning (block)	18,679	2.63-18.25	Mental health, substance use, and injury

[a]RR: relative risk.

Acute Kidney Failure (N17)

We identified 19,111 patients with acute kidney failure. Analysis revealed a multifaceted clinical picture involving kidney disease progression, infection, cardiorenal interactions, and metabolic disturbances. Table 2 presents selected high-impact patterns.

Table 2. Temporal patterns associated with acute kidney failure (N17).

Preceding pattern	Patterns with N17	Total with pattern	Relative risk (95% CI)	Relative width
N19	1038	14,520	15.13 (14.24-16.07)	0.121
N18	1910	38,282	11.02 (10.53-11.54)	0.092
E11 and I50	1894	38,839	10.76 (10.27-11.27)	0.093
A41	1525	33,817	9.76 (9.27-10.27)	0.102
I10 and N18	1162	25,347	9.74 (9.19-10.32)	0.116
E87	1628	38,818	9.12 (8.67-9.58)	0.100
I48 and I50	2692	73,308	8.42 (8.09-8.77)	0.080
I50 and J18	1377	38,797	7.60 (7.21-8.03)	0.108
A49	1869	62,240	6.58 (6.27-6.89)	0.094
E11 and J18	1057	35,383	6.29 (5.92-6.69)	0.122
D64	1456	52,506	5.95 (5.64-6.27)	0.105
I70	1348	52,354	5.49 (5.20-5.80)	0.109
N10	1814	78,325	5.03 (4.80-5.28)	0.096
E11	5248	323,580	4.11 (3.98-4.24)	0.063

Kidney disease represented the most common pattern within the observed trends; the RRs were 15.13 (95% CI 14.24-16.07) for unspecified kidney failure (N19) and 11.02 (95% CI 10.53-11.54) for chronic kidney disease (N18). Both conditions can be part of a continuum in which acute events occur on top of existing chronic disease or are simply classified as unspecified kidney failure upon hospital admission. A higher level of risk was seen when hypertension

occurred in combination with chronic kidney disease (I10 and N18); the RR was 9.74 (95% CI 9.19-10.32).

Sepsis (A41) had an RR of 9.76 (95% CI 9.27-10.27), and bacterial infections (A49) had an RR of 6.58 (95% CI 6.27-6.89). Pneumonia (J18) also appeared in several combinations, including with heart failure (I50, J18) at an RR of 7.60 (95% CI 7.21-8.03) and with type 2 diabetes (E11 and J18) at an RR of 6.29 (95% CI 5.92-6.69). This is consistent with clinical literature reporting severe infection and sepsis as major causes of acute kidney injury [16,17].

Cardiorenal clustering was observed. Diabetic patients with heart failure (E11 and I50) had an RR of 10.76 (95% CI 10.27-11.27). Patients having both heart failure and atrial fibrillation (I48 and I50) had an RR of 8.42 (95% CI 8.09-8.77). These stacked patterns reflect a decline in cardiac output and resultant venous congestion, leading to decreased renal perfusion [18-20]. There is evidence that diabetic and hypertensive patients are at greater risk for developing acute kidney injury than nondiabetic, nonhypertensive patients with heart failure [21].

Electrolyte disorders (E87) have an RR of 9.12 (95% CI 8.67-9.58) and can both cause and result from kidney damage [22]. An RR of 5.95 (95% CI 5.64-6.27) was found for anemia (D64), 5.49 (95% CI 5.20-5.80) for atherosclerosis (I70), and 5.03 (95% CI 4.80-5.28) for pyelonephritis (N10).

Osteomyelitis (M86)

We identified 3733 patients with osteomyelitis. Despite the relatively small cohort, the analysis revealed patterns with very high RR values and narrow CIs. These patterns revealed diabetic foot pathways and pressure ulcer complications as dominant themes. Table 3 shows selected associations.

Table 3. Temporal patterns associated with osteomyelitis (M86).

Preceding pattern	Patterns with M86	Total with pattern	Relative risk (95% CI)	Relative width
E10 and L97	188	1296	157.02 (137.03-179.93)	0.273
E11 and L97	442	6067	84.84 (77.07-93.39)	0.192
E11, I70, and L97	204	2989	74.18 (64.71-85.03)	0.274
A46 and L97	261	4198	68.66 (60.77-77.58)	0.245
E11 and S91	210	3607	63.38 (55.35-72.56)	0.271
L89	291	8328	38.88 (34.57-43.74)	0.236
E11 and I70	488	19,467	29.50 (26.85-32.41)	0.189
E11 and L03	233	9907	25.73 (22.57-29.33)	0.263
A46 and E11	411	19,538	24.18 (21.85-26.77)	0.203
L98	217	11,548	20.45 (17.85-23.43)	0.273
N08	186	10,791	18.60 (16.07-21.53)	0.293
A41	302	35,323	9.48 (8.43-10.66)	0.235
E11	1386	324,619	6.39 (5.98-6.83)	0.133

The diabetic foot pathway represented most of our findings. The combination of type 1 diabetes (E10) and lower-limb ulcer (L97) had the largest RR of 157.02 (95% CI 137.03-179.93). Type 2 diabetes with a lower-limb ulcer (E11 and L97) had an RR of 84.84 (95% CI 77.07-93.39). The addition of atherosclerosis (I70) resulted in 3 combinations with a high RR value of 74.18 (95% CI 64.71-85.03) for E11, I70, and L97. Type 2 diabetes with an open wound of the ankle/foot (E11 and S91) had an RR of 63.38 (95% CI 55.35-72.56). The diabetic foot pathway appears to be the classic medical pathway of progression from lower-limb ulcer to deep tissue infection to bone infection [23].

Outside of diabetes, both pressure ulcers and chronic skin ulcers contributed substantially to the risk of developing osteomyelitis [24,25]. Pressure ulcers (L89) had an RR of 38.88 (95% CI 34.57-43.74), and chronic skin ulcers (L98) produced an RR of 20.45 (95% CI 17.85-23.43). Soft tissue surrounding infections were also found through the presence of cellulitis (L03), with an RR of 10.84 (95% CI 9.88-11.88). Erysipelas (A46) was present in numerous high-risk combinations and produced an RR of 68.66 (95% CI 60.77-77.58) when combined with ulcers.

Systemic infection markers included sepsis (A41) with an RR of 9.48 (95% CI 8.43-10.66). The cardiovascular and metabolic comorbidities included in our study demonstrated a pattern in which type 2 diabetes (E11) produced an RR of 6.39 (95% CI 5.98-6.83), atherosclerosis (I70), in various combinations, produced RRs between 17.05 (95% CI 15.49-18.77) and 29.50 (95% CI 26.85-32.41), and glomerular disorders (N08) produced an RR of 18.60 (95% CI 16.07-21.53), likely indicating diabetic nephropathy in this patient population.

Cerebrovascular Diseases (I60-I69 Block)

At the block level, we analyzed 137,544 patients with cerebrovascular disease codes (I60-I69). The grouping captures the full spectrum from acute stroke to chronic sequelae. Table 4 summarizes our most important findings. In this table, preceding patterns represent their entire block; for example, F00 refers to the entire F00 to F09 range.

Table 4. Temporal patterns for cerebrovascular disease (I60-I69).

Preceding pattern	Patterns with I60-I69	Total with pattern	Relative risk (95% CI)	Relative width
F00	10,444	103,785	2.96 (2.90-3.01)	0.038
G30	8994	96,818	2.70 (2.65-2.76)	0.041
I30	34,285	444,657	2.53 (2.50-2.56)	0.024
I70	7327	83,733	2.52 (2.47-2.58)	0.045
I10	48,922	702,033	2.47 (2.44-2.49)	0.021
E10 and I10	14,183	200,671	2.08 (2.05-2.12)	0.034
E70 and I10	14,529	212,142	2.02 (1.99-2.05)	0.033
G40	21,269	355,382	1.79 (1.77-1.82)	0.028

The risk of cardiovascular disease is strong and consistent in our data. For example, hypertension (I10) had an RR of 2.47 (95% CI 2.44-2.49) and a very narrow CI. The RR of 2.53 (95% CI 2.50-2.56) observed for other heart disease (I30) was similar to that seen for atherosclerosis (RR 2.52, 95% CI 2.47-2.58). The observed RRs are in agreement with the considerable body of epidemiological research, which demonstrates that hypertension and coronary artery disease account for a substantial portion of the increased risk of stroke; atrial fibrillation has been identified as a particular mechanism leading to embolic strokes [26-28].

Neurodegenerative patterns were prominent in the data. RRs of 2.96 (95% CI 2.90-3.01) for dementia (F00) and 2.70 (95% CI 2.65-2.76) for Alzheimer disease and other degenerative conditions (G30) indicate considerable overlap between vascular damage and the neurodegenerative consequences of cerebrovascular events. As mentioned above, poststroke dementia has been identified in several systematic reviews [29].

Diabetes appeared in several different combinations and also co-occurred with hypertension (E10 and I10) at an RR of 2.08 (95% CI 2.05-2.12). Lipid disorders (E70) were also associated with an increased RR of 2.02 (95% CI 1.99-2.05) when combined with hypertension.

Additional Conditions

In addition to the detailed example blocks listed above, we reviewed a number of other medical conditions that illustrated unique patterns of association. We begin with single-disease codes first and then continue with block-level conditions.

Retinal disorders (H36) were highly associated with type 1 diabetes mellitus (RR 48.38, 95% CI 47.09-49.70) and type 2 diabetes mellitus (RR 6.81, 95% CI 6.63-7.00), as expected with diabetic retinopathy. Similarly, glomerular disorders (N08) had an RR of 21.56 (95% CI 20.36-22.84), as they are also associated with common pathologies of the microvasculature of the kidney and retina [30].

Amyloidosis (E85) was identified as one of the most highly ranked RRs in the entire dataset; cerebral amyloid angiopathy (I68) had an RR of 1492.12 (95% CI 1294.16-1720.36), largely due to both the *ICD-10* coding guidelines indicating that individuals should be coded for amyloidosis first when amyloidosis is present and the true pathological overlap between the two. Additionally, ocular involvement (H19) had an RR of 58.76 (95% CI 47.45-72.76), glomerular disorders (N08) had an RR of 31.97 (95% CI 26.47-38.63), and rheumatoid arthritis (M05) had an RR of 13.35 (95% CI 11.31-15.78); these reflected pathways related to renal amyloidosis and AA amyloidosis [31]. The cardiac involvement was evident in heart failure (I50) with an RR of 3.74 (95% CI 3.17-4.40) and in atrial fibrillation (I48) with an RR of 2.10 (95% CI 1.79-2.46).

Vitamin B12 deficiency (D51) was grouped with other anemias (D64 at an RR of 5.88, 95% CI 5.47-6.32), gastritis (K29 at an RR of 5.79, 95% CI 5.31-6.32), and type 2 diabetes (E11 at an RR of 2.95, 95% CI 2.81-3.09), the latter reflecting metformin-associated B12 depletion. Hypothyroidism (E03) showed an RR of 2.82 (95% CI 2.64-3.02), consistent with autoimmune clustering patterns [32,33].

Hypotension (I95) was shown to have strong neurodegenerative relationships with Parkinson disease (G20 at an RR of 10.05, 95% CI 9.51-10.61), dementia (G30 at an RR of 4.91, 95% CI 4.72-5.11; F00 at an RR of 4.84, 95% CI 4.61-5.09), and atrial fibrillation (I48 at an RR of 4.44, 95% CI 4.30-4.57), likely reflecting underlying autonomic dysfunction and/or the influence of medications [34-37].

Both antibiotic poisoning (T36) and the larger group of drug poisoning (T36-T50) demonstrated patterns of mental health and substance abuse. Specifically at the single-code level, psychoactive substance use (F19) had an RR of 27.23 (95% CI 25.74-28.80), personality disorders (F60) had an RR of 14.62 (95% CI 13.87-15.40), and alcohol use (F10) had an RR of 12.29 (95% CI 11.86-12.74). RR values for mood disorders (F30) at the block level were 7.24 (95% CI 7.03-7.46) and for schizophrenia (F20) were 6.26 (95% CI 5.96-6.57). Patterns combining alcohol with mood disorders resulted in an RR of 18.25 (95% CI 17.51-19.01), which matched hospital records regarding intentional self-poisoning [38-40].

Block-level analysis of lung diseases due to external agents (J60-J69) demonstrated pleural effusion (J90) with an RR of 10.18 (95% CI 9.41-11.01). This may reflect either a complication of pneumoconiosis leading to pleural effusion or a shared consequence of the same occupational exposure, such as dust, causing both lung and pleural disease. Pneumonia (J09) had an RR of 5.70 (95% CI 5.47-5.94), and the neurologic pattern suggested a higher likelihood of aspiration with an RR of 4.25 (95% CI 3.90-4.62) for Parkinson disease (G20) and an RR of 3.77 (95% CI 3.54-4.01) for dementia (G30).

Renal failure at the block level (N17-N19) reinforced the findings observed at the single-code level, with glomerular disease (N00) having an RR of 16.90 (95% CI 16.40-17.41) and combinations of cardiovascular factors resulting in RR values greater than 5.

Understudied Patterns

Some patterns were also observed that did not clearly reflect temporal association (Table 5). Upon closer inspection, many of these appear to be symptom-disease associations. For example, other functional intestinal disorders (K59), such as constipation or diarrhea, may occur secondary to malignant neoplasm of the rectosigmoid junction (C19), and excessive, frequent, or irregular menstruation (N92) may be a symptom of salpingitis or oophoritis (N70). Other co-occurrences likely reflect shared risk factors rather than direct causality, such as benign prostatic hyperplasia (N40) and degenerative diseases of the basal ganglia (G23), both associated with older male populations, or osteoarthritis of different sites (M17—knee, M18—first carpometacarpal joint), which tend to appear in the same patients. More intriguing are the co-occurrences of migraine (G43) with uterine or pelvic inflammatory diseases (N71 and N73), in which hormonal changes could predispose to migraine, and inflammation could in turn affect hormones; in these cases, migraine is more likely a consequence than a cause.

Table 5. Examples of associations that do not clearly reflect temporal associations in single-level analysis.

Preceding pattern	Target disease	Target with pattern	Total people with pattern	Relative risk (95% CI)
G43	N71	G43, N71=608	76,803	3.04 (2.80-3.30)
K59	C19	K59, C19=80	78,017	2.99 (2.39-3.75)
N92	N70	N92, N70=192	55,608	5.35 (4.62-6.19)
N40	G23	N40, G23=119	115,796	3.27 (2.71-3.95)
G43	N73	G43, N73=252	77,075	2.75 (2.43-3.13)
N83	N70	N83, N70=503	30,930	29.04 (26.37-31.99)
M17	M18	M17, M18=999	207,176	2.12 (1.98-2.26)
N92	L68	N92, L68=128	55,645	6.38 (5.32-7.65)

Sensitivity Analyses

The minimum support threshold determines which patterns are reported, so we repeated the mining at 1%, 2%, 5%, and 10%, holding the minimum of 5 patients fixed. The threshold is a reporting criterion only and does not enter the RR, CI, or RW. The estimates for any given pattern are therefore identical at every threshold, and only the number of retained patterns changes. Among the 7 single-code target diseases reported in Table 1, lowering the threshold from 10% to 1% increased the retained set by a factor of 47 to 78, for example, from 15 to 1171 patterns for acute kidney failure. These additional patterns are mostly combinations of associations already represented at 5%, or weaker associations with wider CIs, and we therefore retained 5% for the main analysis. Table S1 in Multimedia Appendix 1 gives the counts at each threshold.

Because the data begin in 2015, a target disease recorded at a patient's first visit may already have been present before the study window. We therefore repeated the analysis after excluding these patients. Between 8.6% and 25.9% of each target cohort were excluded, ranging from 2040 of 23,672 patients for hypotension to 364 of 1408 patients for amyloidosis. No reported pattern fell below the support threshold, and none lost statistical significance; RRs increased slightly, as expected, because the excluded patients had the target disease but little or no preceding history. Table S2 in Multimedia Appendix 1 reports the effects for the same 7 target diseases. Since these patients cannot be distinguished from those who were genuinely newly diagnosed, they were retained in the main analysis, which is therefore the more conservative choice.

Discussion

In the following subsections, we discuss our findings, limitations, and future possibilities, and we compare our work with other studies.

Principal Findings

This study identified temporal associations among chronic disease diagnoses in a large cohort of Finnish people. Millions of patient records and thousands of diagnosis codes were included in the analysis, showing that the statistical associations were very consistent with those found in the literature. Most findings corresponded to well-established clinical associations.

This study demonstrated that it is possible to use pattern mining methods at a large scale using real-world health care data. Both expected and some unanticipated patterns were identified. Additionally, a statistical validation framework was used to help distinguish meaningful associations from random co-occurrences.

Comparison With Related Work

Population-based studies using pattern mining approaches have used either network analysis or association rule learning on datasets consisting of thousands to millions of patients. In contrast, our approach combined the use of FP-Growth with both RR validation and RW filtering to provide both pattern discovery and robust statistical support. Temporal sequences were also used in place of co-occurrence windows to distinguish between concurrent conditions.

Choudhary and Fränti [41] used the same data as in our study to build a disease network using a network model, achieving an area under the curve value of 0.65 for predicting the next disease and 0.68 for a ranked list of future diseases. The key difference between their work and ours is that they take a patient's medical history as input and predict possible future disease. In contrast, our study takes a target disease and analyzes potential patterns leading to that target. They also guided their identification of frequent disease by age group and sex, which is missing in our study. Despite the methodological differences, they reported disease progressions that align with our findings, such as dementia with neurodegenerative disorders, stress-related disorders with mood disorders, and ocular/retinal pathways involving diabetes and hypertension.

Uddin et al [42] conducted longitudinal analyses of 29,280 patients over 24 years (1995-2018), validating patterns using *t* tests. However, they used a smaller cohort, which prevented them from conducting analyses of rare diseases. Su et al [43] identified asthma-cardiovascular associations, and Rocca and Anjum [44] mapped multiple chronic disease patterns. Neither of these studies used temporal ordering nor computed RR at scale.

Xiao et al [45] model disease progression in continuous time using a hypergraph neural ordinary differential equation. Lachinov et al [46] fit a partial differential equation to imaging data to model disease progression. Hu et al [47] combine graph neural networks with a transformer to model disease progression over time. These methods train a model to predict how a single patient will progress. Our study answers a different question. We describe, for a whole population, which diagnoses come before a target disease, and we report RR, CI, and RW for each pattern, without training a model. The 2 approaches are complementary because predicting for 1 patient and describing a population serve different needs.

Hidden Factors and Interpretation Limits

Temporal associations are not necessarily a direct measure of causality. Many factors may mediate disease associations, yet these are often not captured in diagnostic coding data. For instance, heart failure can reduce appetite, which leads to anemia [48]. Although our results indicate that heart failure precedes anemia, decreased appetite is not observable through the diagnostic coding. Consequently, the observed temporal association does not establish direct causality, as key mediating factors may be missing from the data.

The timing of a patient's diagnosis also plays a role in determining the results. A diagnosis date represents the time when a doctor records the diagnosis, not the actual time the disease begins. Certain chronic diseases take years to develop prior to the detection of symptoms. Screening tests to identify chronic diseases may provide evidence that reverses the apparent relationship between diseases. For example, although diabetes develops years prior to the identification of kidney disease, the evaluation of kidney disease may result in the identification of diabetes as well.

Therefore, these limitations suggest that the relationships identified in this study should be used to generate hypotheses rather than draw causal conclusions. Additionally, the temporal sequence of events may suggest possible avenues for future research using different methods.

Methodological Considerations

The design of the study made several methodological choices that affected the results. The minimum support threshold was set at 5% of the patients with the target disease, with a minimum of 5 patients. This threshold allowed us to identify enough patterns while keeping the analysis manageable. Lower thresholds would have identified many additional patterns, which would have been much harder to interpret.

The requirement of at least 5 patients per pattern avoided rare cases, which increased the stability of the analysis. We calculated RRs without adjustment for age, sex, socioeconomic status, or the number of prior diagnoses. These factors are expected to influence the strength of the observed associations. For example, the association between dementia and cerebrovascular disease reported in Table 4 (RR 2.96, 95% CI 2.90-3.01) is likely influenced by age, since both conditions become more common in older patients. However, the data represent real-world health care data and therefore reflect the actual burden of co-occurrence of diseases in the health care system, including the contribution of factors such as age, sex, and socioeconomic status. All timing in the analysis was based on the date of diagnosis, not the actual onset of disease. The mining produced 39,826 patterns (33,467 at the single-code level and 6359 at the block level) across all target diseases, and their CIs were not corrected for multiple comparisons. Some of the weaker associations may therefore be false positives. However, 23,536 of the 39,826 patterns had a 95% CI above 1, and 19,592 of these 23,536 (83.2%) remained significant after Bonferroni correction (α=.05/39,826) [49], including all patterns discussed in the Results. The other 3944 of these 23,536 (16.8%) all had an RR below 2 or an RW above 0.5, indicating weak or less reliable associations. Statistical significance alone does not establish clinical relevance, and the patterns should therefore be interpreted as hypotheses for further study.

Limitations

A major advantage of this study was its large dataset size, which provided precise estimates. The study used both specific diseases and higher-level blocks of diseases to allow different levels of observation. However, there are also some limitations to this study. First, all data were derived from a single country, Finland. The health care system and population demographics may differ considerably from one country to another. The second limitation is that the coding of diagnoses may not have been precise, because there are variations in how the codes are used in practice. Some diseases may have been underdiagnosed, and some misdiagnosed. Third, the study used only overall statistics and did not use any clinical information about individual patients. Fourth, the data covered only 2015 to 2018, so earlier diagnoses were not visible. Conditions that already existed before this period

were simply dated to the patient's first visit within it, which may have shifted their apparent order.

Future Work

Possible areas of future work include validating the observed patterns using data from other countries. Laboratory test results and medication data could provide further insight into patterns, although it is not trivial how they could be incorporated into the population-level analyses. Further study could consider age, sex, and socioeconomic factors in the analysis. The volume of data is sufficient for even further subsampling based on more fine-tuning factors. Our analysis identifies which diagnoses precede a target disease but not the order of the diagnoses within a pattern. Sequential pattern mining algorithms such as PrefixSpan (prefix-projected sequential pattern mining) and cSPADE (constrained sequential pattern discovery using equivalence classes), along with temporal rule mining methods such as multistream dependency detection (MSDD), can recover that order. This can be another direction for future work. Finally, predictive machine learning models could be developed based on these findings and applied to a new cohort of patients. Once their predictive performance has been formally evaluated, these models could be integrated into clinical decision support systems to alert health care professionals when specific temporal diagnostic patterns are associated with an increased risk of subsequent severe disease, thereby enabling timely preventive or therapeutic interventions.

Conclusions

Our study used Finnish health records to find disease patterns that happen before a target disease. We mined those patterns using the FP-Growth algorithm and measured their risk by RR, 95% CI, and RW. The patterns match with existing studies, but our study provides added value over existing studies due to its data size. The patterns are also clinically aligned with known paths, such as chronic/unspecified kidney failure (N18/N19) and sepsis (A41), which were strongly seen before acute kidney failure (N17). The cardiorenal combinations, such as heart failure and diabetes, are also clustered with acute kidney failure. Similarly, osteomyelitis (M86) was also strongly associated with diabetes with lower-limb ulcer (E10/E11+L97). In addition, our study provides a public web interface for accessing the generated patterns and statistical measures for research purposes.

Generative AI Use

ChatGPT (OpenAI) and Claude (Anthropic) were used for language editing of the manuscript and to assist with writing code for data visualization and the web interface. All AI-assisted output was reviewed, tested, and edited by the authors, who take full responsibility for the accuracy, originality, and integrity of the content, including references and citations.

Funding

The acquisition and initial processing of the data were funded by the Strategic Research Council at the Academy of Finland (grant numbers 312703, 312706, 336325, and 336330).

Data Availability

The data used in this study come from the registers of the Finnish Institute for Health and Welfare. These datasets are not publicly available because their use is restricted by Finnish data protection legislation. The data were accessed under a granted license. For these analyses, the data were anonymized. The resulting patterns are publicly available for research purposes [13].

Authors' Contributions

Conceptualization: SNR (lead), PF (equal)
Data curation: SNR (lead), JPT (equal), TL (supporting)
Formal analysis: SNR
Funding acquisition: TL
Investigation: SNR
Methodology: SNR (lead), PF (supporting)
Resources: JPT
Software: SNR
Validation: TL (lead), JPT (supporting), PF (supporting)
Visualization: SNR
Writing - original draft: SNR
Writing - review & editing: PF (lead), JPT (equal), TL (equal), SNR (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1

Sensitivity analyses for support-threshold selection and registry-entry exclusion effects on relative risk estimates.
[DOCX File (Microsoft Word File), 25 KB-Multimedia Appendix 1]

References

1. Noncommunicable diseases. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases [Accessed 2025-03-08]
2. Woolf SH, Schoomaker H. Life expectancy and mortality rates in the United States, 1959-2017. JAMA. Nov 26, 2019;322(20):1996-2016. [doi: 10.1001/jama.2019.16932] [Medline: 31769830]
3. Preventing noncommunicable diseases. World Health Organization. 2024. URL: https://www.who.int/activities/preventing-noncommunicable-diseases [Accessed 2025-03-08]
4. Zhang J, Yu KF. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. JAMA. Nov 18, 1998;280(19):1690-1691. [doi: 10.1001/jama.280.19.1690] [Medline: 9832001]
5. Katz D, Baptista J, Azen SP, Pike MC. Obtaining confidence intervals for the risk ratio in cohort studies. Biometrics. Sep 1978;34(3):469-474. [doi: 10.2307/2530610]
6. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3rd ed. Lippincott Williams & Wilkins; 2008. ISBN: 9780781755641
7. National treatment notification register (hilmo). Finnish Institute for Health and Welfare (THL). URL: https://aineistokatalogi.fi/catalog/studies/7567e45d-72b7-428b-be9e-510440336edf [Accessed 2026-11-25]
8. Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. SIGMOD Rec. 2000;29(2):1-12. [doi: 10.1145/335191.335372]
9. Agresti A, Coull BA. Approximate is better than “exact” for interval estimation of binomial proportions. Am Stat. May 1998;52(2):119-126. [doi: 10.1080/00031305.1998.10480550]
10. Wald A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans Amer Math Soc. 1943;54(3):426-482. [doi: 10.2307/1990256]
11. O’Neill RT. On sample sizes to estimate the protective efficacy of a vaccine. Stat Med. Dec 1988;7(12):1279-1288. [doi: 10.1002/sim.4780071208] [Medline: 3231951]
12. Bose M, Biswas A. Sample sizes required to estimate the protective efficacy of a vaccine when there is an unequal allocation of individuals across the vaccine and placebo groups. Stat Methods Med Res. Oct 2023;32(10):1859-1879. [doi: 10.1177/09622802231176807] [Medline: 37647224]
13. Disease pattern. University of Eastern Finland (UEF). URL: https://cs.uef.fi/ml/impro/disease-pattern/ [Accessed 2026-11-04]
14. Data Protection Act. Finlex. URL: https://www.finlex.fi/en/legislation/translations/2018/eng/1050 [Accessed 2026-11-11]
15. Laki sosiaali- ja terveystietojen toissijaisesta käytöstä [Article in Finnish]. Finlex. URL: https://www.finlex.fi/en/legislation/2019/552 [Accessed 2026-11-11]
16. Poston JT, Koyner JL. Sepsis-associated acute kidney injury. BMJ. Jan 9, 2019;364:k4891. [doi: 10.1136/bmj.k4891] [Medline: 30626586]
17. Zarbock A, Nadim MK, Pickkers P, et al. Sepsis-associated acute kidney injury: consensus report of the 28th Acute Disease Quality Initiative Workgroup. Nat Rev Nephrol. Jun 2023;19(6):401-417. [doi: 10.1038/s41581-023-00683-3] [Medline: 36823168]
18. Mullens W, Abrahams Z, Francis GS, et al. Importance of venous congestion for worsening of renal function in advanced decompensated heart failure. J Am Coll Cardiol. Feb 17, 2009;53(7):589-596. [doi: 10.1016/j.jacc.2008.05.068] [Medline: 19215833]
19. Damman K, Navis G, Smilde TDJ, et al. Decreased cardiac output, venous congestion and the association with renal impairment in patients with cardiac dysfunction. Eur J Heart Fail. Sep 2007;9(9):872-878. [doi: 10.1016/j.ejheart.2007.05.010] [Medline: 17586090]
20. Rangaswami J, Bhalla V, Blair JEA, et al. Cardiorenal syndrome: classification, pathophysiology, diagnosis, and treatment strategies: a scientific statement from the American Heart Association. Circulation. Apr 16, 2019;139(16):e840-e878. [doi: 10.1161/CIR.0000000000000664] [Medline: 30852913]
21. Chen JJ, Lee TH, Kuo G, et al. Acute kidney disease after acute decompensated heart failure. Kidney Int Rep. 2022;7(3):526-536. [doi: 10.1016/j.ekir.2021.12.033] [Medline: 35257065]
22. Chen X, Xu J, Li Y, et al. Risk scoring systems including electrolyte disorders for predicting the incidence of acute kidney injury in hospitalized patients. Clin Epidemiol. 2021;13:383-396. [doi: 10.2147/CLEP.S311364] [Medline: 34093042]
23. Giurato L, Meloni M, Izzo V, Uccioli L. Osteomyelitis in diabetic foot: a comprehensive overview. World J Diabetes. Apr 15, 2017;8(4):135-142. [doi: 10.4239/wjd.v8.i4.135] [Medline: 28465790]

24. Chicco M, Singh P, Beitverda Y, Williams G, Hirji H, Rao GG. Diagnosing pelvic osteomyelitis in patients with pressure ulcers: a systematic review comparing bone histology with alternative diagnostic modalities. J Bone Jt Infect. 2020;6(1):21-32. [doi: 10.5194/jbji-6-21-2020] [Medline: 32983845]
25. Kaka AS, Beekmann SE, Gravely A, Filice GA, Polgreen PM, Johnson JR. Diagnosis and management of osteomyelitis associated with stage 4 pressure ulcers: report of a query to the Emerging Infections Network of the Infectious Diseases Society of America. Open Forum Infect Dis. Nov 2019;6(11):ofz406. [doi: 10.1093/ofid/ofz406] [Medline: 31696138]
26. O'Donnell MJ, Chin SL, Rangarajan S, et al. Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study. Lancet. Aug 20, 2016;388(10046):761-775. [doi: 10.1016/S0140-6736(16)30506-2] [Medline: 27431356]
27. Lewington S, Clarke R, Qizilbash N, Peto R, Collins R, Prospective Studies Collaboration. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. Lancet. Dec 14, 2002;360(9349):1903-1913. [doi: 10.1016/s0140-6736(02)11911-8] [Medline: 12493255]
28. Wolf PA, Abbott RD, Kannel WB. Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. Stroke. Aug 1991;22(8):983-988. [doi: 10.1161/01.str.22.8.983] [Medline: 1866765]
29. Pendlebury ST, Rothwell PM. Prevalence, incidence, and factors associated with pre-stroke and post-stroke dementia: a systematic review and meta-analysis. Lancet Neurol. Nov 2009;8(11):1006-1018. [doi: 10.1016/S1474-4422(09)70236-4] [Medline: 19782001]
30. Chung YR, Ha KH, Lee K, Kim DJ. Diabetic retinopathy and related clinical practice for people with diabetes in Korea: a 10-year trend analysis. Diabetes Metab J. Dec 2020;44(6):928-932. [doi: 10.4093/dmj.2020.0096] [Medline: 32662253]
31. Merlini G, Bellotti V. Molecular mechanisms of amyloidosis. N Engl J Med. Aug 7, 2003;349(6):583-596. [doi: 10.1056/NEJMra023144] [Medline: 12904524]
32. de Jager J, Kooy A, Lehert P, et al. Long term treatment with metformin in patients with type 2 diabetes and risk of vitamin B-12 deficiency: randomised placebo controlled trial. BMJ. May 20, 2010;340:c2181. [doi: 10.1136/bmj.c2181] [Medline: 20488910]
33. American Diabetes Association Professional Practice Committee. 9. Pharmacologic approaches to glycemic treatment: Standards of Care in Diabetes-2024. Diabetes Care. Jan 1, 2024;47(Suppl 1):S158-S178. [doi: 10.2337/dc24-S009] [Medline: 38078590]
34. Pfeiffer RF. Autonomic dysfunction in Parkinson's disease. Neurotherapeutics. 2020;17(4):1464-1479. [doi: 10.1007/s13311-020-00897-4] [Medline: 32789741]
35. Cutsforth-Gregory JK, Low PA. Neurogenic orthostatic hypotension in Parkinson disease: a primer. Neurol Ther. 2019;8(2):307-324. [doi: 10.1007/s40120-019-00152-9] [Medline: 31456212]
36. Mendoza-Velásquez JJ, Flores-Vázquez JF, Barrón-Velázquez E, Sosa-Ortiz AL, Illigens BMW, Siepmann T. Autonomic dysfunction in α-synucleinopathies. Front Neurol. 2019;10:363. [doi: 10.3389/fneur.2019.00363] [Medline: 31031694]
37. Ma Y, Zhang Y, Coresh J, et al. Orthostatic blood pressure change, dizziness, and risk of dementia in the ARIC study. Hypertension. Jan 2024;81(1):96-106. [doi: 10.1161/HYPERTENSIONAHA.123.21438] [Medline: 37869909]
38. Hendrix L, Verelst S, Desruelles D, Gillet JB. Deliberate self-poisoning: characteristics of patients and impact on the emergency department of a large university hospital. Emerg Med J. Jan 2013;30(1):e9. [doi: 10.1136/emermed-2011-201033] [Medline: 22328636]
39. Geulayov G, Kapur N, Turnbull P, et al. Epidemiology and trends in non-fatal self-harm in three centres in England, 2000-2012: findings from the Multicentre Study of Self-harm in England. BMJ Open. Apr 29, 2016;6(4):e010538. [doi: 10.1136/bmjopen-2015-010538] [Medline: 27130163]
40. Miller TR, Teti LO, Lawrence BA, Weiss HB. Alcohol involvement in hospital-admitted nonfatal suicide acts. Suicide Life Threat Behav. Oct 2010;40(5):492-499. [doi: 10.1521/suli.2010.40.5.492] [Medline: 21034211]
41. Choudhary GI, Fränti P. Predicting onset of disease progression using temporal disease occurrence networks. Int J Med Inform. Jul 2023;175:105068. [doi: 10.1016/j.ijmedinf.2023.105068] [Medline: 37104895]
42. Uddin S, Wang S, Khan A, Lu H. Comorbidity progression patterns of major chronic diseases: the impact of age, gender and time-window. Chronic Illn. Jun 2023;19(2):304-313. [doi: 10.1177/17423953221087647] [Medline: 35306857]
43. Su X, Ren Y, Li M, Zhao X, Kong L, Kang J. Prevalence of comorbidities in asthma and nonasthma patients: a meta-analysis. Medicine (Baltimore). May 2016;95(22):e3459. [doi: 10.1097/MD.0000000000003459] [Medline: 27258489]
44. Rocca E, Anjum RL. Complexity, reductionism and the biomedical model. In: Anjum RL, Copeland S, Rocca E, editors. Rethinking Causality, Complexity and Evidence for the Unique Patient. Springer; 2020:75-94. [doi: 10.1007/978-3-030-41239-5_5]

45. Xiao T, Lee YA, Xu Z, et al. Temporally detailed hypergraph neural odes for disease progression modeling. arXiv. Preprint posted online on Oct 20, 2025. [doi: 10.48550/arXiv.2510.17211]
46. Lachinov D, Pinetz T, Bogunovic H. PINN-EM: physics-guided disease progression model of geographic atrophy. IEEE Trans Biomed Eng. Oct 2026;73(10):3555-3566. [doi: 10.1109/TBME.2026.3655164] [Medline: 41553889]
47. Hu J, Zhang B, Xu T, Yang H, Gao M. Structure-aware temporal modeling for chronic disease progression prediction. arXiv. Preprint posted online on Aug 20, 2025. [doi: 10.48550/arXiv.2508.14942]
48. Puri P, Kachhadia MP, Singla N, et al. Anemia in heart failure: a systematic review of prevalence and prognostic impact. Cureus. Nov 2025;17(11):e96363. [doi: 10.7759/cureus.96363] [Medline: 41367439]
49. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ. Jan 21, 1995;310(6973):170. [doi: 10.1136/bmj.310.6973.170] [Medline: 7833759]

Abbreviations

cSPADE: constrained sequential pattern discovery using equivalence classes
Eclat: equivalence class clustering and bottom-up lattice traversal
FP-Growth: frequent pattern growth
FP-Tree: frequent pattern tree
***ICD-10*:** *International Statistical Classification of Diseases, Tenth Revision*
MSDD: multistream dependency detection
PrefixSpan: prefix-projected sequential pattern mining
RR: relative risk
RW: relative width
WHO: World Health Organization

Edited by Andrew Coristine, Arriel Benis; peer-reviewed by Hao Xu, Tingsong Xiao; submitted 04.Mar.2026; initial comments to author 14.Jun.2026; final revised version received 25.Sep.2026; accepted 20.Oct.2026; published 26.Nov.2026

Please cite as:
Ratul SN, Tuononen JP, Laatikainen T, Fränti P
Frequent Pattern Mining for Early Risk Signal Detection in Population Health Data: Nationwide Retrospective Cohort Study
JMIR Med Inform 2026;14:e94284
URL: https://medinform.jmir.org/2026/1/e94284
doi: 10.2196/94284

© Sazid Nur Ratul, Jimi Patrik Tuononen, Tiina Laatikainen, Pasi Fränti. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 26.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.