7.3.2022

to appear in *JMIR Medical Informatics*

Clustering Diagnoses from 58M Patient Visits in Finland 2015–2018

Pasi Fränti[1], Sami Sieranoja[1], Katja Wikström[2,3] and Tiina Laatikainen[2,3]

[1]*Machine Learning Group*

School of Computing, University of Eastern Finland

P.O. Box 111, FIN-80101 Joensuu, FINLAND

pasi.franti@uef.fi, sami.sieranoja@uef.fi,

[2]*Institute of Public Health and Clinical Nutrition,*

University of Eastern Finland

PO Box 1627, FI-70211 Kuopio, FINLAND

[3]*The Department of Public Health and Welfare,*

Finnish Institute for Health and Welfare,

PO Box 30, FI-00271 Helsinki, FINLAND

katja.wikstrom@thl.fi, tiina.laatikainen@thl.fi

Abstract

Background: Patients with multiple chronic diseases cause a major burden to the health service system. Currently, diseases are mostly treated separately without paying enough attention to their relationships, which results in a fragmentation of the care process. Better integration of services can lead to more effective organization of the overall health care system.

Objective: To analyze the connections between diseases based on their co-occurrences in order to support decision-makers in better organizing health care services.

Methods: We performed cluster analysis of diagnoses using data from the Finnish Health Care Registers for primary and specialized health care visits and inpatient care. The target population of this study comprised those 4.3 million individuals (78% of whole population) aged 18 years or older who used health care services during the years 2015–2018. They produced a total of 58M visits. Clustering was performed based on the co-occurrence of diagnoses. The more the same pair of diagnoses appears in the records of same patients, the more the diagnoses correlate. Based on the co-occurrences, we calculated the relative risk of each pair of diagnoses and clustered the data using a graph-based clustering algorithm called M-algorithm, a variant of k-means.

Results: The results reveal multimorbidity clusters, of which some are expected, for example one representing hypertensive and cardiovascular diseases. Other clusters are more unexpected, such as a cluster containing lower respiratory tract diseases and systemic connective tissue disorders. The annual costs of all clusters total 10.0 billion euros and the costliest cluster is Cardiovascular and metabolic problems with 2.3 billion euros.

Conclusions: The method and achieved results provide new insight to identify key multimorbidity groups, especially ones resulting in burden and costs in health care services.

Keywords: Multimorbidity, cluster analysis, disease co-occurrence, comorbidity network, healthcare data analysis**,** graph clustering, k-means.

1. Introduction

1.1 Multimorbidity

Patients with multiple chronic diseases cause a major burden to the health service system both in terms of service utilization and costs [1]. In many service systems diseases are mostly treated separately without paying enough attention to their relationships, which results in a fragmentation of the care process. Better integration of services can lead to more effective organization of the overall health care system. To support this, we analyzed the connections between diseases based on their co-occurrence and performed clustering analysis to find multimorbidity patterns.

Multimorbidity is often defined as the *coexistence of two or more chronic conditions* within a patient [2], [3] but the number of medical conditions included in this definition ranges widely [4]. Systematic reviews have shown that multimorbidity reduces self-rated health, quality of life and functional ability and increases the risk of premature death, hospitalizations, and use of health services causing substantial economic burden for societies and health care systems [5]. Wang et al. [6] reported that multimorbidity cases, defined as a patient having two or more chronic conditions, have 2–16 times higher costs than non-multimorbidity cases . Brettschneider et al. [7] analyzed the impact of 45 conditions on health-related life quality. The authors measured multimorbidity by a weighted count score and assessed its association to decrease in the health-related quality of life. The strongest impact was observed by Parkinson's disease, depression, and obesity.

One active research area is to measure the severity of multimorbidity. Stirland et al. [8] reviewed 35 multimorbidity measures. Most measures (25) in their review are based on simple (weighted or unweighted) counts of diseases; some measures (4) use drug counts, and some (5) are based on expert generated grouping of diagnoses, mainly based on frequencies. Such measures have been used for measuring mortality, health care use, cost, and quality of life.

1.2 Diagnose groups

The number of possible multimorbidities (connections between all diagnoses) is too large for a human analyst to examine them individually. It is easier to analyze by first dividing the diagnoses into smaller groups that contain related diagnoses and then examine the connections between diagnoses in these groups.

The diagnose groups can also predict future costs for a patient. Farley [11] discovered that simply by counting the number of diagnose clusters a patient belongs to is a good predictor for high costs in future. When combined with other measures such as the number of prescriptions it outperformed more complex comorbidity indices such as the Charlson, Elixhauser and RxRisk-V indices [11].

Diagnose groups have previously been created manually by experts by joining diagnoses of clinical similarity together. Travers et al. [9] studied how well four groupings covered emergency medicine. The authors discovered that Agency for Healthcare Research and Quality (AHRQ) grouping for inpatient care provides the best coverage (99%) while National Center for Health Statistics (NCHS) vital statistics grouping covered only 88%. They also criticized that most clusters (76%) are small, and that there are large clusters containing dissimilar conditions. Open questions are how to evaluate a cluster system and how to find its clinical relevance. Travers et al. further argued that a good clustering system should collapse individual ICD-9-CM codes into clinically meaningful clusters.

The number of groups is also problematic. Schneeweiss et al. [10] argued that 367 clusters are too many for comparative analysis while 17 clusters are too broad for the purpose. The authors reduced the number of ICD categories to 110 diagnosis clusters by cross-tabulation between ICD-9-CM and ICHPPC-2 classifications, covering about 90% of all diagnoses of their records made by family physicians.

1.3 Clustering to detect multimorbidity patterns

An alternative to manual grouping of diagnoses is to use computer algorithms to create the groups. Cluster is a group of objects which are similar to each other while objects in different clusters are expected to be far from each other, or at least less similar than those in the same cluster [Jain1998]. Clustering can be used to detect multimorbidity patterns by grouping either patients or diseases [18]. If we group the diagnoses, one diagnosis belongs to only one cluster but a patient can belong to several clusters. If we group the patients, the reverse is true: one diagnosis can belong to several groups but a patient can belong to only one cluster. This paper focuses on grouping the diagnoses.

The data used in the clustering can be either numerical values or text. Here we follow Hidalgo et al. [46] and represent the diagnoses as nodes and their relations as links in a network. We refer to this as *multimorbidity network*. In this network, the weight of the links between two diagnoses measure how strongly they correlate in a patient record database.

Although clustering algorithms have been widely used elsewhere in health care, the existing literature lacks reliable, automatic, computer-generated clusters. Estiri et al. [12] used clustering to detect anomalies in health records by combining agglomerative clustering and a k-means algorithm. The idea is to detect small clusters and flag them as anomalies. The authors reported a significantly smaller number of false positive cases than simple anomaly detections based on standard deviation and the Mahalanobis distance.

Huang et al. [13] clustered patients into five clinically meaningful groups based on the similarity of their diagnoses and the geographical locations of the hospitals. Their motivation was to build machine learning models trained for each group separately to provide better prediction of mortality and intensive care unit (ICU) stay time.

Kalgotra et al. [14] used co-occurrence statistics to build a *multimorbidity network* to study the disparity of gender. The statistics were extracted from treatment data of more than 22.1 million patients. They created networks separately for males and females and compared the structures of the two networks. Female patient networks had more connections to mental health.

Folino et al. [15] clustered patients based on a multimorbidity network built from co-occurrence statistics. They used the k-means clustering algorithm with Jaccard distance. A representative of each cluster was chosen as the set of all diseases whose relative frequency in the cluster exceeded a user-defined threshold (e.g., 0.8). The clustering was used to predict future diseases and was tested with records of 1,462 patients of a small town in South Italy.

In [16], the same prediction system was revised by using common neighbors in the network. Records of 2,541 patients during 2000–2009 were used to build a network from ICD-9-CM codes. The resulting network had 492 nodes and 21,676 connections. Two separate sub-networks were created. The first included only connections with *relative risk* value >20 (2,330 connections), and the other included those with a Pearson correlation value ≤ 0.06 (7,242 connections). Future patient diseases were predicted by calculating the number of common neighbors shared by the two diseases.

Ding et al. **[**17**]** extended the previous prediction model using ICD-10 and demographic data. Based on data collected between 2007 and 2014 in a (unnamed) provincial capital in China, they reported that 71% of acute diseases and 82% of chronic diseases are predictable.

John et al. [19] applied clustering for 1,039 American Indians using data from interview-based questionnaire. Cornell et al. [20] used ICD-9 codes from data obtained from administrative databases of primary care clinics. Marengoni et al. [21] used electronic medical records of acute care wards of 38 internal medicine and geriatric wards in Italy during 2008.

Marengoni et al. [21] calculated clusters of diseases to detect groups of patients that are at risk of in-hospital death. Their data consisted of 1,332 elderly people hospitalized in acute care wards. This small data set had 19 diagnoses which they grouped to eight clusters using a correlation matrix and average linkage agglomerative clustering. The result included four cluster consisting

of a disease and its possible consequences. For example, diabetes clustered with cerebrovascular diseases and coronary heart diseases; thyroid dysfunction with anxiety; chronic renal failure (CRF) with anemia. The combination of CRF and anemia had the highest likelihood of in-hospital death with an odds ratio of 6.1.

Most of the existing studies on clustering are based on hierarchical agglomerative method using heuristic criteria, either *average linkage* or *complete linkage* [18]. Wartelle et al. [22] extended the hierarchical agglomerative clustering optimizing the clustering directly using *relative risk*. This is by-default more solid approach than any linkage criterion (single, average, complete). They applied the method on data collected from the emergency department (ED) of Troyes hospital in Eastern France during a two-year period between 2017 and 2019. A network consisting of 151 ICD-10 blocks was created using 114,391 hospital visits of 72,666 patients.

1.4 Proposed methodology

In this paper, instead of agglomerative clustering, we apply a *k-means* based algorithm. Previously k-means has been used for clustering patients [23]. We apply the algorithm for clustering diseases, using data consisting of 45 million health care visits covering all public health service use (both primary and secondary care) of population aged 18 or older in entire Finland from 2015 to 2018. This data set is significantly larger than in any of the previous studies.

We constructed a multimorbidity network which consists of diseases represented as blocks of ICD-10 codes. Correlating diseases are linked in the network. The strength of the links between the diseases are measured using *relative risk* which estimates how much higher the observed prevalence is in relation to the expected. Clustering is used to find multimorbidity patterns by dividing the network into subgroups that have high relative risk values within. These groups can contain previously unknown multimorbidity patterns.

Similar to [22], our study is also based on relative risk. However, there are two main differences. First, the agglomerative clustering algorithm in [22] needs to access the original data after each merge to re-calculate the relative risk values, which becomes very time consuming with large data. We construct the network only once without any need to access the original data after that. This approach scales better as the network is remarkably smaller than the original data (205 nodes vs. 58 million patients). K-means itself may require multiple runs [24] to create accurate clustering, but we avoid this by using a more robust derivation called an M-algorithm [25].

The second difference is that the results of [22] are from emergency visits. While the resulting clusters can be valid in this context, the generated clusters are different from what we obtained from all general health care visits.

The main contributions of our paper can be summarized as follows:

- We use a k-means based algorithm called M-algo, which is shown to provide highly accurate clustering with controlled validation datasets and scaling up to large-scale data [25].
- We use inverse internal weight in the network as a cost function as it has been shown to provide more balanced cluster sizes than the other counterparts [25].
- We apply the algorithm for large-scale data consisting of 58 million health care visits in all of Finland from 2015 to 2018.
- We make the data publicly available, including the multimorbidity network and the clusters.[1]

These contributions directly support several of the goals described by Whitty and Watt [26]. These objectives include strengthening statistical methods to detect clusters, application to large data sets, and to treat clusters of disease more effectively. In this paper, we describe the content of the

[1] http://cs.uef.fi/ml/impro/DiagnosisClusters

generated clusters and their relationships with the nearby clusters. We report the most significant observations and their effect on both the service utilization and costs in the health care system. The paper follows the TRIPOD guidelines [63] in all the relevant items except the ones that relates to prediction.

2. Methods

Graph clustering has been used in physics [27,28], engineering [29], image processing [30], medical [31] and social sciences [32]. The technique has several names including *network community detection* [33-39], *graph clustering* [40] or *graph partitioning* [30,41,42]. These methods can be directly applied to diseases by considering the co-occurrence matrix of diseases as a graph.

By grouping data into meaningful clusters, by finding co-occurring diagnoses, it is possible to plan the treatment processes of multimorbid patients and resources needed in service provision. It is known that diseases often cluster due to a common risk factor, but only a small number of the possible clusters and the connections between the clusters are well known [26].

2.1 Data

A summary of the patient record database is presented in Table 1. The data was extracted from the National Administrative Care Register for Health Care, covering all inpatient and outpatient care both in primary and specialized care between 2015 and 2018. Finnish Health Care Registers include data on the patient's age, gender, municipality of residence as well as information concerning the service event, such as the type of contact (visit, phone call, inpatient admission), the reason for the visit, treatment, and procedures. Reasons for the visits are recorded using ICD-10 or ICPC-2 codes.

The entire patient record database contains information on 4.3 million patients over 18 years old. For the cluster analysis, we included only patients with a medical diagnosis (excluding external cause diagnoses), which totaled 3.8 million. The full database includes almost 312 million contacts with the health services. The visits are divided into 272,090,337 contacts with primary care and 39,631,625 contacts with special care services. The primary care contacts include 142,874,297 home visits, 71,658,708 visits to a health center, 26,849,249 phone calls, and 30,708,083 other types of contacts.

Table 1 Summary of the patient database

Whole database

All patients	4,280,985
Patients with ICD-10 codes	3,987,382
Time range	2015–2018
Total visits	311,721,962
Visits with ICD-10 codes	69,306,854
Diagnoses per visit (if any)	1.6
Total cost of all visits / year	€9,685 M

Included in clustering

Visits	58,391,604
Costs/year	€6,596 M
Patients	3,835,531

Females/Males	54%/46%
Median age	54
Patients over 70 years	25%
Mean cost of patient per year	€2,414

For the clustering analysis, from all the visits (311,721,962), we included only those having ICD-10 diagnoses recorded (69,306,854). We excluded all the symptom codes (R00–R99), external causes for injuries, diseases and deaths (V01–Y92), health factors and contacts to the service providers (Z00–ZZB), as they do not represent any disease themselves, and special diagnosis codes (U00–U99). After filtering these out, the remaining data included 58,391,604 visits with a diagnosis code.

The costs for each diagnosis were calculated using computational standard cost [43,44] employing patient grouping methods and standard unit costs calculated from national level cost accounting projects. Hospitalizations and hospital outpatient visits were grouped using the NordDRG grouper. The NordDRG cost weights for hospitalizations and outpatient visits were based on individual-level cost accounting data from several hospitals and used in the national price lists by the Finnish Institute for Health and Welfare [45]. The unit cost estimates for each type of primary care contact were obtained from the national standard price list for primary care encounters. The unit cost estimates for social care encounters and community care bed-days were derived from the national price list for unit costs of health care services in Finland.

The total annual health service cost in Finland, during the period 2015–2018 was €9,685 M for a total of 311M visits. The cost estimation for the data used in cluster analysis totals to €6,596 M per year. Annual cost of each year has an increasing trend between 2015-2017 but decreases in 2018: €6,579 M (2015), €6,626 M (2016), €6,723 M (2017), €6,455 M (2018). Some changes may originate from changes in recording practices. Also, some patients that were hospitalized for longer periods (weeks or months) are not included in the 2018 data if they were not discharged by the end of 2018.

2.2 Measuring Relative risk

There are several possibilities to measure the strength of relation between two diseases (see Table 2). These include *φ-correlation* (Pearson's correlation) [31,46], *Co-occurrence correlation* [47], *Jaccard coefficient* [20], *Yule Q* [19,21], *Salton Cosine index* [14], and multiple variants of *Relative Risk* [15,16,25]. For a good review, see [47].

Several authors [14,22,47] have noted that existing measures contain biases. For example, Relative Risk overemphasizes the connection of infrequent diseases. Pearson's correlation underestimates the relation between common and infrequent diseases. Because of these problems, Srinivasan et al. [47] ended up proposing their own method, called *co-occurrence correlation*.

We use the Relative Risk (variant 1 in Table2) because this measure has been widely used in the literature and its values are clear to understand. It has been previously used by several authors [15,22,46] to study the relation of diagnoses. It can also be used for other purposes, for example to to study market baskets [48].

Table 2 Ways of measuring disease connectivity

Name	Formula	References
Relative Risk (1)	$\frac{P_{xy}N}{P_xP_y}$	[46,48]

Relative Risk (2)	$\frac{(P_{xy} - N)N}{P_x P_y}$	[15]
Relative Risk (3)	$\frac{P_{xy}N}{P_x P_y - C_{xy}}$	[49]
Co-occurrence correlation	$\frac{P_{xy}\sqrt{2}}{\sqrt{P_x^2 + P_y^2}}$	[47]
φ-correlation	$\frac{P_{xy}(N - P_{xy}) - P_x P_y}{\sqrt{P_x P_y (N - P_x)(N - P_y)}}$	[15,31,46] (slight variation:[49])

Notations:

N	Number of patients
P_x	Number of patients with diagnosis x (prevalence)
P_{xy}	Number of patients with both diagnosis x and y (prevalence)
E[xy]	Expected frequency of xy
$p(x)$	$= P_x / N$. Probability of a random patient having a diagnosis x
$p(xy)$	$= P_{xy} / N$. Probability of a random patient having both diagnosis x and y

Relative risk is defined based on the diagnose prevalences as follows:

$$\mathrm{RR}_{xy} = \frac{\text{observed}}{\text{expected}} = \frac{O(\mathrm{xy})}{E[\mathrm{xy}]} = \frac{p(\mathrm{xy})}{p(x)p(y)} = \frac{P_{xy}/N}{(P_x/\mathrm{N})(P_y/\mathrm{N})} = \frac{P_{xy}N}{P_x P_y}$$

Where p(x) and p(y) are the probabilities that a randomly chosen patient has the disease x and y, respectively, and p(xy) is the probability that a randomly chosen patient has them both. Fig. 1 demonstrates the detailed calculation of the relative risk values in case of asthma and sleep disorders. An RR-value > 1.0 indicates that the two diseases are related.

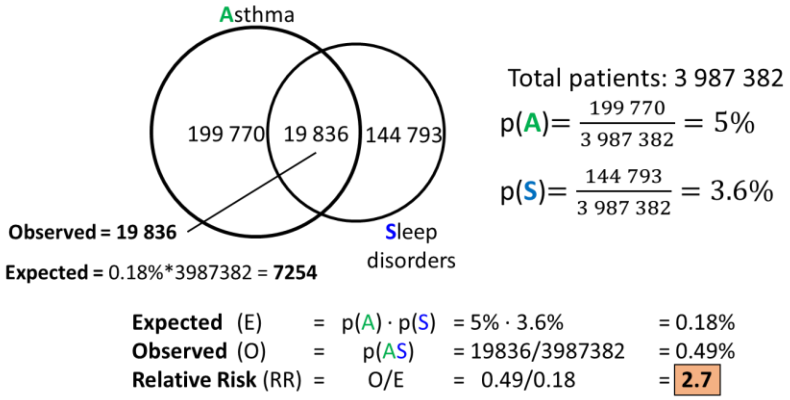

Fig. 1. Example of measuring comorbidity by relative risk. Here asthma and sleep disorders are highly correlated. If they were independent of each other, the probability of a person having both should be p(A)· p(B) = 0.18% while their observed co-occurrence is 0.49%. The relative risk to have both is therefore 2.7 times higher than by random chance.

Most relative risk values are between 0.5–5.0 but they can also be over 100. These outlier values would dominate the clustering cost function optimization, and for this reason, we normalize them to the range of [0,1] by using the following variant of the generalized symmetrical sigmoid function [65]:

$$w'_{xy} = \frac{w_{xy}}{w_{xy}+1}$$

The normalization function $f=x/(x+1)$ behaves similarly as the log-function, but it caps overly large outlier values more aggressively and is limited between [0,1]. Large RR-values (> 6) are not as important to the clustering process as differences in the range 1-6.

2.3 Multimorbidity network

A *multimorbidity network* is formed by connecting all pairs of diagnoses whose RR-value > **??** (see Fig. 2). Each node in this network, corresponds to a medical diagnosis and the strength of the connections is measured by using relative risk. The name multimorbidity network follows the choice of Aguado et al. [66] The network has also been called a *disease co-occurrence network* [45], *phenotypic disease network* [46], *comorbidity network* [14], and *disease comorbidities network* [31].

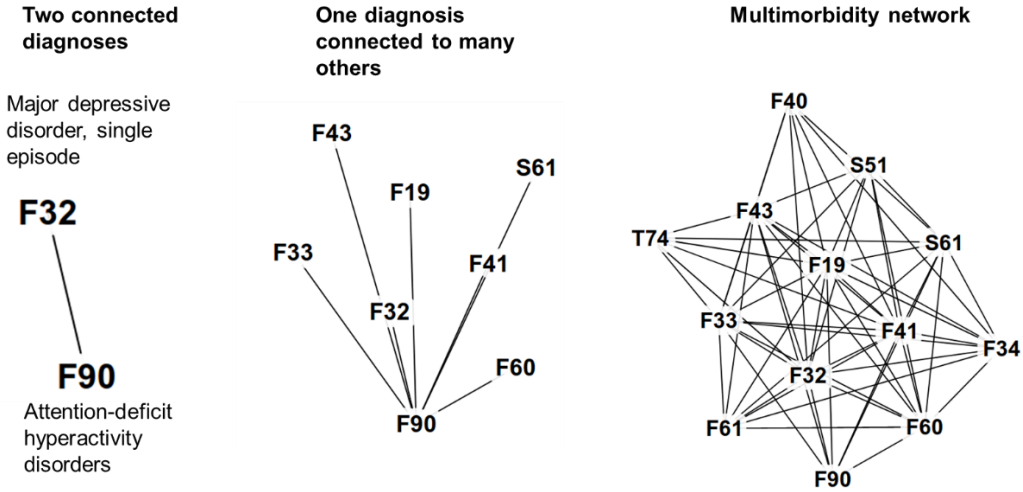

Figure 2. Multimorbidity network is formed by finding related diagnoses for all diagnoses in the data set.

There are several previous works that have used a multimorbidity network [14,15,31,46,47,66]. Also, Klimek et al. [50] and Moni and Liò [49] studied comorbidity associations, although they did not explore much of the network analysis portion. Moni and Liò [49] created an R language software called *comoR* for disease comorbidity risk analysis. Divo et al. [31] studied Chronic obstructive pulmonary disease (COPD) for disease screening and management. Folino et al. [15] predicted future diseases based on past medical history. Srinivasan et al. [47] used multimorbidity network to extract features for high cost patient prediction. Hidalgo et al. [46] also published their co-occurrence network data (based on 13 million patients) [51].

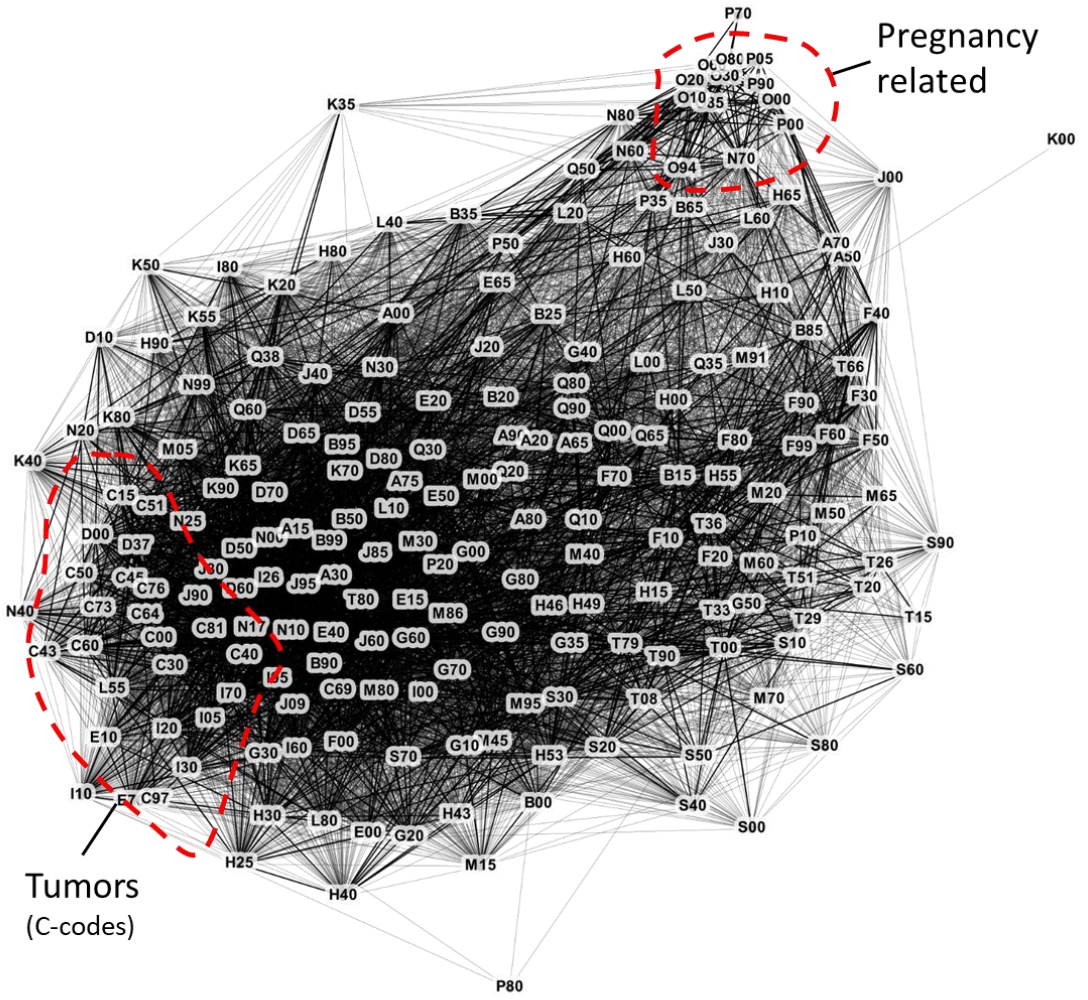

Figure. 3. The full network is overwhelming to analyze, with 205 disease subgroups and 14,254 connections overall. Here we show only the 8,895 connections with RR >1.5. Connections with RR > 3.0 are drawn with bold. ICD-10 subgroups are represented by the first diagnosis of the group (see Appendix I). Image created using the Gephi software [52]. Only very tight groups such as pregnancy related diagnoses and tumors can be recognized from the network.

We form the multimorbidty network (Figure. 3) by calculating the relative risk value for all pairs of diagnoses and including those with RR value $\geqslant$ 1.0 and at least 10 patients that have both diagnoses. The accuracy used for diagnoses is the subgroups of ICD-10 classification (e.g., I20–I25). We also filter out diagnoses that indicate symptoms and external causes (those starting with Z,W,Y and R). After filtering, we have 205 disease subgroups in the graph (see Appendix for the full list).

2.4 Clustering

The main motivation for clustering is that the multimorbidity network is too large (205 nodes and 14,254 connections) to analyze in detail. For this reason, we clustered the graph to form more compact entities of related diseases. The goal is to assign strongly related diseases into the same cluster but keep uncorrelated diseases in different clusters. To achieve this goal, an evaluation criterion is necessary to measure the effectiveness of the clustering.

Cost function

Instead of using heuristic criteria such as average or complete linkage, it is better to define an exact cost function that the clustering algorithm optimizes directly. When clustering numerical data, the typical goal is to measure the compactness of the clusters. For example, both Ward's method and k-means minimize the *sum-of-squared distances* between the data objects to the cluster mean (*centroid*). However, calculating the mean of a subgraph is not possible directly but would require some indirect solution such as vectorizing the nodes by graph embedding [59]. Moreover, calculating a distance between two nodes is not possible if they are not connected. Graph-specific cost functions have therefore been developed to overcome these issues.

Three cost functions were evaluated in [25] with controlled data: *conductance*, *mean internal weight*, and *inverse internal weight* (IIW). The last function produced the most accurate clustering result with balanced clusters sizes and is therefore chosen in this paper as well. When k is the number of clusters, W_i is the internal weights of cluster i and M is the total weight (mass) of the whole graph, the cost is calculated as follows:

$$\textit{Inverse internal weight} \qquad = \frac{1}{k}\sum_{i=1}^{k}\frac{M}{\mathrm{kW}_i} = \frac{M}{k^2}\sum_{i=1}^{k}\frac{1}{W_i} \qquad (1)$$

In the multimorbidity network analysis, it is desirable to have clusters of roughly the same size. This can be controlled by specifying the number of clusters. Since the cost function induces balanced cluster sizes, we aim at grouping the N nodes to k clusters of size $N/k = n$. In our case, we have $N = 205$ diseases and $k = 15$ clusters with $205/15 = 13.7$ diseases, on average. This size is small enough so that we can investigate the clusters manually.

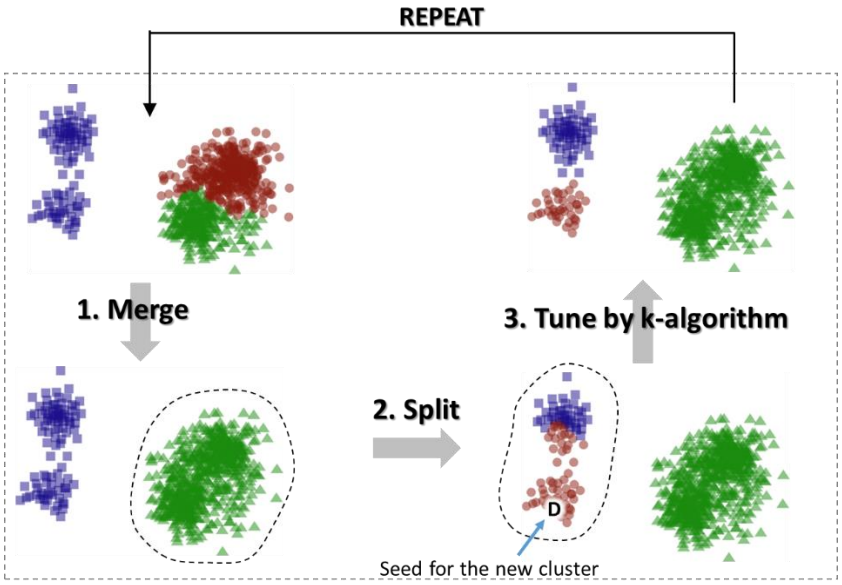

Figure 4. The M-algorithm merges two random clusters, splits one random cluster and fine-tunes the result by using the k-algorithm. The network in this example is the kNN graph of the presented 2D dataset.

Clustering algorithm

We used the recently developed *M-algorithm* in [25], which combines a k-means type of iterative optimization with an additional merge-and-split strategy to escape from local minima. *Inverse internal weights* was the recommended cost function.

K-means uses two optimization steps in turn: assignment and centroid steps. In the assignment step, every point is put into the cluster whose mean (centroid) is closest. However, the assignment of the points is not independent from the assignment of other points. Their joint effect may cause the cost value to fluctuate so that the total value increases even if the single assignment would decrease. To avoid this problem, we used the sequential variant of k-means where every assignment takes immediate effect on the centroids. This technique prevents the fluctuation.

The k-means variant applied for graphs is called the *K-algorithm*, which is like the original k-means algorithm but without the centroids. We replaced the distance calculations by evaluating the effect of the assignment to the cost function directly. Most cost functions are based on maximizing the weights inside the cluster or minimizing the external weights. The effect of a node joining to a cluster can therefore be calculated using only its edges and the size of the cluster.

The K-algorithm iteratively improves the initial solution by processing the nodes sequentially in random order. For each node, the method considers all clusters and checks if changing the partition of this node improves the cost function. If it does, the cluster assignment is changed. After all nodes have been processed, the algorithm starts another iteration. The iterations continue until no changes happen.

The M-algorithm differs from the K-algorithm by an additional merge-and-split step. The M algorithm first merges two random clusters and then splits one random cluster. The clustering solution is fine-tuned by the K-algorithm. If the new solution improves the cost function value, it is kept as the current solution; otherwise the process continues from the previous solution. The merge and split process is repeated depending how much computing time is wanted. The pseudocode for the algorithms is presented below:

K-algorithm(graph,k,clustering):

```
IF cluster == NULL
  cluster = InitialPartition(graph,k)
DO
  FOR i=1:N // Loop all nodes
    cluster[i] = find optimal cluster for node i according to cost function IIW
WHILE cluster improved
```

M-algorithm(graph,k,R):

```
cluster = K-algorithm(graph,k,NULL)
FOR i=1:R
  newClu = cluster
  newClu = Merge random pair of clusters A,B
  newClu = Split random cluster C
  newClu = K-algorithm(graph,k,newClu)
  IF newClu better than cluster
    cluster = newClu
RETURN cluster
```

Because the network itself is quite small (205 diagnoses), the clustering algorithm takes only little time. The time complexity of the M-algorithm is $O(RIN(k+|E|/N))$, where R is the number of repeats, N the number of diagnoses (nodes), k the number of clusters, $|E|/N$ the average number of connections for each node (diagnosis) and I is a small number reflecting the number of iterations to converge. We run the M-algorithm for 20,000 repeats which took 27 minutes (single thread) on Intel(R) Xeon(R) W-2255 CPU @ 3.70GHz. The bottleneck was the $O(N_v)$ network construction which needed to process all N_v = 58M patient visits and took 52 minutes.

The number of clusters k must be fixed by the researcher beforehand. Small number is likely to generate large mixed clusters of many diseases, losing the capability to make meaningful observations. Large number of clusters tend to cluster mainly diseases from the same ICD group

which might lose the chance to detect relevant multimorbidity patterns. We tried clustering with several different *k*-values and chose *k*=15 as it produced clusters with convenient size to analyze in the form of similarity matrices (see example in Fig. 5).

It is also possible to let the algorithm recommend the number of clusters using suitable cluster validity index that measures the ratio of the within cluster and between clusters similarities as in [53]. Wartelle et al derived validity index from the relative risk and obtained *k*=16 clusters in their data [22]. We tried *silhouette coefficient* [64] for our data and in the range 5..25 it obtained *k*=17 clusters. They are both close to our choice of *k*=15.

3. Results

3.1 Relative risk

Table 3 shows ten pairs of disease sub-groups having the highest relative risk values. They are diagnoses with the highest probability to appear jointly relative to the expected probability with the independent assumption. Some connections are obvious, often representing the same or closely related conditions (C40-C41 and C45-C49). Some have known explanations in medical science (F70-F79 and Q90-Q99) or have a clear causal relationship (D80-D89 and N00-N08). There are also connections with smaller relative risk values that are not so obvious at first sight, but they are clinically meaningful (I26-I28 and M30-M36). In addition to using the ICD-10 subgroups, we calculated the relative risk values for diagnoses in three-character precision as well. Some RR-values <1.0 were also found for such diagnoses like E10 and E11, which are exclusive for each other.

Table 3 Ten disease pairs with the highest relative risk value
Full list is available here: http://cs.uef.fi/ml/impro/DiagnosisClusters

RR	Count	Diagnose A		Diagnose B	
170.1	484	A80-A89	Viral infections of the central nervous system	G00-G09	Inflammatory diseases of the central nervous system
110.7	132	A15-A19	Tuberculosis	B90-B94	Sequelae of infectious and parasitic diseases
98.3	107	C40-C41	Malignant neoplasms of bone and articular cartilage	C45-C49	Malignant neoplasms of mesothelial and soft tissue
91.0	893	T20-T25	Burns and corrosions of external body surface, specified by site	T29-T32	Burns and corrosions of multiple and unspecified body regions
79.7	945	F70-F79	Mental retardation	Q90-Q99	Chromosomal abnormalities, not elsewhere classified
50.7	811	G35-G37	Demyelinating diseases of the central nervous system	H46-H48	Disorders of optic nerve and visual pathways
47.2	2386	D80-D89	Certain disorders involving the immune mechanism	N00-N08	Glomerular diseases
45.7	866	J85-J86	Suppurative and necrotic conditions of lower respiratory tract	J90-J94	Other diseases of pleura
45.3	328	N25-N29	Other disorders of kidney and ureter	Q60-Q64	Congenital malformations of the urinary system
42.0	238	F70-F79	Mental retardation	Q00-Q07	Congenital malformations of the nervous system

3.2 Clustering results

The overall clustering result is visualized as a graph in Fig. 4. The graph shows connections within the clusters, but all connections between the clusters have been eliminated for clarity.

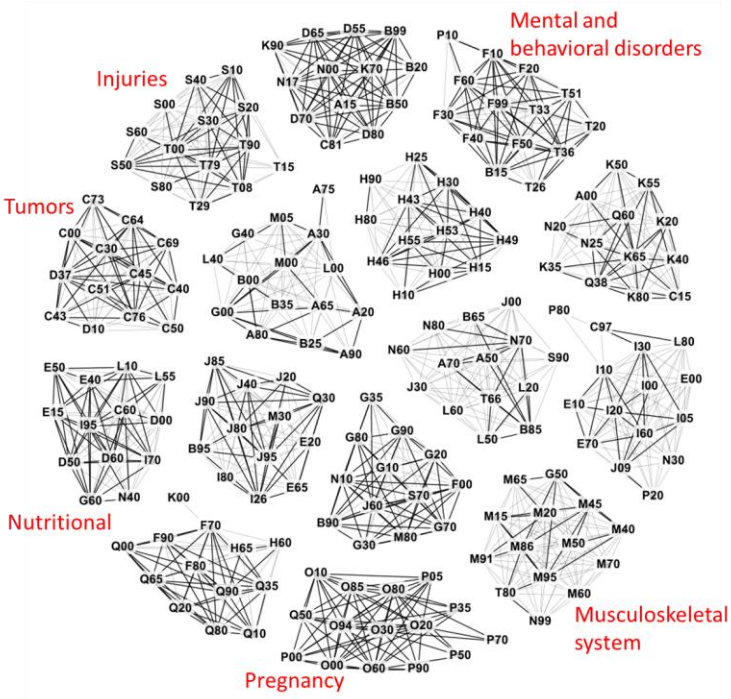

Fig. 5. Clusters obtained from the multimorbidity network. Subjective labels of six clusters are also shown. Showing all 205 diagnoses and only those 1,144 connections with RR ≥ 1.5. Cases when RR ≥ 3 are shown with thicker lines. ICD-10 blocks are represented by the first diagnosis of the block (e.g., F10–F19 by F10).

We have fixed the number of clusters to 15 for the M-algorithm [25]. This roughly matches the number 16 used in a study by Wartelle [22]. The main characteristics of the resulting clusters are summarized in Tables 4 and 5. The strength of associations between the diagnosis subgroups inside two example clusters and the connections between the two clusters can be observed in Fig. 5. The number of patients in each cluster, number of visits in health services, total costs, cost per visit, and cost per patient are reported in Table 6.

Most clusters are slightly dominated by records from female patients. Cluster 1 (100%) includes only females because it consists of pregnancy related diagnoses. Other clusters having >60% from females are Cluster 14 (67%) of mixed diseases (sexual and urinary), and Cluster 4 (63%) of malignant tumors. The only cluster with significantly higher proportion of diagnoses from males is Cluster 7 (67%), consisting of diagnoses mainly related to nutrition. In most other clusters, the proportion of men and women is about equal.

The main reasons for the female dominance are that the full database includes 1,999,325 males and 2,253,669 females, and that females had on average 6.6 diagnoses while males had only 5.4. A possible reason is that there is a lower threshold for females to seek help from health services than for males. For example, Corrigan's study [54] suggested that social factors discourage men from seeking mental health care, which can lead to absence of mental health-related multimorbidities among men.

In that all diagnoses were forced to belong to some cluster, there are several mixed clusters. For example, the largest cluster (Cluster 3) consists of 838,208 patients including all those with dental health problems (K00-K14). If this subgroup of diagnoses was removed, the number of patients would decrease to only 87,634 and would mainly consist of diagnoses related to mental retardation, congenital malformations, and chromosomal abnormalities. However, it is quite logical that dental health-related diagnoses are clustered with mental retardation, congenital malformations, and abnormalities as such patients usually also having malformations in the oral cavity, jaws, and teeth, which is a patient group treated in the public health service system.

The second largest cluster (Cluster 15) consisting of 773 406 patients includes cardiovascular diseases, endocrine and metabolic diseases. It also has the highest number of visits to health care (3.3 million annual visits). The third largest (Cluster 13) has 616 550 patients but is more focused on diagnoses related to diseases of the musculoskeletal system and connective tissues. Other more clearly focused clusters include tumors (Cluster 4), mental disorders (Cluster 6), injuries (Cluster 12), diseases related to nutrition (Cluster 7), and pregnancy (Cluster 1). These clusters are easy to explain based on the morbidity and mortality data in Finland. Cardiovascular diseases are still the major cause of death [55], and mental disorders are the main cause of disability pensions followed by musculoskeletal disorders [56].

The clusters also have clear age profiles. The average age of most clusters is rather high, being 60 or higher in the case of ten clusters. The exceptions are Cluster 6 (mental; 46y), Cluster 12 (injuries; 55y), mixed Clusters 3 (mental, ear and oral cavity; 49y) and 14 (sexual and urinary; 48y), and of course, Cluster 1 (pregnancy; 33y).

Table 4 Content of the 15 clusters (ICD-10 blocks) and their strength as the mean RR-value of diagnoses within the cluster.

Cluster	1	2	3	4	5	6	7
Mean RR	11.3	8.1	7.8	7.6	5.7	5.4	4.8
ICD-10 codes	O85-O92 O30-O48 O20-O29 O10-O16 O60-O75 O94-O99 O80-O84 P05-P08 P00-P04 O00-O08 P35-P39 P90-P96 Q50-Q56 P70-P74 P50-P61	B50-B64 N00-N08 D70-D77 C81-C96 D55-D59 D80-D89 D65-D69 B99-B99 A15-A19 N17-N19 B20-B24 K70-K77 K90-K93	F70-F79 Q90-Q99 F80-F89 Q00-Q07 Q35-Q37 Q80-Q89 Q65-Q79 F90-F98 Q20-Q28 Q10-Q18 H65-H75 H60-H62 K00-K14	C40-C41 C45-C49 C76-C80 C30-C39 D37-D48 C69-C72 C00-C14 C51-C58 C64-C68 C73-C75 C50-C50 C43-C44 D10-D36	J95-J99 J85-J86 J90-J94 J80-J84 Q30-Q34 I26-I28 J40-J47 B95-B98 M30-M36 J20-J22 E65-E68 E20-E35 I80-I89	T36-T50 B15-B19 F60-F69 F10-F19 F99-F99 T51-T65 F20-F29 F30-F39 T33-T35 T26-T28 F40-F48 T20-T25 F50-F59 P10-P15	E40-E46 E50-E64 D60-D64 I95-I99 D50-D53 L55-L59 D00-D09 E15-E16 G60-G64 I70-I79 L10-L14 C60-C63 N40-N51

Cluster	8	9	10	11	12	13	14	15
Mean RR	4.6	4.5	4.3	3.8	3.0	2.9	2.9	2.1
ICD-10 codes	G80-G83 G10-G14 J60-J70 G90-G99 F00-F09 G70-G73 G30-G32 N10-N16 B90-B94 S70-S79 M80-M85 G20-G26 G35-G37	Q60-Q64 N25-N29 K65-K67 Q38-Q45 C15-C26 K80-K87 K55-K64 K40-K46 N20-N23 K20-K31 K50-K52 A00-A09 K35-K38	G00-G09 A80-A89 A90-A99 A65-A69 M00-M03 A30-A49 B25-B34 A20-A28 M05-M14 B00-B09 L00-L08 L40-L45 B35-B49 G40-G47 A75-A79	H53-H54 H46-H48 H55-H59 H49-H52 H43-H45 H30-H36 H15-H22 H40-H42 H25-H28 H00-H06 H10-H13 H90-H95 H80-H83	T00-T07 T90-T98 T79-T79 S10-S19 S30-S39 S20-S29 T08-T14 T29-T32 S50-S59 S40-S49 S80-S89 S60-S69 S00-S09 T15-T19	M95-M99 M40-M43 M45-M49 M86-M90 T80-T88 G50-G59 M15-M19 M20-M25 M50-M54 M91-M94 M65-M68 M70-M79 N99-N99 M60-M63	A50-A64 A70-A74 B85-B89 N70-N77 B65-B83 T66-T78 L50-L54 L20-L30 N80-N98 L60-L75 J30-J39 N60-N64 J00-J06 S90-S99	I30-I52 I20-I25 I60-I69 I10-I15 L80-L99 I05-I09 J09-J18 E70-E90 N30-N39 E10-E14 E00-E07 I00-I02 P20-P29 C97-C97 P80-P83

Table 5 Summarization of the cluster content with their age and gender distributions

Cluster	Dominant gender	Median age	Age ≥70	Description
1 Pregnancy	females 100%	33	0%	Pregnancy, childbirth and the puerperium (O codes), certain conditions and disorders originating in perinatal period (P05-P08 P00-P04 P35-P39 P90-P96 P70-P74 P50-P61), and congenital malformations of genital organs (Q50-56)
2 Immune system and blood-forming organs	males 51%	69	50%	Infectious diseases strongly affecting the immune system (B50-B64, B20-24, B99-B99, A15-19), malignant neoplasms of lymphoid, hematopoietic and related tissue (C81-96), diseases of the kidneys (N00-N08, N17-N19), liver (K70-77), blood and blood-forming organs and disorders of the immune mechanism (D70-D77 D55-D59 D80-D89 D65-D69, (except nutritional and aplastic and other anemias), other diseases of the digestive system (K90-K93).
3 Mixed cluster. Includes mental disorders, malformations, ear and oral cavity diseases	females 55%	49	17%	Mental retardation (F70-79) and disorders of psychological development/unspecified disorder (F80-F89, F99-F99) and congenital malformations (Q codes except codes for congenital malformations of respiratory system, digestive system, genital organs and urinary system), diseases of the ear (H65-H75, H60-H62), diseases of oral cavity, salivary glands and jaws (K00-K14)
4 Tumors	females 63%	66	42%	Malignant neoplasms (all C codes, except codes for malignant neoplasm in digestive organs, male genital organs, lymphoid, hematopoietic and related tissue, independent multiple sites) and benign neoplasms (D10-D36)
5 Lower respiratory system	females 59%	64	38%	Lower respiratory tract diseases and related inflammatory conditions (J95-J99 J85-J86 J90-J94 J80-J84 J40-J47 J20-J22), congenital malformations of the respiratory system (Q30-Q36), pulmonary heart disease and diseases of pulmonary circulation (I26-I28), bacterial, viral and other infectious agents (B95-B98), systemic connective tissue disorders (M30-M36), obesity (E65-E68) and disorders of other endocrine glands (E20-E35), diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified (I80-I89)
6 Mental and behavioral disorders	females 58%	46	15%	Mental and behavioral disorders and substance abuse problems (F60-F69, F10-F19, F20-F29 F30-F39 F40-F48 F50-F59, F99), poisonings (T36-T50, T51-T65) and certain viral infections (B15-B19), and related burns (T20-T25, T26-T28), frostbite injuries (T33-T35), and birth trauma (P10-P15).
7 Nutritional	males 67%	72	58%	Malnutrition (E40-E46) and nutritional deficiencies (E50-64), anemias (D50-D53, D60-D64), other and unspecified disorders of the circulatory system (I95-I99), certain skin diseases (L55-L59, L10-L14), in situ neoplasms (D00-D09), other disorders of glucose regulation and pancreatic internal secretion (E15-E16), polyneuropathies (G60-G64), diseases of arteries, arterioles and capillaries (I70-I79), diseases and malignant neoplasms of male genital organs (C60-C63, N40-N51)
8 Diseases related to aging	females 60%	76	64%	Cerebral palsy, memory disorders, other diseases of the central nervous system/neurodegenerative diseases (included G-codes), lung diseases due to external agents (J60-J70), organic mental disorders (F00-F09), renal tubulo-interstitial diseases (N10-N16), changes in bone structure (M80-85) and injuries (hip and thigh S70-S79), other infections (B90-B94)
9 Mixed cluster. Includes organ malformations and digestive system disorders	females 54%	63	37%	Congenital malformations of the urinary system and digestive system (Q60-Q64, Q38-Q45), some disorders of kidney and ureter (N25-N29) and genitourinary system (N20-N23), diseases of the digestive system (all K codes, except diseases of oral cavity, salivary glands and jaw, and diseases of liver, malignant neoplasms of

				digestive organs (C15-C26), and intestinal infectious diseases (A00-A09)
10 Infections and inflammation	females 54%	61	33%	Inflammatory diseases (G00-G09)/viral infections (A80-A89) of the central nervous system, hemorrhagic fevers (A90-A99), certain other infectious and parasitic diseases (A65-A69, A30-A49, A20-A28, A75-A79, B00-B09, B35-B49), infectious arthropathies/inflammatory polyarthropathies (M00-M03, M05-M14), infections of the skin and subcutaneous tissue/ papulosquamous disorders (L00-L08, L40-L45), episodic and paroxysmal disorders (G40-G47)
11 Eye and ear	females 59%	67	45%	Diseases of the eye and adnexa (all H codes) and diseases of inner ear (H80-H83) and other disorders of ear (H90-H90)
12 Injuries	males 51%	55	26%	Injuries in different parts of the body (all S codes, except injuries to the hip and thigh) and in multiple body regions (T00-T07)/unspecified parts (T08-T14, T29-T32), effects of foreign body entering through natural orifice (T15-T19), and some of their consequences (T79-T79, T90-T98)
13 Musculoskeletal system	females 59%	60	31%	Diseases of the musculoskeletal system and connective tissue (all M codes, except infectious and inflammatory arthropathies/poly arthropathies, systemic connective tissue disorders, disorders of bone density and structure), complications of surgical and medical care (T80-T88), nerve, nerve root and plexus disorders (G50-G59), and other disorders of the genitourinary system (N99-N99)
14 Mixed cluster: Includes sexually transmitted, parasitic, urinary tract diseases	females 66%	48	19%	Sexually transmitted diseases (A50-A64, A70-A74), parasitic diseases (B85-B89, B65-B83), unspecified effects of external causes (T66-T78), inflammatory diseases of female pelvic organs (N70-N77), disorders of breast (N60-N64), non-inflammatory disorders of female genital tract (N80-N98), some diseases of the skin (L50-L54, L20-L30, L60-L75), acute and some other upper respiratory infections (J30-J39, J00-J06), injuries to the ankle and foot (S90-S99)
15 Cardiovascular and metabolic	females 56%	68	47%	Diseases of the circulatory system (all I codes, except pulmonary heart disease and diseases of pulmonary circulation (I26-I28), and diseases of arteries and veins (I70-I79, I 80-I89, I95-I99)), other disorders of the skin and subcutaneous tissue (L80-L99), influenza and pneumonia(J09-J18), metabolic disorder (E70-E90), disorders of thyroid gland (E00-E07), diabetes mellitus (E10-E14), other diseases of urinary system (N30-N39), respiratory and cardiovascular disorders specific to the perinatal period (P20-P29), malignant neoplasms of independent (primary) multiple sites (C97-C97), conditions involving the integument and temperature regulation of fetus and newborn (P80-P83)

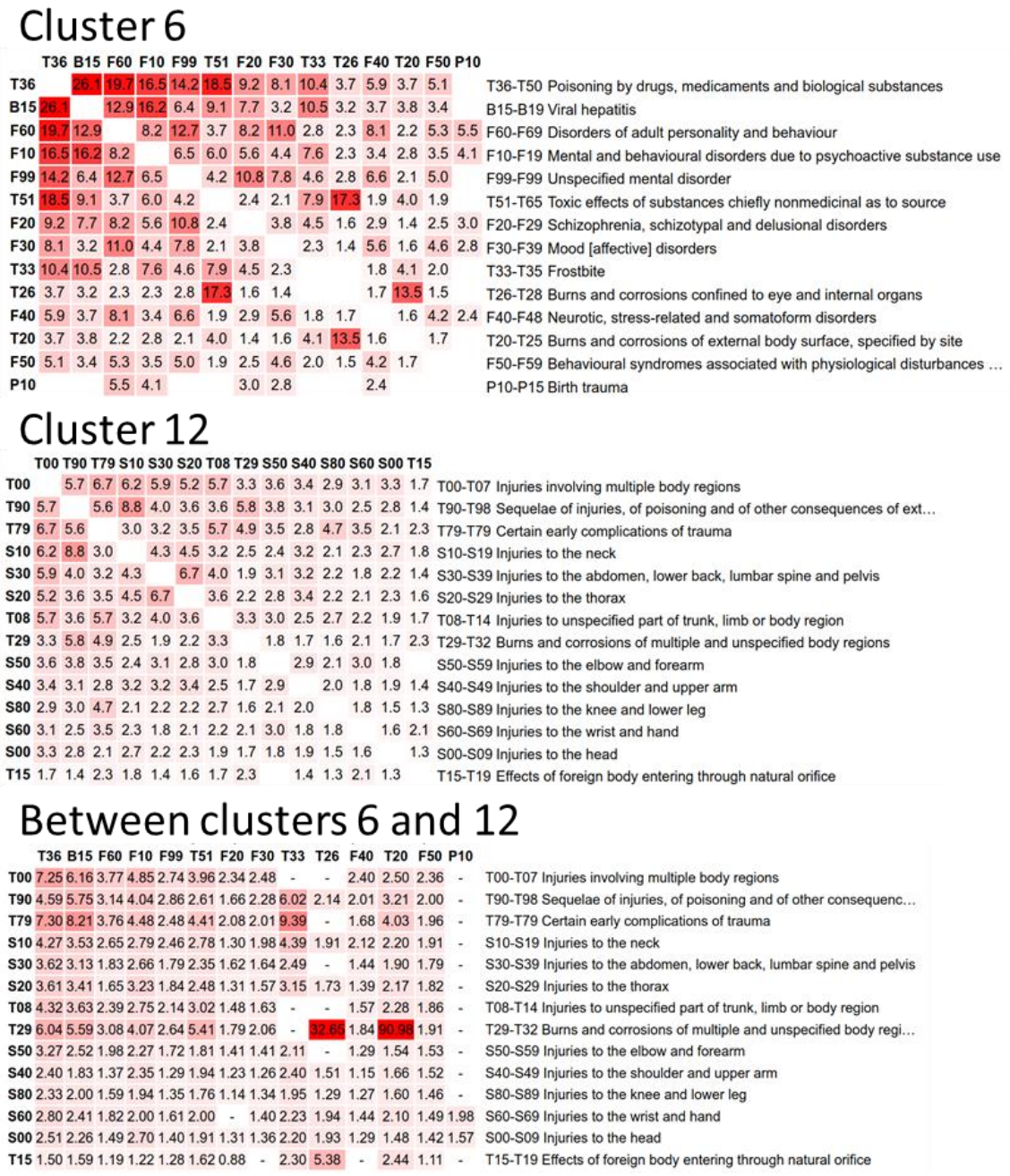

Fig. 6. Two example clusters and their connections in between. The numbers are RR-values. High values and red color signify stronger relationships. The blocks are represented by the first diagnosis code (e.g., T36 represents block T36–T50).

Although clustering captures many connections between the diseases, it does not capture all information. In fact, many interesting connections can be found by analyzing how strongly the clusters are connected to each other (see Fig. 6). Cluster 7 (nutritional problems) is the most central cluster, having strong connection to ten other clusters. Cluster 1 (pregnancy) is also connected to Cluster 6 (mental and behavioral disorders). For example, pregnancy with abortive outcome (O00–O08) has five connections with RR > 2 to Cluster 6 (mental and behavioral disorders), including neurotic, stress-related, mood disorders, and drug poisoning (T36–T50).

Cluster 12 (injuries) has strong connections to Clusters 6, 7, and 8. For example, the connection to the nutritional problems cluster has 56 links with RR > 2. Nine of these links come from the connection to other and unspecified disorders of the circulatory system (I95–I99).

Fig. 5 shows the connections between Clusters 6 and 12 in more detail. Cluster 6 consists of mental health (e.g., F30–F39, F60–F69) and substance abuse-related (T36–T50, F10–F19) diagnoses. Cluster 12 consists of fractures and other injuries. The clusters have a strong connection. A possible explanation is that mental health and substance abuse problems often lead to painful, fracture-causing accidents.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1 Pregnancy	-	16	14	3	9	**21**	0	1	13	6	1	1	4	**63**	10
2 Immune system and blood-forming organs	16	-	9	86	**122**	20	**129**	97	**100**	89	41	31	40	13	83
3 Mental disorders, malformations, ear and mouth	14	9	-	10	23	**35**	9	**30**	15	25	**32**	8	22	22	10
4 Tumors	3	**86**	10	-	**75**	3	**98**	63	54	26	23	2	31	6	68
5 Lower respiratory system	9	**122**	23	75	-	15	**115**	93	85	71	34	36	62	11	**96**
6 Mental and behavioral disorders	21	20	35	3	15	-	30	**36**	6	22	7	**89**	16	**54**	4
7 Nutritional	0	**129**	9	98	115	30	-	**132**	98	64	64	56	66	14	**123**
8 Diseases related to aging	1	**97**	30	63	93	36	**132**	-	58	53	46	72	57	1	**100**
9 Organ malformations and digestive system	13	**100**	15	54	**85**	6	**98**	58	-	28	14	8	39	9	58
10 Infections and inflammation	6	**89**	25	26	**71**	22	**64**	53	28	-	26	21	41	38	44
11 Eye and ear	1	41	32	23	34	7	**64**	**46**	14	26	-	7	30	19	**45**
12 Injuries	1	31	8	2	36	**89**	**56**	**72**	8	21	7	-	48	27	18
13 Musculoskeletal system	4	40	22	31	**62**	16	**66**	**57**	39	41	30	48	-	14	46
14 Sexually transmitted, parasitic, urinary tract	**63**	13	22	6	11	**54**	14	1	9	**38**	19	27	14	-	8
15 Cardiovascular and metabolic	10	83	10	68	**96**	4	**123**	**100**	58	44	45	18	46	8	-
Sum:	162	876	264	548	847	358	**998**	839	585	554	389	424	516	299	713

Fig. 7. Connections between clusters. Each cluster is represented in the rows with the number and description and in the columns with the number. Values in the table represent the number of links with RR > 2.0 between the clusters. Higher values signify a stronger connection and are emphasized by red color. Three clusters with the highest values for each row are highlighted with boldface font.

3.3 Cost effect

Costs of all visits, ward stays, and other contacts of patients belonging to the cluster were calculated for those contacts in services having a diagnosis belonging to the cluster. The estimated costs of each cluster are summarized in Table 6.

In general, the cost depends on the number of patients and the number of visits. The largest cluster (cardiovascular and metabolic Cluster 15) has 3.3 M visits and €2.3 B in total costs. However, the cost per patient (€2920) is not the highest and the cost per visit (€679) is only slightly above average. The diseases in the cluster, such as cardiovascular and metabolic disorders, are to a large extent treated in primary health care and thus the average visit cost remains relatively low.

Per patient, the highest costs are in Cluster 2 (€5,435) including infectious diseases strongly affecting the immune system, diseases of blood and blood-forming organs and other disorders involving the immune mechanism. These diseases are likely to need frequent contacts in specialized care. Per-patient costs were high also in Cluster 8 (diseases related to aging), including diagnoses of neurodegenerative diseases and memory disorders requiring frequent health care contacts and intensive care. The cheapest clusters per patient are Cluster 3 (Mental disorders, malformations, ear and mouth; €387) and Cluster 14 (Sexually transmitted, parasitic, urinary tract; €611). However, if dental diagnoses were removed, then the cost for Cluster 3 would be €1,144.

Per visit, the highest cost (€829) was in Cluster 9 including organ malformations and diseases of the digestive system. The second highest cost per visit was observed in Cluster 1 (pregnancy) where the cost per visit was €810. This is likely due to the delivery-related hospital stays, operations, and other specialized care. Regular maternity care visits are not usually recorded using ICD-10 coding. The clusters having the lowest cost per visit were the same as clusters having the lowest cost per patient.

Table 6 Estimated (annual) costs of each cluster. One patient can belong to multiple clusters. Visits and costs include only visits and related costs for diagnoses in cluster. Cost/visit is calculated as average for the whole 4-year period; all other values are annual

Cluster	Description	Patients	Total visits	Total cost (m€)	Cost/visit (€)	Cost/ptn (€)
1	Pregnancy	78 159	255 902	207	**810**	2 648
2	Immune system and blood-forming organs	95 865	653 500	521	**798**	**5 435**
3	Mental disorders, malformations, ear and mouth	838 208	1 899 209	324	171	**387**
4	Tumors	210 272	1 046 147	704	673	3 348
5	Lower respiratory system	299 482	953 199	620	651	2 070
6	Mental and behavioral disorders	280 450	2 094 496	908	434	3 238
7	Nutritional	194 250	708 930	525	741	2 703
8	Diseases related to aging	172 194	1 105 325	730	661	**4 239**
9	Organ malformations and digestive system	262 362	867 971	720	**829**	2 744
10	Infections and inflammation	359 738	1 110 728	627	564	1 743
11	Eye and ear	320 947	827 680	298	359	929
12	Injuries	324 191	720 282	417	579	1 286
13	Musculoskeletal system	616 550	1 704 486	836	490	1 356
14	Sexually transmitted, parasitic, urinary tract	474 604	955 465	290	303	**611**
15	Cardiovascular and metabolic	773 406	3 326 018	2 258	679	2 920
	Average	**353 378**	**1 215 289**	**666**	**583**	**2 377**

Table 8 shows how the costs of some clusters have developed during the years relative to the total cost of all clusters in the same year. Only clusters with a visible trend (increasing or decreasing) are shown. Clusters including tumors, lower respiratory system, and eye and ear have steadily increased their proportion of all costs from 2015 to 2018, as well as the cluster including inflammatory diseases and infections, among a few others. The diseases included in these clusters are increasing with age and thus the increase in costs is most likely due to the aging of the population.

Relative costs of mental and behavioral disorders have decreased most (from 9.5% to 8.8%) but also injuries (4.4% to 4.0%) and pregnancy related diseases (2.4% to 2.0%) show a clear decrease. There can be several explanations for the decline observed in costs of care related to mental and behavioral disorders, including the current tendency to prefer outpatient services and difficulties in appropriate service provision. The absolute cost values for pregnancy-related issues were €219 M, €213 M, €202 M, and €194 M during 2015 to 2018. The decrease is therefore real, which could be explained by the decrease in the birth rate from 1.65 to 1.41 during the same period (1.65, 1.57, 1.49, 1.41) [57].

Table 8 Trends of the annual costs (relative to all costs) of selected clusters from 2015 to 2018

	2015	2016	2017	2018
	Increasing trend			
Tumors	7.0 %	7.1 %	7.2 %	7.4 %
Mixed cluster 10	6.1 %	6.3 %	6.4 %	6.6 %
Lower respiratory system	6.1 %	6.2 %	6.4 %	6.4 %
Eye and ear	2.9 %	3.0 %	3.1 %	3.2 %
	Decreasing trend			
Mental and behavioral disorders	9.5 %	9.0 %	9.0 %	8.8 %

Mixed cluster 8	6.9 %	6.7 %	6.7 %	6.3 %
Injuries	4.4 %	4.2 %	4.2 %	4.0 %
Pregnancy	2.4 %	2.2 %	2.0 %	2.0 %

4. Discussion

4.1 Main findings

We analyzed the data by clustering the diagnoses to 15 clusters. All of the clusters where consistent with expert knowledge on the domain. Some of them were expected. For example, mental and behavioral disorders were so closely associated with substance abuse problems that they formed one cluster. Some of the clusters also showed interesting and unexpected connections such as cluster including lower respiratory tract diseases and systemic connective tissue disorders. Although some connections are easily justified by the close relation of the diagnoses, they are not necessarily considered when planning the current service processes and resources. For example, understanding the strong connections of many disorders related to aging could improve the treatment processes of elderly multimorbid patients.

Analyzing the connections between clusters also provided interesting details. For example, the mental health and substance abuse cluster was very closely connected to the cluster consisting of fractures and other injuries. A possible explanation is that mental health and substance abuse problems often lead to painful, fracture-causing accidents. The nutritional problems cluster was the most central in the data, having strong connection to ten other clusters. This is an interesting finding addressing the connection of nutritional status to various health disorders.

Per patient, the highest costs were in Cluster 2 (€5,435) which includes infectious diseases strongly affecting the immune system, diseases of blood and blood-forming organs and other disorders involving the immune mechanism. These diseases are likely to need frequent contacts in specialized care.

Clusters associated with aging population have increased their proportion of all costs from 2015 to 2018. These clusters include diseases related to tumors, lower respiratory system, and eye and ear. Relative costs of mental and behavioral disorders have decreased most (from 9.5% to 8.8%) which might be partly explained by the current tendency to prefer outpatient services.

4.2 Limitations

The underlying data reflects how patients are using health services and are diagnosed during the health care contacts, which may not always accurately reflect the true relation of diseases. For example, a person who visits the health services only for caries treatment may not be as easily diagnosed with alcohol-related disorders (F10) or problems related to metabolic disorders (E66) as a person who visits because of mental health issues or maternity.

The clustering methodology itself has a few limitations. Although the chosen clustering algorithm and cost function was shown to have a good clustering accuracy with validation data, it forces every diagnosis to belong to some cluster even if it does not have any connections to other diagnoses. A possible improvement could be to apply outlier detection as pre-processing to remove such cases.

Another limitation is that every diagnosis can belong to only one cluster, though it can be connected to diseases in several clusters. For example, dental health diagnoses are clustered with mental retardation and malformations but are clearly very relevant co-morbidities for other chronic conditions such as diabetes. Also, many infectious disease subgroups are likely to have significant connections with many chronic conditions that decrease the immune response, such as tumors.

Data might also be biased by domestic characteristics within Finnish population and traditions in recording diagnoses. For example, some conditions such as substance abuse disorders are still highly stigmatized and thus underdiagnosed. Then again, the research goal was exactly to find relevant multimorbidity diseases that have a high cost effect on the Finnish health care system. Even though some bias might exist, we expect most multimorbidity patterns to appear in other developed countries, and the main results might therefore be globally generalizable. This finding was partly confirmed by similar studies in the United States [58] and France [22].

Comparison to other clustering results in earlier studies was challenging mainly because there are many variations in the definition and measures of multimorbidity as well as the data sources, such as registers, health records, and self-reports, which have been used to obtain information on comorbidities. These differences make comparison difficult but still possible into some degree as shown in [18, 22].

4.3 Comparison with prior work

Comparison of clusters

Wartelle et al. [22] obtained 16 clusters (vs. 15 in our case). Some of them are similar to ours. For example, cluster 5 contains diagnoses related to mental disorders, substance abuse and fractures. In our results substance abuse and mental problems also formed one cluster which was closely connected to another cluster with different types of fractures. Their data also has one women specific cluster with pregnancy related diagnoses. However, mostly the clusters are very different from ours.

Their clusters are more unbalanced in size, five of the clusters contain only one diagnosis and the largest cluster has 13 diagnoses. In our case, the smallest cluster was size 13 and largest size 15. This is partly due to our choice of a clustering cost function that favors more balanced clusters, but also the choice of emergency department data in [22] is expected to generate larger clusters for trauma diagnoses.

Most of the differences originate from the data. Our data is everyday health care visits, while the data studied by Wartelle et al. [22] comes from emergency department visits (called ED). They have a smaller number of diagnoses (162 vs. 205). These include also symptom codes (R00-R99) and factors influencing health status (Z00-Z99) which we removed because we found them confusing the analysis. These data related factors produce several clear differenses in the results which we report below.

The first difference from [22] is that our data has female majority (54 %). We have only three clusters with more male than female visits (nutritional 67%, injuries 51%, immune system and blood-forming organs, 51%). ED data has ten clusters with male majority (52–64%). The likely explanation is that these clusters are either directly or indirectly related to trauma commonly treated in emergency departments, whereas our data represent the service used in primary health care, which has only one cluster (12) related to injuries.

Patients in ED data are also much younger than in our data (mean age 40 vs. 51 years). There are three clusters where the average age of patients exceeds 50 years. One cluster mostly (~50%) consisting of children younger than 5 years. Our data was restricted to adult patients. ED data also lacks a clear pregnancy cluster and pregnancy related diagnoses are merged with digestive and menstruation related diagnoses.

Busija et. al. [60] conducted a meta-analysis study by investigating 51 different articles of multimorbidity profiles. They constructed a similarity matrix of health conditions by counting the number of times each pair of diseases appeared in the same group. The similarity matrix was then projected to 2D surface by using multidimensional scaling (SPSS/PROXSCAL). This was done separately to 4 different types of studies grouped by methodology: exploratory factor analysis, cluster analysis of diseases, latent class analysis and cluster analysis of people.

Overall, their data had less diagnoses and clusters. The largest case (factor analysis) included only 70 diagnoses and they manually distinguished 5 clusters (and a group of mental health problems as one axis) from the 2D projection. They reported the clustering of vision, hearing impairment and fractures in two of the four cases. In our data, vision and hearing problems were in one cluster and fractures in another. These were also only weakly connected. A mental health group was visible in all of the four cases, and it was closely associated with addictions. This is consistent with our results where mental health and substance abuse problems formed one cluster.

Comparison of costs

We compared the cost of our data to those reported by Milken Institute in the United States in 2016 [58]. The costliest (both direct and indirect costs) chronic disease in United States is Diabetes Type 2 with the direct costs of $185 B. When indirect costs are included the four most costly diseases were hypertension ($1,042 B), diabetes type 2 ($526 B), chronic back pain ($440 B), and osteoarthritis ($430 B).

The costliest diseases (hypertension and Type 2 diabetes) are in accordance with our results where the costliest is the Cluster 15 (cardiovascular and metabolic), which includes also hypertension and diabetes-related diagnoses (I10–I15, E10–E14) as well as other related cardiovascular diseases common in Finnish population. The costs of the cluster become high as the patient population is large as well as the need for frequent contacts to health care even though costs per visit are close to average.

4.4 Conclusions

To the best of our knowledge, this is the first clustering study having such a rich data set including all health care visits of Finnish adult population aged 18 years or older covering both primary and secondary-level care. The good coverage is important as the tendency in development of health service systems is to seek better integration of services, including the integration of primary health care, specialized care, and social services.

Identifying multimorbidity clusters, related characteristics and especially the burden they cause for the service utilization and costs is helpful in estimating the resources needed in the service system including the specialties and other knowledge profiles of professionals. Such information could also be applied in estimating future needs when for example the projections of population aging, and other demographics are known.

To the best of our knowledge, this is also the first study using k-means based clustering of diseases. While the standard k-means algorithm can be unstable, we used a recent modification called M-algorithm which was shown to be accurate on controlled validation datasets. It directly optimizes for relative risk. Existing studies rely mainly on agglomerative clustering, either using a heuristic cost function such as average or complete linkage, or a slow calculation of the relative risk. The used methodology is accurate and scalable to large-scale data.

As a future study, we will consider clustering the patients and comparing whether the same diagnoses will group together. Another future idea is to study geographical differences within Finland. The data itself is large, and as it is publicly available, it has high potential for others to find more interesting results by data mining.

Conflict of interests

None.

Author contributions

Dr. Sieranoja and Prof. Fränti developed the analytical methods. Prof. Laatikainen (MD) and Dr. Wikström performed the data acquisition and medical interpretation, and provided guidance related to health service system. Prof Fränti and Dr. Sieranoja drafted the manuscript. All authors contributed to the writing and editing of the manuscript and approved the final version.

Acknowledgments

The project was funded by the Strategic Research Council (SRC) at the Academy of Finland (grant numbers 312703, 312706 and 336325). We thank Prof. Miika Linna for help with the cost estimations.

References

[1] S. H. Van Oostrom, H. S. J. Picavet, S. R. De Bruin, I. Stirbu, J. C. Korevaar, F. G. Schellevis, and C. A. Baan, "Multimorbidity of chronic diseases and health care utilization in general practice," BMC family practice, vol. 15, no. 1, pp. 1–9, 2014.

[2] M. van den Akker, F. Buntinx, and J. A. Knottnerus, "Comorbidity or multimorbidity: what's in a name? a review of literature," The European journal of general practice, vol. 2, no. 2, pp. 65–70, 1996.

[3] M. Van den Akker, F. Buntinx, J. F. Metsemakers, S. Roos, and J. A. Knottnerus, "Multimorbidity in general practice: prevalence, incidence, and determinants of co-occurring chronic and recurrent diseases," Journal of clinical epidemiology, vol. 51, no. 5, pp. 367–375, 1998.

[4] T. G. Willadsen, A. Bebe, R. Køster-Rasmussen, D. E. Jarbøl, A. D. Guassora, F. B. Waldorff, S. Reventlow, and N. d. F. Olivarius, "The role of diseases, risk factors and symptoms in the definition of multimorbidity–a systematic review," Scandinavian journal of primary health care, vol. 34, no. 2, pp. 112–121, 2016.

[5] X. Xu, G. D. Mishra, and M. Jones, "Evidence on multimorbidity from definition to intervention: an overview of systematic reviews," Ageing research reviews, vol. 37, pp. 53–68, 2017.

[6] L. Wang, L. Si, F. Cocker, A. J. Palmer, and K. Sanderson, "A systematic review of cost-of-illness studies of multimorbidity," Applied health economics and health policy, vol. 16, no. 1, pp. 15–29, 2018.

[7] C. Brettschneider, H. Leicht, H. Bickel, A. Dahlhaus, A. Fuchs, J. Gensichen, W. Maier, S. Riedel-Heller, I. Schäfer, G. Schön, et al., "Relative impact of multimorbid chronic conditions on health-related quality of life–results from the multicare cohort study," PloS one, vol. 8, no. 6, p. e66742, 2013.

[8] L. E. Stirland, L. González-Saavedra, D. S. Mullin, C. W. Ritchie, G. Muniz-Terrera, and T. C. Russ, "Measuring multimorbidity beyond counting diseases: systematic review of community and population studies and guide to index choice," Bmj, vol. 368, 2020.

[9] D. A. Travers, S. W. Haas, A. E. Waller, and J. E. Tintinalli, "Diagnosis clusters for emergency medicine," Academic emergency medicine, vol. 10, no. 12, pp. 1337–1344, 2003.

[10] R. Schneeweiss, D. C. Cherkin, L. G. Hart, et al., "Diagnosis clusters adapted for icd-9-cm and ichppc-2," J Fam Pract, vol. 22, no. 1, pp. 69–72, 1986.

[11] J. F. Farley, C. R. Harley, and J. W. Devine, "A comparison of comorbidity measurements to predict healthcare expenditures," American Journal of Managed Care, vol. 12, no. 2, pp. 110–118, 2006.

[12] H. Estiri, J. G. Klann, and S. N. Murphy, "A clustering approach for detecting implausible observation values in electronic health records data," BMC medical informatics and decision making, vol. 19, no. 1, pp. 1–16, 2019.

[13] L. Huang, A. L. Shea, H. Qian, A. Masurkar, H. Deng, and D. Liu, "Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records," Journal of biomedical informatics, vol. 99, p. 103291, 2019.

[14] P. Kalgotra, R. Sharda, and J. M. Croff, “Examining health disparities by gender: A multimorbidity network analysis of electronic medical record,” International journal of medical informatics, vol. 108, pp. 22–28, 2017.

[15] F. Folino, C. Pizzuti, and M. Ventura, “A comorbidity network approach to predict disease risk,” in Information Technology in Bio-and Medical Informatics, ITBAM 2010, pp. 102–109, Springer, 2010.

[16] F. Folino and C. Pizzuti, “Link prediction approaches for disease networks,” in International Conference on Information Technology in Bio-and Medical Informatics, pp. 99–108, Springer, 2012.

[17] R. Ding, F. Jiang, J. Xie, and Y. Yu, “Algorithmic prediction of individual diseases,” International Journal of Production Research, vol. 55, no. 3, pp. 750–768, 2017.

[18] A. Prados-Torres, A. Calderón-Larrañaga, J. Hancco-Saavedra, B. Poblador-Plou, and M. van den Akker, “Multimorbidity patterns: a systematic review,” Journal of clinical epidemiology, vol. 67, no. 3, pp. 254–266, 2014.

[19] R. John, D. S. Kerby, and C. Hagan Hennessy, “Patterns and impact of comorbidity and multimorbidity among community-resident american indian elders,” The Gerontologist, vol. 43, no. 5, pp. 649–660, 2003.

[20] J. E. Cornell, J. A. Pugh, J. W. Williams Jr, L. Kazis, A. F. Lee, M. L. Parchman, J. Zeber, T. Pederson, K. A. Montgomery, and P. H. Noël, “Multimorbidity clusters: clustering binary data from multimorbidity clusters: clustering binary data from a large administrative medical database,” Applied multivariate research, vol. 12, no. 3, pp. 163–182, 2008.

[21] A. Marengoni, F. Bonometti, A. Nobili, et al., “In-hospital death and adverse clinical events in elderly patients according to disease clustering: the reposi study,” Rejuvenation research, vol. 13, no. 4, pp. 469–477, 2010.

[22] A. Wartelle, F. Mourad-Chehade, F. Yalaoui, J. Chrusciel, D. Laplanche, and S. Sanchez, “Clustering of a health dataset using diagnosis co-occurrences,” Applied Sciences, vol. 11, no. 5, p. 2373, 2021.

[23] C. Violán, A. Roso-Llorach, Q. Foguet-Boreu, et al., “Multimorbidity patterns with k-means nonhierarchical cluster analysis,” BMC family practice, vol. 19, no. 1, pp. 1–11, 2018.

[24] P. Fränti and S. Sieranoja, “How much can k-means be improved by using better initialization and repeats?,” Pattern Recognition, vol. 93, pp. 95–112, 2019.

[25] S. Sieranoja, P. Fränti, “Adapting k-means for graph clustering,“ Knowledge and Information Systems (KAIS), DOI: 10.1007/s10115-021-01623-y (accepted for publication)

[26] C. J. Whitty and F. M. Watt, “Map clusters of diseases to tackle multimorbidity,” Nature, vol. 579, pp. 494-496, 2020.

[27] M. E. Newman, “Modularity and community structure in networks,” Proceedings of the national academy of sciences, vol. 103, no. 23, pp. 8577–8582, 2006.

[28] M. E. Newman, “Analysis of weighted networks,” Physical review E, vol. 70, no. 5, p. 056131, 2004.

[29] B. W. Kernighan and S. Lin, “An efficient heuristic procedure for partitioning graphs,” Bell system technical journal, vol. 49, no. 2, pp. 291–307, 1970.

[30] J. Shi and J. Malik, “Normalized cuts and image segmentation,” IEEE Transactions on pattern analysis and machine intelligence, vol. 22, no. 8, pp. 888–905, 2000.

[31] M. J. Divo, C. Casanova, J. M. Marin, V. M. Pinto-Plata, J. P. de Torres, J. J. Zulueta, C. Cabrera, J. Zagaceta, P. Sanchez-Salcedo, J. Berto, et al., “Chronic obstructive pulmonary disease comorbidities network,” European Respiratory Journal, pp. ERJ–01716, 2015.

[32] H. Hromic, N. Prangnawarat, I. Hulpuș, M. Karnstedt, and C. Hayes, “Graph-based methods for clustering topics of interest in twitter,” in International Conference on Web Engineering, pp. 701–704, Springer, 2015.

[33] S. Fortunato, “Community detection in graphs,” Physics reports, vol. 486, no. 3-5, pp. 75–174, 2010.

[34] S. Fortunato and D. Hric, “Community detection in networks: A user guide,” Physics reports, vol. 659, pp. 1–44, 2016.

[35] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” Journal of statistical mechanics: theory and experiment, vol. 2008, no. 10, p. P10008, 2008.

[36] A. Lancichinetti and S. Fortunato, “Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities,” Physical Review E, vol. 80, no. 1, p. 016118, 2009.

[37] J. J. Whang, D. F. Gleich, and I. S. Dhillon, “Overlapping community detection using neighborhood-inflated seed expansion,” IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 5, pp. 1272–1284, 2016.

[38] Z. Lu, Y. Wen, and G. Cao, “Community detection in weighted networks: Algorithms and applications,” in 2013 IEEE International Conference on Pervasive Computing and Communications (PerCom), pp. 179–184, IEEE, 2013.

[39] A. Lancichinetti and S. Fortunato, “Community detection algorithms: a comparative analysis,” Physical review E, vol. 80, no. 5, p. 056117, 2009.

[40] W. Zhang, X. Wang, D. Zhao, and X. Tang, “Graph degree linkage: Agglomerative clustering on a directed graph,” in European Conference on Computer Vision, pp. 428–441, Springer, 2012.

[41] D. LaSalle and G. Karypis, “A parallel hill-climbing refinement algorithm for graph partitioning,” in 2016 45th International Conference on Parallel Processing (ICPP), pp. 236–241, IEEE, 2016.

[42] S. S. Tabatabaei, M. Coates, and M. Rabbat, “Ganc: Greedy agglomerative normalized cut for graph clustering,” Pattern Recognition, vol. 45, no. 2, pp. 831–843, 2012.

[43] E. Mustonen, I. Hörhammer, P. Absetz, K. Patja, J. Lammintakanen, M. Talja, R. Kuronen, and M. Linna, “Eight-year post-trial follow-up of health care and long-term care costs of tele-based health coaching,” Health services research, vol. 55, no. 2, pp. 211–217, 2020.

[44] M. Linna, T. Mikkola, A. Peltokorpi, T. Tyni, “Rekistereistä tietoa vanhuspalvelujen johtamiseen?, Ikääntyneen väestön sosiaali- ja ter-veyspalveluiden käytön arviointi rekisteriaineis-toja hyödyntämällä,“ Suomen Kunta-liitto, 2016. (in Finnish)

[45] S. Kapiainen, A. Väisänen, and T. Haula, “Terveyden-ja sosiaalihuollon yksikkökustannukset suomessa vuonna 2011,” 2014.

http://www.julkari.fi/bitstream/handle/10024/114683/THL_RAPO3_2014_web.pdf?sequence=1&isAllowed=y. Accessed 10 May 2019.

[46] C. A. Hidalgo, N. Blumm, A.-L. Barabási, and N. A. Christakis, “A dynamic network approach for the study of human phenotypes,” PLoS computational biology, vol. 5, no. 4, p. e1000353, 2009.

[47] K. Srinivasan, F. Currim, and S. Ram, “Predicting high-cost patients at point of admission using network science,” IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 6, pp. 1970–1977, 2018.

[48] S. Brin, R. Motwani, and C. Silverstein, “Beyond market baskets: Generalizing association rules to correlations,” Acm Sigmod Record, vol. 26, no. 2, pp. 265–276, 1997.

[49] M. A. Moni and P. Liò, “comor: a software for disease comorbidity risk assessment,” Journal of clinical bioinformatics, vol. 4, no. 1, p. 8, 2014.

[50] P. Klimek, A. Kautzky-Willer, A. Chmiel, I. Schiller-Frühwirth, and S. Thurner, “Quantification of diabetes comorbidity risks across life using nation-wide big claims data,” PLoS computational biology, vol. 11, no. 4, p. e1004125, 2015.

[51] https://web.archive.org/web/20160426043824/http://hudine.neu.edu/

[52] M. Bastian, S. Heymann, and M. Jacomy, “Gephi: an open source software for exploring and manipulating networks,” in Third international AAAI conference on weblogs and social media, 2009.

[53] Q. Zhao and P. Fränti, “Wb-index: A sum-of-squares based index for cluster validity,” Data & Knowledge Engineering, vol. 92, pp. 77–89, 2014.

[54] P. Corrigan, “How stigma interferes with mental health care,” American psychologist, vol. 59, no. 7, p. 614, 2004.

[55] Statistics Finland, “Findicator - Mortality from ischaemic heart disease,“ 2020. https://findikaattori.fi/en/83

[56] T. Karolaakso, R. Autio, T. Näppilä, K. Nurmela, and S. Pirkola, “Socioeconomic factors in disability retirement due to mental disorders in finland,” European journal of public health, vol. 30, no. 6, pp. 1218–1224, 2020.

[57] Statistics Finland, “Official Statistics of Finland (OSF): Births,“ [e-publication], ISSN=1798-2413, 2019. [referred: 14.6.2021]. http://www.stat.fi/til/synt/2019/synt_2019_2020-04-24_tie_001_en.html

[58] H. Waters and M. Graf, “The costs of chronic disease in the us,” Santa Monica, CA: The Milken Institute, 2018. https://milkeninstitute.org/sites/default/files/reports-pdf/ChronicDiseases-HighRes-FINAL.pdf

[59] B. Rozemberczki, R. Davies, R. Sarkar, C. Sutton, “Gemsec: graph embedding with self clustering”, Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining, pp 65–72, 2019.

[60] L. Busija, K- Lim, C. Szoeke, K.M. Sanders, M.P. McCabe, “Do replicable profiles of multimorbidity exist? Systematic review and synthesis.” European Journal of Epidemiology, vol 34, pp. 1025-1053, 2019.

[61] A. Amritphale, G.C. Fonarow, N. Amritphale, B. Omar, E.D. Crook, “All-Cause Unplanned Readmissions in the United States. Insights from the Nationwide Readmission Database.” Internal Medicine Journal, 2021.

[62] N. Amritphale. A. Amritphale, D. Vasireddy, M. Bantra, M. Sehgal, D. Gremse, “Age-and Diagnosis-Based Trends for Unplanned Pediatric Rehospitalizations in the United States.”, Cureus 13(12), 2021.

[63] K. Moons et al., “Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration.” Annals of internal medicine 162(1), W1-W73, 2015.

[64] M. Zoubi, M. Rawi, An efficient approach for computing silhouette coefficients, J. Comput. Sci. 4 (3) 252–255, 2008.

[65] A.J. Dunning, J. Kensler, L. Coudeville and F. Bailleux, "Some extensions in continuous models for immunological correlates of protection," BMC medical research methodology, 15(1), pp.1-11, 2015.

[66] A. Aguado, F. Moratalla-Navarro, F. López-Simarro and V. Moreno, "MorbiNet: multimorbidity networks in adult general population. Analysis of type 2 diabetes mellitus comorbidity," Scientific reports, 10(1), pp.1-12, 2020.

Appendix I: ICD-10 blocks

ICD-10 block	Description
A00-A09	Intestinal infectious diseases
A15-A19	Tuberculosis
A20-A28	Certain zoonotic bacterial diseases
A30-A49	Other bacterial diseases
A50-A64	Infections with a predominantly sexual mode of transmission
A65-A69	Other spirochaetal diseases
A70-A74	Other diseases caused by chlamydiae
A75-A79	Rickettsioses
A80-A89	Viral infections of the central nervous system
A90-A99	Arthropod-borne viral fevers and viral haemorrhagic fevers
B00-B09	Viral infections characterized by skin and mucous membrane lesions
B15-B19	Viral hepatitis
B20-B24	Human immunodeficiency virus [HIV] disease
B25-B34	Other viral diseases
B35-B49	Mycoses
B50-B64	Protozoal diseases
B65-B83	Helminthiases
B85-B89	Pediculosis, acariasis and other infestations
B90-B94	Sequelae of infectious and parasitic diseases
B95-B98	Bacterial, viral and other infectious agents
B99-B99	Other infectious diseases
C00-C14	Malignant neoplasms of lip, oral cavity and pharynx
C15-C26	Malignant neoplasms of digestive organs
C30-C39	Malignant neoplasms of respiratory and intrathoracic organs
C40-C41	Malignant neoplasms of bone and articular cartilage
C43-C44	Melanoma and other malignant neoplasms of skin
C45-C49	Malignant neoplasms of mesothelial and soft tissue
C50-C50	Malignant neoplasm of breast
C51-C58	Malignant neoplasms of female genital organs
C60-C63	Malignant neoplasms of male genital organs
C64-C68	Malignant neoplasms of urinary tract
C69-C72	Malignant neoplasms of eye, brain and other parts of central nervous system
C73-C75	Malignant neoplasms of thyroid and other endocrine glands
C76-C80	Malignant neoplasms of ill-defined, secondary and unspecified sites
C81-C96	Malignant neoplasms, stated or presumed to be primary, of lymphoid, haematopoietic and related tissue
C97-C97	Malignant neoplasms of independent (primary) multiple sites
D00-D09	In situ neoplasms
D10-D36	Benign neoplasms
D37-D48	Neoplasms of uncertain or unknown behaviour
D50-D53	Nutritional anaemias
D55-D59	Haemolytic anaemias
D60-D64	Aplastic and other anaemias
D65-D69	Coagulation defects, purpura and other haemorrhagic conditions
D70-D77	Other diseases of blood and blood-forming organs
D80-D89	Certain disorders involving the immune mechanism
E00-E07	Disorders of thyroid gland
E10-E14	Diabetes mellitus
E15-E16	Other disorders of glucose regulation and pancreatic internal secretion
E20-E35	Disorders of other endocrine glands
E40-E46	Malnutrition
E50-E64	Other nutritional deficiencies
E65-E68	Obesity and other hyperalimentation
E70-E90	Metabolic disorders

F00-F09	Organic, including symptomatic, mental disorders
F10-F19	Mental and behavioural disorders due to psychoactive substance use
F20-F29	Schizophrenia, schizotypal and delusional disorders
F30-F39	Mood [affective] disorders
F40-F48	Neurotic, stress-related and somatoform disorders
F50-F59	Behavioural syndromes associated with physiological disturbances and physical factors
F60-F69	Disorders of adult personality and behaviour
F70-F79	Mental retardation
F80-F89	Disorders of psychological development
F90-F98	Behavioural and emotional disorders with onset usually occurring in childhood and adolescence
F99-F99	Unspecified mental disorder
G00-G09	Inflammatory diseases of the central nervous system
G10-G14	Systemic atrophies primarily affecting the central nervous system
G20-G26	Extrapyramidal and movement disorders
G30-G32	Other degenerative diseases of the nervous system
G35-G37	Demyelinating diseases of the central nervous system
G40-G47	Episodic and paroxysmal disorders
G50-G59	Nerve, nerve root and plexus disorders
G60-G64	Polyneuropathies and other disorders of the peripheral nervous system
G70-G73	Diseases of myoneural junction and muscle
G80-G83	Cerebral palsy and other paralytic syndromes
G90-G99	Other disorders of the nervous system
H00-H06	Disorders of eyelid, lacrimal system and orbit
H10-H13	Disorders of conjunctiva
H15-H22	Disorders of sclera, cornea, iris and ciliary body
H25-H28	Disorders of lens
H30-H36	Disorders of choroid and retina
H40-H42	Glaucoma
H43-H45	Disorders of vitreous body and globe
H46-H48	Disorders of optic nerve and visual pathways
H49-H52	Disorders of ocular muscles, binocular movement, accommodation and refraction
H53-H54	Visual disturbances and blindness
H55-H59	Other disorders of eye and adnexa
H60-H62	Diseases of external ear
H65-H75	Diseases of middle ear and mastoid
H80-H83	Diseases of inner ear
H90-H95	Other disorders of ear
I00-I02	Acute rheumatic fever
I05-I09	Chronic rheumatic heart diseases
I10-I15	Hypertensive diseases
I20-I25	Ischaemic heart diseases
I26-I28	Pulmonary heart disease and diseases of pulmonary circulation
I30-I52	Other forms of heart disease
I60-I69	Cerebrovascular diseases
I70-I79	Diseases of arteries, arterioles and capillaries
I80-I89	Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified
I95-I99	Other and unspecified disorders of the circulatory system
J00-J06	Acute upper respiratory infections
J09-J18	Influenza and pneumonia
J20-J22	Other acute lower respiratory infections
J30-J39	Other diseases of upper respiratory tract
J40-J47	Chronic lower respiratory diseases
J60-J70	Lung diseases due to external agents
J80-J84	Other respiratory diseases principally affecting the interstitium
J85-J86	Suppurative and necrotic conditions of lower respiratory tract
J90-J94	Other diseases of pleura
J95-J99	Other diseases of the respiratory system
K00-K14	Diseases of oral cavity, salivary glands and jaws

K20-K31	Diseases of oesophagus, stomach and duodenum
K35-K38	Diseases of appendix
K40-K46	Hernia
K50-K52	Noninfective enteritis and colitis
K55-K64	Other diseases of intestines
K65-K67	Diseases of peritoneum
K70-K77	Diseases of liver
K80-K87	Disorders of gallbladder, biliary tract and pancreas
K90-K93	Other diseases of the digestive system
L00-L08	Infections of the skin and subcutaneous tissue
L10-L14	Bullous disorders
L20-L30	Dermatitis and eczema
L40-L45	Papulosquamous disorders
L50-L54	Urticaria and erythema
L55-L59	Radiation-related disorders of the skin and subcutaneous tissue
L60-L75	Disorders of skin appendages
L80-L99	Other disorders of the skin and subcutaneous tissue
M00-M03	Infectious arthropathies
M05-M14	Inflammatory polyarthropathies
M15-M19	Arthrosis
M20-M25	Other joint disorders
M30-M36	Systemic connective tissue disorders
M40-M43	Deforming dorsopathies
M45-M49	Spondylopathies
M50-M54	Other dorsopathies
M60-M63	Disorders of muscles
M65-M68	Disorders of synovium and tendon
M70-M79	Other soft tissue disorders
M80-M85	Disorders of bone density and structure
M86-M90	Other osteopathies
M91-M94	Chondropathies
M95-M99	Other disorders of the musculoskeletal system and connective tissue
N00-N08	Glomerular diseases
N10-N16	Renal tubulo-interstitial diseases
N17-N19	Renal failure
N20-N23	Urolithiasis
N25-N29	Other disorders of kidney and ureter
N30-N39	Other diseases of urinary system
N40-N51	Diseases of male genital organs
N60-N64	Disorders of breast
N70-N77	Inflammatory diseases of female pelvic organs
N80-N98	Noninflammatory disorders of female genital tract
N99-N99	Other disorders of the genitourinary system
O00-O08	Pregnancy with abortive outcome
O10-O16	Oedema, proteinuria and hypertensive disorders in pregnancy, childbirth and the puerperium
O20-O29	Other maternal disorders predominantly related to pregnancy
O30-O48	Maternal care related to the fetus and amniotic cavity and possible delivery problems
O60-O75	Complications of labour and delivery
O80-O84	Delivery
O85-O92	Complications predominantly related to the puerperium
O94-O99	Other obstetric conditions, not elsewhere classified
P00-P04	Fetus and newborn affected by maternal factors and by complications of pregnancy, labour and delivery
P05-P08	Disorders related to length of gestation and fetal growth
P10-P15	Birth trauma
P20-P29	Respiratory and cardiovascular disorders specific to the perinatal period
P35-P39	Infections specific to the perinatal period
P50-P61	Haemorrhagic and haematological disorders of fetus and newborn
P70-P74	Transitory endocrine and metabolic disorders specific to fetus and newborn
P80-P83	Conditions involving the integument and temperature regulation of fetus and newborn
P90-P96	Other disorders originating in the perinatal period
Q00-Q07	Congenital malformations of the nervous system

Q10-Q18	Congenital malformations of eye, ear, face and neck
Q20-Q28	Congenital malformations of the circulatory system
Q30-Q34	Congenital malformations of the respiratory system
Q35-Q37	Cleft lip and cleft palate
Q38-Q45	Other congenital malformations of the digestive system
Q50-Q56	Congenital malformations of genital organs
Q60-Q64	Congenital malformations of the urinary system
Q65-Q79	Congenital malformations and deformations of the musculoskeletal system
Q80-Q89	Other congenital malformations
Q90-Q99	Chromosomal abnormalities, not elsewhere classified
S00-S09	Injuries to the head
S10-S19	Injuries to the neck
S20-S29	Injuries to the thorax
S30-S39	Injuries to the abdomen, lower back, lumbar spine and pelvis
S40-S49	Injuries to the shoulder and upper arm
S50-S59	Injuries to the elbow and forearm
S60-S69	Injuries to the wrist and hand
S70-S79	Injuries to the hip and thigh
S80-S89	Injuries to the knee and lower leg
S90-S99	Injuries to the ankle and foot
T00-T07	Injuries involving multiple body regions
T08-T14	Injuries to unspecified part of trunk, limb or body region
T15-T19	Effects of foreign body entering through natural orifice
T20-T25	Burns and corrosions of external body surface, specified by site
T26-T28	Burns and corrosions confined to eye and internal organs
T29-T32	Burns and corrosions of multiple and unspecified body regions
T33-T35	Frostbite
T36-T50	Poisoning by drugs, medicaments and biological substances
T51-T65	Toxic effects of substances chiefly nonmedicinal as to source
T66-T78	Other and unspecified effects of external causes
T79-T79	Certain early complications of trauma
T80-T88	Complications of surgical and medical care, not elsewhere classified
T90-T98	Sequelae of injuries, of poisoning and of other consequences of external causes