Crisafulli et al. *EPJ Data Science* (2026) 15:50
https://doi.org/10.1140/epjds/s13688-026-00651-4

EPJ.org

EPJ Data Science

RESEARCH

Open Access

Large scale statistically validated comorbidity networks

Paride Crisafulli[1], Tobias Galla[1], Antti Karlsson[2], Salvatore Micciché[3], Jyrki Piilo[4] and Rosario N. Mantegna[3,5*]

Handling Editor: Eugenio Valdano

*Correspondence:
rosario.mantegna@unipa.it
[3]Dipartimento di Fisica e Chimica Emilio Segrè, Università degli Studi di Palermo, Palermo, Italy
[5]Complexity Science Hub, Vienna, Austria
Full list of author information is available at the end of the article

Abstract

We obtain comorbidity networks starting from medical information stored in electronic health records collected by the Wellbeing Services County of Southwest Finland (Varha). Based on the data, we connect each patient to one or more diseases and construct complex comorbidity networks associated with large patient cohorts characterized by an age interval and sex. The information about diseases in electronic health records is coded using the highest granularity present in the international classification of diseases (ICD codes) provided by the World Health Organization. We statistically validate links in each cohort's comorbidity network and furthermore partition the networks into communities of diseases. These are characterized by the over-expression of a few disease categories, and communities from different age or sex cohorts show various similarities in terms of these disease classes. Moreover, the detected communities for all the cohorts can be organized into a hierarchical tree. This allows us to observe a number of clusters of communities — originating from diverse age and sex cohorts — that group together communities characterized by the same disease classes. We also perform a dismantling procedure of statistically validated comorbidity networks to highlight those categories of diseases that are most responsible for the compactedness of the comorbidity networks for a given cohort of patients.

Keywords: Electronic Health Records; Comorbidity; Complex networks; Statistically Validated Networks

1 Introduction

Network modeling is a powerful and flexible tool for the investigation of complex systems. A complex network is a system composed of interconnected elements whose interactions give rise to non-trivial structural and dynamic properties [1–3]. Distinguished from simple or regular networks, complex networks often exhibit heterogeneous degree distributions, clustering, and modularity, reflecting real-world behaviors found in biological, social, and technological systems. Key characteristics include small-world behavior, where most nodes can be reached through relatively short paths, and scale-free topology, in which a few highly connected nodes (hubs) dominate connectivity patterns. These emergent properties enable complex networks to demonstrate both robustness and vul-

Springer

nerability, depending on the configuration of their nodes and links, making them a central focus in the study of distributed systems, resilience, and collective dynamics [1–3].

The investigation of complex networks of biological and/or medical interest has led to research fields referred to as 'network medicine' [4] and 'network physiology' [5]. One line of research of network medicine concerns comorbidity networks.

Comorbidity networks are networks of diseases (or conditions – we will use both terms synonymously). Nodes represent diseases, and the presence of a link between a pair of nodes indicates the co-occurrence of both diseases in the medical history of a patient. Comorbidity networks can be constructed from information in electronic health records (EHRs) of large health organizations providing health services to patients of regions or countries.

EHRs originating from the hospitalization process are standardized in most of the countries. Diseases are coded using the international classification of diseases (ICDs) published by the World Health Organization (WHO). This classification is used worldwide to estimate morbidity and mortality statistics. The classification is periodically updated, and it is aimed at facilitating international comparability. The current version is ICD-11. Other versions that are still being used are ICD-9 and ICD-10 [6].

Since the early work investigating Medicare claims [7, 8], comorbidity networks have been investigated by using small [9, 10] and large [11–13] sets of EHRs of different regional areas and countries. A comparative study of comorbidity patterns in China and the UK has been reported recently [14]. Comorbidity networks are often obtained for specific cohorts of patients; typical stratification is in terms of age and sex. Results for specific cohorts of patients can contribute to devising cohort-specific medical protocols.

In the present study, we analyze large comorbidity networks starting from the dataset of EHRs collected by the Wellbeing Services County of Southwest Finland (Varha) [15]. These data can be accessed for research purposes via Auria Clinical Informatics [16]. Auria Clinical Informatics is a company situated in Turku, Finland, interacting with Turku University Hospital and the University of Turku and operating under license from Valvira, the National Supervisory Authority for Welfare and Health in Finland.

Previous studies of comorbidity networks [8–14, 17] have been performed by considering ICD classification at so-called "level-3" (i.e., a classification level with a set of 3 characters as, for example I10). Here we investigate different cohorts of EHR of Finnish patients by using a ICD classification at "level-4" (i.e., with a resolution of four characters as, for example F32.2). Level-3 codes are used for high-level categorization in clinical diagnosis or health statistics while a level-4 classification provides a more specific diagnosis, often used by healthcare providers and researchers to determine exact treatments and prognosis. Level-4 classification is therefore more detailed and closer to the practice of medical doctors. By moving from level-3 to level-4 we improve the resolution of the comorbidity network description. This choice allows us to analyze each disease at the highest level of description that is present in medical records. Further details can be found in Sect. 2.2.

In our analysis, we start from a bipartite patient-disease network in which all hospital diagnosis are recorded as links between a patient and the conditions the patient has been diagnosed with. We then obtain a projected network comprising of only disease nodes, each representing one level-4 ICD code. Two ICD codes are connected in this network if there is at least one patient in the target cohort who has been diagnosed with both conditions. We refer to this network as 'PROJ' (for projected network).

Our working hypothesis is that not all links in this projected network are medically informative to the same degree. To highlight comorbidity relations that are not explainable as arising from the prevalence of diseases, for each link in the PROJ network, we perform a statistical validation [18–27]. For each link in the original PROJ network, we assess if the link is compatible with a null hypothesis of random co-occurrence, taking into account the prevalences of the different conditions in the patient cohort for which the network is being constructed. If a link is compatible with random co-occurrence to a specified level of statistical confidence, we remove the link. By performing this statistical test for all links in the original PROJ network, we extract what we will refer to as the 'statistically validated network' (SVN) of diseases.

We therefore perform in our study a parallel analysis of PROJ networks and the SVNs for different cohorts of patients. Using community-detection methods [28], we look for communities of diseases in both types of comorbitity network.

While the original PROJ networks are so dense that the chosen community detection algorithm is unsuccessful in detecting distinct clusters of diseases, the same algorithm finds several clusters of diseases in the SVNs. For SVNs, disease communities have sizes ranging from small to medium to large number of ICD codes, and we find that they are informative about different groups of comorbidity for different cohorts of patients. We also provide examples of ICD community analysis by discussing in some detail two disease communities with an over-expressed presence of mental and behavioral disorders.

Healthcare policy decisions focused on specific cohorts of the patient population benefit from information about comorbidity relationships specific to those cohorts. In different cohorts, different diseases are characterized by a high value of node betweenness. From network science, it is known that high betweenness nodes are those nodes to be targeted and removed to drastically decrease network interconnection [29]. Therefore highlighting them provide information about those diseases that are more responsible in sustaining the interconnection of the main component of the comorbidity network. To highlight the categories of these diseases, we perform an investigation of ICD nodes contributing to the robustness of comorbidity in PROJ networks and SVNs. This analysis allow us to highlight what are the ICD code categories that have the most prominent role in keeping the comorbidity network cohesive.

Specifically, we perform a so-called 'dismantling procedure' [30, 31] for the PROJ networks and the SVNs. Network dismantling mimics preventive medicine approaches applied to different cohorts of patients aimed at the minimization of medical comorbidity. Using this procedure, we highlight some prominent roles for specific categories of disease in determining the compactness of the comorbidity network. We verify that this information is cohort-specific in several cohorts characterized by age and sex.

The remainder of the paper is organized as follows. In Sect. 2 we present the data and the methodologies used to obtain and analyze comorbidity networks together with some summary information about the PROJ networks and SVNs. Section 3.1 presents briefly the basic properties of the SVNs while in Sect. 3.2 we discuss the disease communities detected in the SVNs. Section 3.3 contains two case studies on mental and behavioral disorders, and in Sect. 3.4 we describe results obtained from a dismantling procedure to PROJ networks and SVNs of different cohorts of patients. In Sect. 4 we present our conclusions.

2 Data and methodology

2.1 Availability of data materials

The data investigated in this study are proprietary data of Auria Clinical Informatics which operates in connection with Varha. Data can be accessed with permission from Varha. The present study analyzes disease networks obtained with the approval of the Institutional Review Board of Turku University Hospital (license number T152/2017 [32]). Informed consent was waived due to the study's retrospective design, according to Finnish legislation on the secondary use of health data. The whole Auria dataset comprises longitudinal patient-level data covering all specialized healthcare contacts since 2004. The data reflect routine clinical practice and include both outpatient visits and inpatient treatment periods, the latter corresponding to continuous hospital stays from admission to discharge. Each healthcare contact included in the whole dataset was characterized by the treating specialty, care unit, and mode of admittance (e.g., emergency or elective), as well as the start and end dates of the encounter. Probable, suspected and confirmed diagnoses were recorded using the International Classification of Diseases, 10th Revision (ICD-10), including both primary and secondary diagnoses assigned during the course of care. From a clinical perspective, these data describe episodes of care as they occur in everyday practice, enabling reconstruction of individual patient pathways across outpatient and inpatient settings. For example, in cardiology, a patient presenting with acute chest pain may be admitted emergently, treated in a coronary care unit, and subsequently transferred to a general ward before discharge, with follow-up outpatient visits; such sequences can be captured longitudinally, including the timing of events and associated diagnoses. The longitudinal structure of the complete data set allows assessment of disease patterns, comorbidity burden, and outcomes over time, reflecting real-world clinical decision-making and patient management within secondary and tertiary care.

The records of the dataset available to us covers the period between the 1st of January 2004 and the 31st of July 2019, i.e., for a time period covering 16 years, and includes a total of 628,831 patients. Each of the more than 20 million line records in the dataset consists of five sets of information: (i) an anonymized patient ID, (ii) the ICD code of the diagnosis, (iii) the timestamp of the visit, and (iv) age and (v) sex of the patient at the time of the recorded event. For the purposes of our analysis. the sex of a patient in the dataset is the sex recorded in the patient's ID document.

2.2 Preprocessing and cohorting of medical data

To classify medical conditions we use the International Classification of Diseases, 10th Revision (ICD-10). This is a globally recognized classification system developed by the WHO for classifying and coding diseases, health conditions, and causes of death. It provides standardized codes that facilitate the documentation, reporting, and analysis of health data across different regions and healthcare settings. In the 10th revision, ICD codes consist of a letter (disease class) followed by two or more digits (which can be separated by decimal points), progressively deepening the identification of the disease (or condition for a subset of categories). For example, codes starting with the letter F refer to mental and behavioral disorders. Within this group, F32 (i.e., a so-called level-3 code composed by three characters) represents the occurrence of depressive episodes, and further, F32.2 (i.e., a level-4 code) stands for severe depressive episode without psychotic symptoms. The full list of Level-4 codes can be found on the WHO website [6]. In Table I of Supplementary Information we list the disease categories classified by the first letter of ICD codes. In our study,

Table 1 Number of patients, nodes and links in the bipartite, projected (PROJ), and validated (SVN) networks for different cohorts of patients

Number of	0-9 F	0-9 M	10-19 F	10-19 M	20-29 F	20-29 M	30-39 F	30-39 M	40-49 F
Patients	46,387	53,130	44,451	45,549	71,057	55,076	64,337	50,887	57,487
Bipartite links	192,414	247,572	236,009	237,602	470,116	236,183	583,306	240,946	469,684
PROJ nodes	3904	4121	5250	4933	5904	5163	6113	5296	6184
PROJ links	215,096	266,002	407,770	381,156	711,053	398,215	911,158	467,127	964,727
SVN nodes	1531	1810	2401	2276	3374	2437	3503	2549	3632
SVN links	9192	11,629	14,861	13,652	31,306	13,178	37,204	16,624	37,187

Number of	0-49 M	50-59 F	50-59 M	60-69 F	60-69 M	70-79 F	70-79 M	80+ F	80+ M
Patients	52,821	66,699	62,776	65,587	63,156	55,947	48,046	41,505	24,463
Bipartite links	292,536	506,556	418,254	568,945	541,195	578,651	530,509	539,798	345,711
PROJ nodes	5561	6349	5910	6130	6047	5860	5713	5429	4878
PROJ links	613,123	1,014,863	832,433	1,083,676	1,020,775	1,051,192	973,948	859,456	682,662
SVN nodes	2897	3805	3438	3709	3632	3405	3103	2574	2122
SVN links	22,068	38,539	32,072	39,181	37,627	34,911	29,397	20,500	12,379

we set the granularity of our study by using level-4 ICD codes. This defines a set of 9303 different codes.

In a first step, we identify all entries for a specific patient and thus construct a list of all conditions this patient is diagnosed with (at any time). For further analysis, we divide patients into cohorts characterized by sex and age. Specifically, we study 10-year age groups, resulting in a total of 18 different cohorts (9 age groups, two sexes). The reference age is always the age at the time of diagnosis. For a given sex and age group, we include all conditions patients were diagnosed with within the age limits of the cohort or earlier. This is because many diseases can have long-term effects, causing comorbidities. For example, if a patient was diagnosed for the first time with ICD code S61.9 (open wound of wrist and hand, part unspecified) when he or she was 17 years old and has a diagnosis with ICD code F50.1 (atypical anorexia nervosa) at age 24, then ICD code S61.9 contributes to the 10-19 cohort, and both diagnoses (S61.9 and F50.1) contribute to the 20-29 age cohort.

2.3 Bipartite and projected networks

Bipartite networks for the different cohorts are built from the dataset by connecting each patient ID with ICD codes reported in her/his medical history if this patient was diagnosed with this condition at least once at an age in the cohort age limits, or earlier. Overall, there is a slight sex imbalance in the number of patients in the dataset, 48.1% are male, and 51.9% female. Some summary statistics for the PROJ networks and SVNs are shown in Table 1 for the different cohorts. The number of links in the bipartite network is larger in the age cohorts from 50 to 79 years compared to other age groups for both sexes. In the age range 20-39, the number of links in the bipartite network is larger for females than for males due to pregnancy-related diagnoses.

As shown in Fig. 1, for all cohorts, the degree of ICD nodes in the bipartite network is more heterogeneous (ranging from one up to more than 10^4) than the degree of patient nodes (where the degree is ranging from one up to about 80).

2.4 Statistically validated networks

We now describe the statistical validation procedure by which we obtain the SVNs from the PROJ networks.

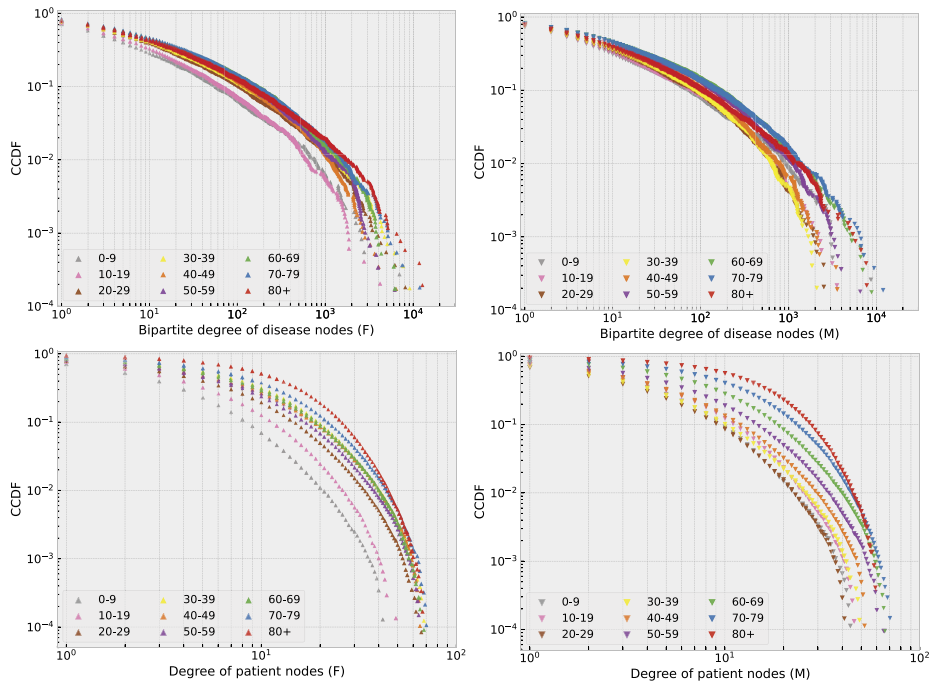

Figure 1 Complementary cumulative distribution function of the degree of ICD nodes (top panels) and of patient nodes (bottom panels) in the bipartite network for all patient cohorts of females (left panels) and males (right panels)

Suppose the initial bipartite network for a cohort has N_D disease nodes (ICD codes) and N_P patient nodes, with a link connecting a patient and a disease if this patient was diagnosed with the condition at an age in the cohort age period or earlier. The PROJ network then consists of N_D nodes, and a link exists between any two nodes if at least one patient in the cohort had both diseases.

We focus now on two nodes in the PROJ network, A and B. These are two ICD codes. Assume now that N_A is the number of patients in the cohort diagnosed with condition A, and N_B the number of patients with condition B. Further assume that N_{AB} patients in the cohort have been diagnosed with both conditions (at an age within the age limits of the cohort, or earlier).

We now formulate the null hypothesis of random co-occurrence of condition A, and B. This can be thought of as follows: there are N_A out of N_P patients with condition A and N_B patients with condition B. We can then ask what the probability is to have a given number ℓ of patients with both conditions, if the N_A and N_B individuals were selected at random from the total of N_P patients. Under the assumption that the heterogeneity of number of diseases for patient is moderate the probability of the null hypothesis of observing both conditions is given with a good approximation by [18, 19]:

$$H_{AB}(\ell|N_P,N_A,N_B) = \frac{\binom{N_A}{\ell}\binom{N_P-N_A}{N_B-\ell}}{\binom{N_P}{N_B}}. \tag{1}$$

Equation (1) approximates to the exact probability to have a given number ℓ of patients with both conditions in the absence of heterogeneity of the degree of the patient's nodes

[18]. Under above hypotheses the p-value of observing N_{AB} or more co-occurrences is:

$$p\text{-value}(N_{AB}) = \sum_{\ell=N_{AB}}^{\min(N_A,N_B)} H_{AB}(\ell|N_P,N_A,N_B). \tag{2}$$

This allows us to test whether or not the observation of N_{AB} patients in the data with both conditions is compatible with the null hypothesis of random co-occurrence at a given level of statistical confidence. The validation procedure implies the execution of a large number of tests (one for each link in the original PROJ network), requiring a multiple-hypothesis test correction [33]. Specifically, we used the so-called 'control of the false discovery rate' (FDR) [34]. To perform the hypothesis test, we set the uni-variate statistical threshold at 0.01 and then apply the FDR correction. If a link in the PROJ network leads to a p-value lower than the corrected threshold, we conclude that the null hypothesis is rejected, i.e., the empirical observation of N_{AB} patients with both conditions A and B is not statistically compatible with random co-occurrence hypothesis. The link is then included in the SVN. Conversely, when the null hypothesis is not rejected, we do not include the comorbidity link in the SVN. Finally, isolated nodes are not included in the SVN.

For example, E03.9 (Hypothyroidism, unspecified) and J01.0 (Acute maxillary sinusitis) are common diseases in the 50-59 cohort of males, so it's quite likely that several patients got both diseases just due to their high prevalence. In fact, in the 50-59 cohort, these two diseases are connected in the PROJ network, but they are not connected in the corresponding SVN suggesting that the medical relevance of this observed comorbidity in some patients might be limited or even absent. The validation was performed by using the Python module accessible at the `svalnet` github repository [35].

2.5 Community detection in comorbidity networks

Having obtained the PROJ networks and the SNVs for different cohorts of patients, we search communities of diseases in both versions of comorbidity network using community-detection methods for complex networks [28]. This provides communities of diseases characterized by pronounced intra-community connectivity.

Community detection in complex networks is used to identify groups of nodes, customary called communities, that are more densely connected internally than with the rest of the network. This task is central to understanding the mesoscopic structure of systems such as biological, medical, social, and technological networks. A wide range of methods has been developed, reflecting different conceptual and mathematical frameworks. Modularity optimization [36] seeks partitions that maximize the difference between observed and expected link densities, with algorithms such as the Louvain and Leiden methods offering efficient implementations [37]. Statistical inference approaches, notably the stochastic block model [38], treat community detection as a probabilistic inference problem, estimating the likelihood that observed edges arise from underlying community structure. Another widely used approach is flow-based. Specifically this is the approach followed by the Infomap [39] algorithm. This algorithm conceptualize communities as regions that efficiently retain information flow, identifying modules by minimizing the description length of random walks on the network.

In our study we use the Infomap algorithm because its search structure in terms of random walk performed on the comorbidity network is mimicking the disease progression

associated with multimorbidity. The Infomap algorithm seeks to minimize the observed length of movements within and between modules. Communities are thus defined as partitions of the network that achieve the most efficient compression of an informational representation. In essence, Infomap identifies network modules by optimizing how concisely the trajectories of information flow can be encoded, providing a dynamic and statistically grounded view of network structure.

While the original PROJ networks are so dense that Infomap algorithm is unsuccessful in detecting distinct clusters of diseases, the same algorithm finds several clusters of diseases in the SVNs. By using the terminology of complex networks we call these clusters "communities". The communities are a large number and have sizes ranging from small to medium to large number of ICD codes for different cohorts of patients.

We label different communities by codes such as c11_60.69_M and c5_70.79_M. For example, the former is the community with numeric label 11 of the cohort of ages 60-69 for male patients, and the latter is the community with numeric label 5 of the cohort of ages 70-79 again for males.

2.6 Similarity between pairs of ICD communities

We verify that different cohorts of patients are characterized by distinct sets of communities in the SVNs.

To quantify the similarity between pairs of SVN communities obtained for the different cohorts of patients, we compute the Jaccard similarity $J(C_{k,a}, C_{\ell,b})$ between community a of cohort k $C_{k,a}$ and community b of cohort ℓ $C_{\ell,b}$ as follows [40, 41]:

$$J(C_{k,a}, C_{\ell,b}) = \frac{|\text{Edges}(C_{k,a}) \cap \text{Edges}(C_{\ell,b})|}{|\text{Edges}(C_{k,a}) \cup \text{Edges}(C_{\ell,b})|} \tag{3}$$

where $\text{Edges}(C_{k,a})$ defines the set of links connecting the nodes of community $C_{k,a}$. The intersection between sets (labeled by the symbol '$\cap$') contains links present in both sets, whereas the union ('$\cup$') contains all links present in at least one of the two sets. The value of Jaccard similarity ranges from zero (when no link is present in both communities) to one (when both communities contain the exact same set of links).

We provide an overall comparison of comorbidity communities found in different patient cohorts by computing the dissimilarity $d(C_{k,a}, C_{l,b}) = 1 - J(C_{k,a}, C_{l,b})$ between each pair of communities and grouping them by using agglomerative hierarchical clustering [42]. Using this distance we perform an agglomerative hierarchical clustering procedure. The clustering procedure starts with n isolated elements, each containing a single element, computes the pairwise distance matrix between all observations and then proceeds with an iterative merging to find the closest pair of elements according to a linkage criterion (e.g., Single, Complete, Ward's or Average linkage). The sequence of merges is summarized in a hierarchical tree (also called dendrogram). In this study we use the average linkage algorithm [42].

2.7 Over-expression of disease category in a set of diseases

The inspection of the hierarchical tree is supported by a statistical test detecting the category or categories of diseases with an over-expressed number of ICD codes in each community (see Sect. 3.2.4).

To estimate over-expressed ICD categories in the ICD comorbidity communities obtained with Infomap we use the method introduced in Ref. [43]. It is a statistical method identifying which categorical features are significantly over-expressed within a given subset (or community) of elements. The method assumes a null hypothesis that heterogeneous attributes are randomly distributed across all elements of the system. For each attribute we test whether the number $N_{C,Q}$ of ICD elements in community C that have the attribute Q is over-expressed with respect to what expected by the null hypohesis of random matching. The probability that X elements in cluster C have the attribute Q, under the null hypothesis that elements in the cluster are randomly selected, is approximately given by the hypergeometric distribution [43]

$$H(X \mid N, N_C, N_Q) = \frac{\binom{N_C}{X}\binom{N - N_C}{N_Q - X}}{\binom{N}{N_Q}}, \tag{4}$$

where N_Q is the total number of ICD elements in the system with attribute Q, N is the total number of ICD elements, and N_C is the number of ICD elements in community C. By using this distribution one can associate a p-value with the observed number $N_{C,Q}$ of elements in cluster C that are classified with the attribute Q according to the equation

$$p\text{-value}(N_{C,Q}) = 1 - \sum_{X=0}^{N_{C,Q}-1} H(X \mid N, N_C, N_Q). \tag{5}$$

If p-value$(N_{C,Q})$ is smaller than a given statistical threshold p_s, we reject the null hypothesis and conclude that the attribute Q is over-expressed in community C. In our test we use $\alpha = 0.01$ as a statistical threshold and we perform the control of the FDR [34].

With this characterization, we note that communities in which the same disease categories are over-expressed tend to group in the same subregions of the hierarchical tree, even when they belong to cohorts of patients of different age and sex.

2.8 Dismantling of a network

Here we consider the procedure of fractioning a connected component of a complex network by performing a so-called "dismantling" procedure [31]. For each network of interest, we remove one node at a time from the largest connected component of it, iterating the removal process until percolation across all remaining nodes of the network breaks down. We monitor the presence of percolation among the remaining nodes of the largest connected component by comparing the size of the first and second largest connected components at each removal step. In the presence of percolation, the largest connected component is much bigger than the second largest connected component. In contrast, when percolation is lost, the first and the second largest connected components have comparable sizes [44].

The sequence of node removal is obtained by using the following procedure: at each step, we compute the betweenness of all nodes, and we remove the node with the highest betweenness [31]. Node betweenness centrality is a fundamental measure in complex network analysis that quantifies the extent to which a node functions as an intermediary within the overall structure of a network. Formally defined as the fraction of shortest

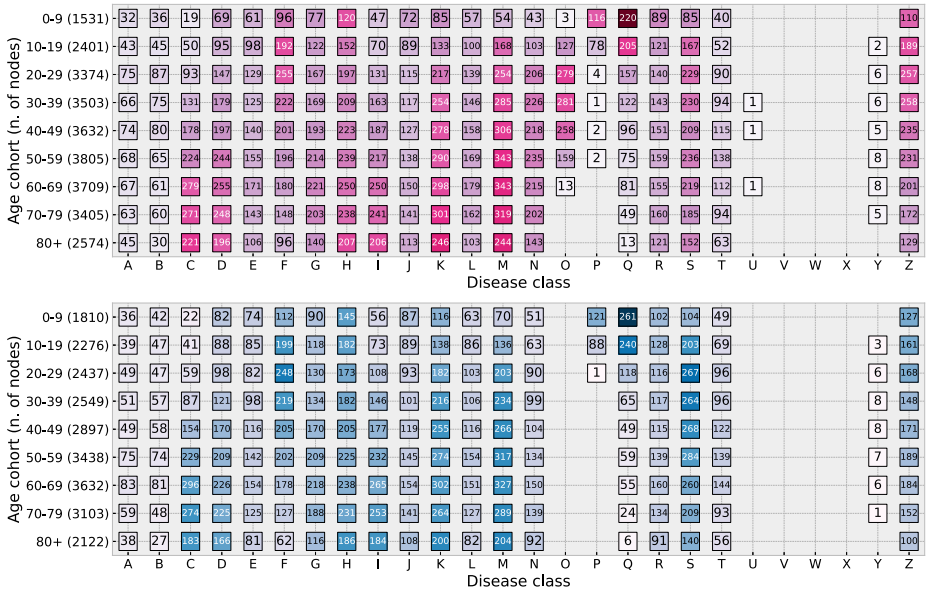

Figure 2 Statistics of ICD category composition of SVNs for all cohorts of patients. ICD categories (columns) for different age cohorts (rows) and sexes (females in the upper panel and males in the lower). The number in brackets next to the age limits of the cohort is the total number of nodes in the SVN. The number inside the squares is the number of nodes of the disease class (the letter in the ICD code) in the given network, while the color is proportional to the fraction of diseases that belong to that class

paths between all pairs of nodes that pass through a given node, betweenness centrality captures the node's potential to control or influence the flow of interactions, resources, or information in the system. High-betweenness nodes often act as critical bridges or brokers connecting otherwise distant or weakly linked regions of the network, and their removal can significantly disrupt network connectivity or efficiency. Because it reflects structural position rather than merely local connectivity, betweenness centrality is widely used to identify key regulators in biological or medical networks, influential actors in social systems, and vulnerable points in technological or infrastructural infrastructures [1–3].

3 Results

3.1 Properties of the validated networks

Table 1 shows that the number of validated links is less than 5% of that in the PROJ network. This large reduction of comorbidity links does not significantly affect the number of ICD nodes present in the SVNs. In fact, the percentage of retained nodes ranges from about 40% to 60% therefore providing information on a large number of ICD level-4 nodes. For all cohorts, both for PROJ networks and SVNs consist of a largest connected component (LCC) containing more than 90% of the network nodes. For a summary of the statistics of the networks see Sec. S2 of the Supplementary Information. The remaining nodes are grouped into several small components each with only a small number of nodes.

For a summary of the statistics of ICD codes present in the SVNs for all cohorts, see Fig. 2. In general, the two sexes do not present strong differences in disease distributions, except for specific categories. Prominent examples of differences are trivially the categories O (pregnancy, childbirth and the puerperium), and further N (diseases of the genitourinary system). A higher number of nodes of this category are seen for each age-

Table 2 Number of communities of any size, size larger than 10 nodes or 25 nodes of SVNs for each cohort. We also report the number of nodes of the SVN and the number of nodes of the largest community detected. The total number of detected communities of size larger than 25 ICD codes is 380

Cohort	Nodes of SVN	# of comm. (any size)	# of comm. (nodes > 10)	# of comm. (nodes > 25)	Size of the largest SVN community
0-9 F	1531	145	24	10	301
0-9 M	1810	152	27	10	384
10-19 F	2401	178	42	20	233
10-19 M	2276	202	41	15	331
20-29 F	3374	282	44	23	666
20-29 M	2437	176	46	20	294
30-39 F	3503	284	50	22	1182
30-39 M	2549	160	56	22	434
40-49 F	3632	253	52	28	373
40-49 M	2897	157	50	26	285
50-59 F	3805	267	46	26	748
50-59 M	3438	227	50	24	482
60-69 F	3709	233	49	27	621
60-69 M	3632	212	48	26	710
70-79 F	3405	237	49	25	984
70-79 M	3103	178	52	20	949
80+ F	2574	166	48	20	798
80+ M	2122	187	39	16	509

specific female network (Fig. 2) compared to the male network for the same age bracket. Categories H, K, and M are present in similar quantities through all ages, while other categories turn out to be more expressed at specific age intervals, e.g. categories P and Q in early ages, F and S in teenagers and young adults, and C, D, and I at later ages.

3.2 Disease categories and community structure

3.2.1 Community detection in comorbidity networks

We now analyze the structure of comorbidity networks by looking at highly interconnected subgroups of ICD codes. To this end, we perform community detection. This is an unsupervised data mining procedure to identify groups in a complex network whose nodes are more connected between each other rather than to other nodes in the network [28]. As discussed in Sect. 2.5, we chose the Infomap algorithm [39] to perform community detection both in the PROJ networks and SVNs. In the PROJ networks, only one very large community is found by the algorithm for each age and sex cohort. In other words, the PROJ network of diseases has a high link density in all regions of the network, and the algorithm is not able to detect distinct regions of the network. On the other hand, in the SVNs many communities are observed for each cohort of patients.

3.2.2 Number of disease communities in different cohorts

In Table 2, we report the number of ICD communities detected by the Infomap algorithm for all cohorts. Communities vary in size ranging from groups of two ICD codes to communities with hundreds of nodes. The difference in the number of communities between male and female SVNs does not reflect the difference in the number of nodes. This hints at the fact that communities are more influenced by disease categories and interconnections than the size of the network. Although medical information can be obtained by the inspection of communities of any size, in the present study, we focus our attention on communities larger than 25 nodes.

3.2.3 Similarity of communities across cohorts, and hierarchical tree

One might expect that communities of ICDs carrying relevant biomedical information would be observed across cohorts. We test this hypothesis by computing the similarity between pairs of SVN communities in different cohorts.

Across all cohorts in Table 2 we find a total of 380 communities with more than 25 nodes each. Across all pairs of communities among this set, the Jaccard similarity runs from 0 to the maximum value of 0.609. Computing the 380×380 matrix of pairwise Jaccard similarities between communities reveals that a large fraction of community pairs have a similarity value close to 0.6, indicating a large overlap of edges present in several communities.

Starting from the similarity matrix and applying the hierarchical clustering method of the average linkage, we obtain the hierarchical tree shown in Fig. 3. For future purposes each of the 380 communities is assigned a running number (1 to 380) in the order as they appear in the tree. The vectorial PDF of Fig. 3 can be downloaded from our repository [35] for a better in-depth visualization.

We find that the hierarchical tree is quite structured, indicating that the SVNs have communities of different patient cohorts that cluster together. In other words, there are similarities among communities of diseases that persist over cohorts of different age intervals and different sexes. The coherence of groups of communities detected by hierarchical clustering in terms of their membership to specific cohorts and categories of diseases can be seen in detail by analyzing the Jaccard similarity matrix shown in Sec. S3 of the Supplementary Information. The rows and columns of the matrix are arranged in the order as they appear in the average linkage hierarchical tree in Fig. 3 to efficiently visualize the clustering of groups of communities. The Jaccard similarity matrix primarily shows a block diagonal structure but also a moderate overlap of different blocks is sometime observed (see Fig. S1 of the Supplementary Information). We will comment more on such overlap in the next sub-section.

3.2.4 Over-expression of ICD codes in different communities

We find that different communities are characterized by an abundance of ICD codes of one or a few categories. To confirm this observation quantitatively, we perform a statistical test evaluating the over-expression of ICD codes of a specific category for each community. The statistical test is discussed in Sect. 2.7. Based on the statistical test, we conclude that 369 communities out of 380 present over-expression of one or more ICD categories. In Fig. 3, when we observe an over-expression of ICD codes with just a single specific letter in a specific community, we draw the given community with a color associated with the over-expressed category.

Large branches of the same color are noticed in Fig. 3, indicating that these sets of communities are rich in a specific category of diseases. An overall analysis of all over-expression shows that 293 communities present over-expression of just one category, 70 communities of two categories, 13 communities of three categories, and one community has over-expression of 6 categories. The high number of communities with over-expression of one (or more than one) ICD category suggests that the partition in communities of SVNs highlights information that can be interpreted in terms of specific groups of diseases.

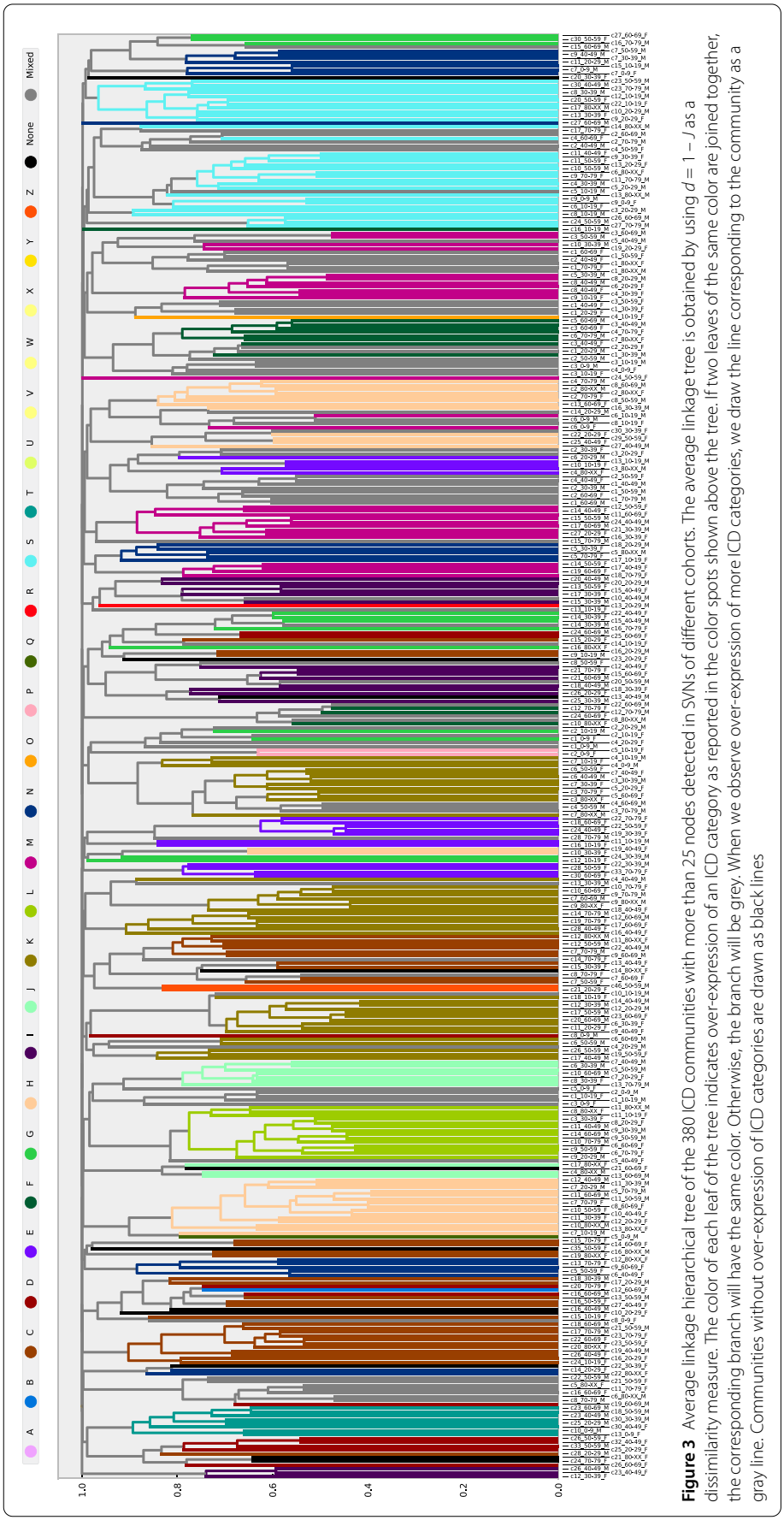

Figure 3 Average linkage hierarchical tree of the 380 ICD communities with more than 25 nodes detected in SVNs of different cohorts. The average linkage tree is obtained by using $d = 1 - J$ as a dissimilarity measure. The color of each leaf of the tree indicates over-expression of an ICD category as reported in the color spots shown above the tree. If two leaves of the same color are joined together, the corresponding branch will have the same color. Otherwise, the branch will be grey. When we observe over-expression of more ICD categories, we draw the line corresponding to the community as a gray line. Communities without over-expression of ICD categories are drawn as black lines

The presence of the statistically validated over-expression suggests a rich pattern of comorbidity involving diseases of different types and/or affecting different organs or physiological districts. The complete list of over-expressions of ICD communities is provided in Sec. S4 of Supplementary Information, where we detect long sequences of communities (when ordered as they appear in the hierachical tree) with over-expression of the same ICD category. The list of over-expressed categories also highlights the presence of a few large clusters characterized by over-expression of more than one ICD category.

It is worth analyzing in parallel the sequence of over-expressed categories of groups of communities and the block like structure of the Jaccard similarity matrix. A detailed analysis shows that there is some overlap between blocks whenever two blocks have at least one common over-expressed category. The presence of this (usually) small overlap suggests that some sub-regions of comorbidity communities are nested into larger communities and such nested structure are similar in otherwise distinct comorbidity communities.

In summary, ICD communities detected in SVN comorbidity networks partition the comorbidity space and provide robust information on the degree of similarity or differences between each pair of ICD communities specific for age interval and sex.

3.3 Analysis of community clusters: the case of mental and behavioral disorders

In this subsection, we provide two examples of analyses of clusters of communities detected in the SVNs and characterized by the over-expression of a specific ICD category. We discuss clusters of communities characterized by the over-expression of category F (mental and behavioral disorders). Two regions of the average linkage hierarchical tree present this over-expression (see complete list of category over-expression in Sec. S4 of Supplementary Information). They are the cluster comprising communities located in the average linkage hierarchical tree from position 199 to 204 (see Fig. 4) and the cluster of communities located from position 291 to 305 (see Fig. 5).

The first group (Fig. 4) consists of 6 communities of ICD codes from cohorts with age intervals from 60-69 to 80+ for females and males. SVNs of these communities are shown in Sec. S5 of Supplementary Information. In addition to the F over-expression in three communities, we also observe over-expression of the G ICD category (diseases of the nervous system). The majority of the diseases present in this cluster concerns degenerative diseases of the nervous system coded in the set of ICD codes ranging from G30.* to G39.*.

In Fig. 6, we show the normalized occurrence (top panel) and the over-expression (bottom panel) of ICD codes present in each disease's community of the cluster. We note that the occurrence of ICD codes is almost exclusive to the categories F, G, R, and Z and a few additional diseases of I, D, and T categories. This cluster is therefore specific for degenerative disease of the nervous system affecting both females and males starting from the age of sixties.

This branch of the hierarchical tree (Fig. 4) has leaves merging at relatively small distances. In fact, the merging of the whole branch occurs at $d = 0.6327$, i.e., at a value lower than the ones observed for the large majority of homogeneous clusters of communities observed in Fig. 3. Communities are characterized by a number of nodes ranging from 26 (c24_60-69_F) to 40 (c8_80-XX_M) and a number of edges ranging from 78 (c24_60-69_F) to 132 (c12_70-79_F). By performing a graph intersection of the six communities (i.e., by selecting those edges that are present in all communities) we select the core of ICD codes and comorbidity relations that are common to all six communities. This subnetwork has

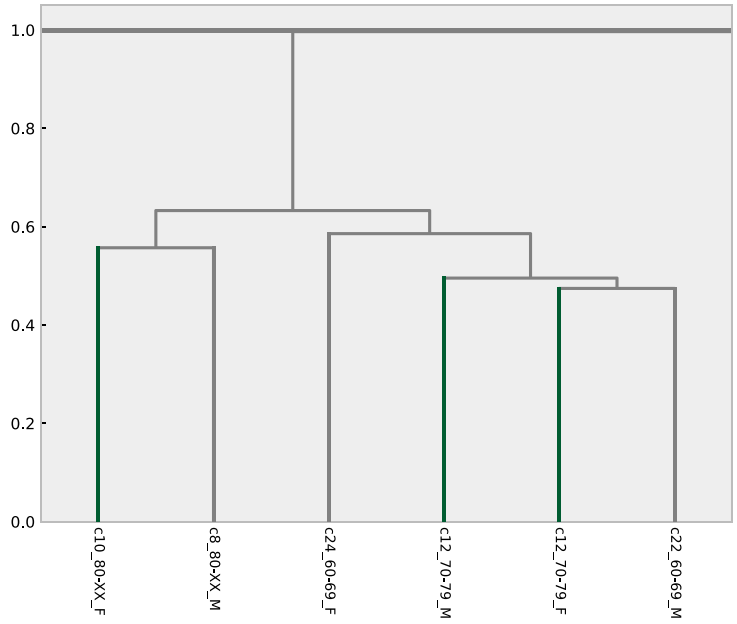

Figure 4 Selected region of the average linkage hierarchical tree with communities located at positions from 199 to 204. The over-expression of disease category for each community starting from left is as follows F, FG, FG, F, F, FG

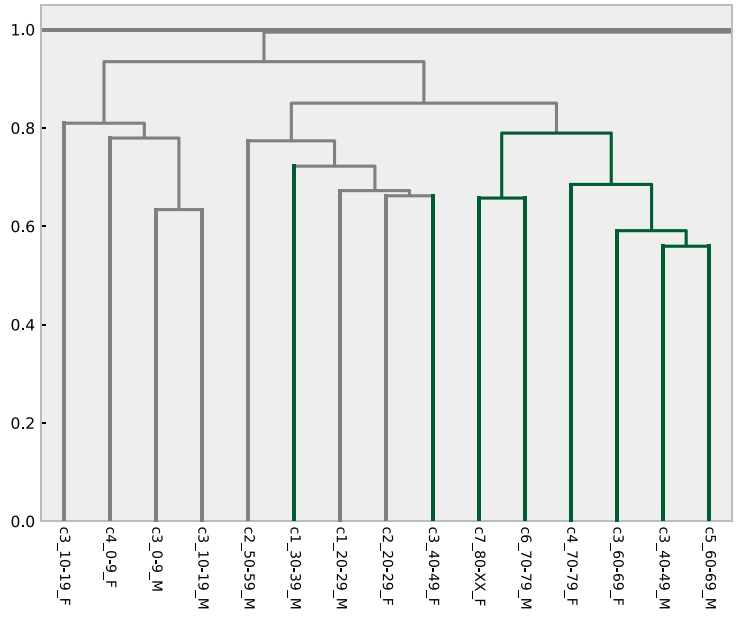

Figure 5 Selected region of the average linkage hierarchical tree with communities located at positions from 291 to 305. The over-expression of categories for each community, starting from left, is as follows FYZ, FZ, FZ, FZ, FS, F, FZ, FZ, F, F, F, F, F, F

11 nodes and 36 edges. The 11 ICD codes of the intersection are the following: F00.0, F00.1, F00.2, F00.9, F03, F06.7, G30.0, G30.1, G30.8, G30.9. The 36 edges of the graph intersection are given in Sec. S6 of Supplementary Information.

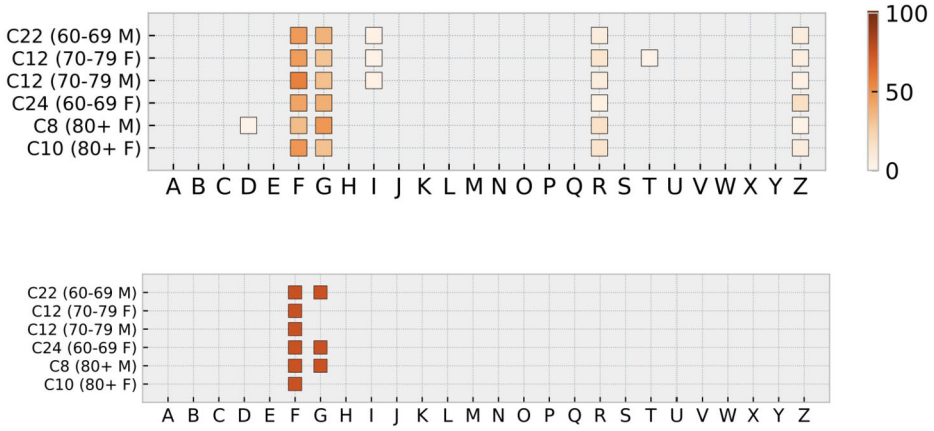

Figure 6 Normalized occurrence (upper panel) and over-expression (lower panel) of the different ICD letter categories in the communities in the cluster of Fig. 4. The color scale in the upper graph represents the percentage of nodes belonging to a particular disease category (letter) over all the category of each community. In the lower graph, the presence of a red square indicates over-expression of the category

The ICD-10 codes associated with the 11 nodes at the intersection are primarily associated with Alzheimer's disease (G30-series) and its related dementias (F00-series), with additional codes referring to unspecified dementia (F03), mild cognitive disorder due to medical causes (F06.7), and mental health screening (Z13.3). Collectively, they describe conditions related to Alzheimer disease and their consequences, ranging from early symptom screening to advanced dementia. A few specific ICD codes are highlighted when one considers graph intersections involving smaller groups of communities. For example, the intersection of communities c10_80-XX_F and c8_80-XX_M (let us label this set as set 1 for this discussion) involves the set of nodes F00, F00.0, F00.1, F00.2, F00.9, F01.0, F01.1, F01.2, F03, F05.1, F06.7, G30.0, G30.1, G30.8, G30.9, G31.1, R41.0, and R41.8 whereas, at the other side of the branch, the intersection of communities c12_70-79_F and c22_60-69_M (labeled as set 2) involves the ICD codes F00.0, F00.1, F00.2, F00.9, F01.2, F01.9, F02.0, F03, F05.1, F06.7, F07.0, G30.0, G30.1, G30.8, G30.9, G31.0, G31.1, G31.8, G31.9, R41.3, and R41.8. As expected by the overall similarity there is a large overlap but specificity for each pair of communities are also observed. Set 1 is more Alzheimer-focused, it shows the presence of vascular dementia (F01.1 is specific to this set) and includes classic cognitive symptoms like disorientation (R41.0 again this node is specific to this set). Overall, Set 1 reflects a more traditional geriatric cognitive profile.

Set 2 has similar Alzheimer coverage but also presents broader neurodegenerative disease categories (G31.8, G31.9), some specific diagnoses like frontotemporal dementia (G31.0) and Pick's disease dementia (F02.0), personality/behavioral changes (F07.0), and memory impairment focus (R41.3). In summary, this branch of the hierarchical clustering suggests the coverage of dementia but a distinct dementia/neurodegenerative profiles in 60-80 yrs versus high-comorbidity dementia in age of more than 80 yrs of set 1.

This observation is supported by results presented in the medical literature. In fact, the epidemiology of frontotemporal dementia (G31.0) [45] suggests that although traditionally considered "younger-onset", a sizeable proportion are diagnosed after age 60-65 (for example, highest prevalence in 60-69 age group). The same conclusion is reached in Ref. [46]. Concerning dementia in the oldest cohorts (age > 80) we highlight the joint pres-

ence of Alzheimer (G30.x) and vascular pathology (F01.1 and F01.9). For this cohort, the observation of this comorbidity is consistent with Ref. [47] where the authors show that comorbidity of Alzheimer and vascular pathology is common for patients with age ≥ 85. Disorientation in time and place (R41.0) in old age is also documented and it is discussed in detail in Ref. [48].

The second group of ICD communities (see Fig. 5) distinctly shows three sub-branches. The first, that is located right of the branch, and comprises the six communities c5_60-69_M, c3_40-49_M, c3_60-69_F, c4_70-79_F, c6_70-79_M, and c7_80-XX_F. This "right" sub-branch merges with the "middle" sub-branch (c3_40-49_F, c2_20-29_F, C1_20-29_M, c1_30-39_M, and c2_50-59_M). The agglomeration of these two sub-branches then merges with the "left" branch (c3_10-19_M, c3_0-9_M, c4_0-9_F, and c3_10-19_F) at a distance $d = 0.9347$. This distance is much higher than the one observed for the cluster of communities of Fig. 4. For the sake of simplicity, we address the three sub-branches as the right, middle and left sub-branch respectively.

In each cohort, numerical labels of communities are set starting from bigger communities. In other words, the number of nodes of a community is high when the numeric label of a community is low. The fact that in the communities of Fig. 5 the numeric labels of communities are pretty low values is therefore associated with the fact that these communities have a large number of nodes and edges. The number of nodes of communities ranges from 36 (c7_80-XX_F) to 433 (c2_50-59_M) and the number of edges ranges from 201 (c7_80-XX_F) to 5676 (c1_30-39_M). In other words, the 15 communities of Fig. 5 are not very specialized with respect to the category of diseases although the F category is the most prominent one in all communities (see top and bottom panels of Fig. 7).

To highlight the disease profile of the three sub-branches we compute the graph intersection among the communities of different sets of communities. The graph intersection among all 15 communities of the branch consists of only 4 nodes (F32.0, F32.1, F41.2, and Z00.4) and 3 edges (Z00.4 linked to the three ICD codes of category F). Therefore, at the most general level co-occurrence was identified among the ICD-10 codes F32.0 (Mild depressive episode), F32.1 (Moderate depressive episode), and F41.2 (Mixed anxiety and depressive disorder). This finding aligns with existing epidemiological evidence indicating substantial symptomatic and diagnostic overlap between depressive and anxiety disorders, as well as their frequent sequential or concurrent presentation over time [49, 50]. The observation of Z00.4 (General psychiatric examination, not elsewhere classified), although not indicative of a pathological diagnosis, provides information on patient interactions with the healthcare system.

It is worth pointing out that our approach includes in our analysis ICD codes of both the type R (Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified) and Z (Factors influencing health status and contact with health services). This inclusion enables the integration of healthcare utilization patterns into the comorbidity network, thereby providing a more comprehensive understanding of both clinical and behavioral dimensions of mental health. In this way, the analysis captures not only the burden of affective and anxiety-related conditions but also the dynamics of healthcare engagement and continuity of psychiatric care within the studied population.

The branch divides in two sub-branches, the left branch and a sub-branch merging the middle and right sub-branch. We first analyze the characteristics that are distinctive of the left sub-branch. All communities of this sub-branch refer to cohorts of patients from 0 to

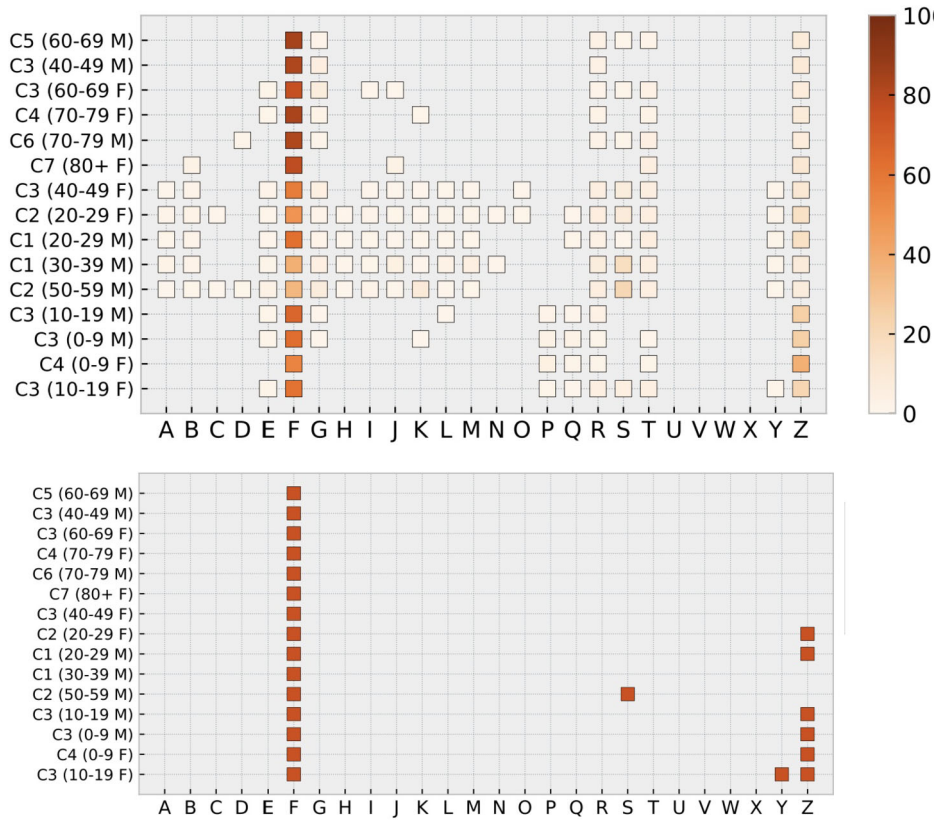

Figure 7 Relative occurrence and over-expression of ICD letter category in the communities in the cluster of Fig. 5. The color scale in the upper graph represents the percentage of nodes belonging to a particular disease category (letter) over all categories of each community. In the lower graph, the presence of a red square indicates over-expression of the category

19 years old. When we consider the four communities c3_10-19_M, c3_0-9_M, c4_0-9_F, and c3_10-19_F we observe that the graph intersection presents a clique of 9 ICD codes (F90.0, F91.3, F92.9, F93.8, F93.9, Z00.4, Z61.8, Z63.2, and Z63.8) as its largest clique.

This set of ICD codes describes childhood behavioral and emotional dysregulation associated with family or psychosocial adversity, ending up contacting with mental health services. In network terms, it is a psychosocial-behavioral clique, tightly connected via both diagnostic (F90.0, F91.3, F92.9, F93.8, and F93.9) and contextual (Z00.4, Z61.8, Z63.2, and Z63.8) nodes, representing a clinically coherent subgroup in the pediatric population. The F90–F93 codes all belong to the child and adolescent mental and behavioral disorders block. They cover Attention Deficit Hyperactivity Disorder ADHD (F90.0), oppositional defiant/conduct problems (F91.3, F92.2), and emotional disorders (F93.8 and F93.9). These commonly co-occur: ADHD and ODD/conduct disorder are highly comorbid (40–60%), and both often overlap with anxiety or emotional dysregulation in youth [51].

The Z-codes reflect encounters and psychosocial circumstances. Specifically, Z00.4 is associated with psychiatric evaluation, suggesting recognition of the psychiatric problem by the healthcare system. Codes Z61.8, Z63.2, Z63.8 point to family adversity, inadequate support, or traumatic experience. Overall, this clique represents a psychosocial–behavioral phenotype of children/adolescents with neurodevelopmental and emotional

dysregulation disorders often under social stress or family dysfunction with engagement in psychiatric assessment or care. In Finland (as in other Nordic countries), these patterns are well documented [51, 52]. Children with ADHD or disruptive disorders have high odds of family stressors, social service involvement, and psychiatric follow-up codes. In summary, the core of the sub-branch likely captures a behavioral–psychosocial care cluster, not merely a medical/biological comorbidity.

The middle sub-branch has 5 communities, and these communities are pretty large in terms of nodes and edges. The graph intersection of these communities presents 9 largest cliques of 21 ICD nodes each. These cliques comprises the following 25 ICD nodes (F10.1, F10.2, F29, F31.3, F31.6, F31.9, F32.0, F32.1, F32.2, F32.3, F32.9, F33.0, F33.1, F33.2, F33.9, F40.1, F41.0, F41.1, F41.2, F41.9, F60.3, F61.0, T36, Z00.4, Z76.0). This sub-branch presents communities observed for adults' cohorts aged 20 to 60 years. A notable degree of comorbidity was observed across diagnostic categories encompassing disorders related to alcohol abuse (F10.1, F10.2), psychotic and affective disorders (F29, F31.x, F32.x, F33.x), anxiety and personality disorders (F40.1, F41.x, F60.3, F61.0), as well as non-psychiatric and contextual codes (T36, Z00.4, Z76.0).

Again, we note that the inclusion of Z codes (notably Z00.4 and Z76.0, with this last code indicating issue of repeat prescription) extends the analysis beyond disease-specific morbidity to encompass patient interactions with the healthcare system. This approach acknowledges that frequent or prolonged contact with healthcare services constitutes an important dimension of comorbidity, potentially reflecting chronicity. By integrating both diagnostic and interaction codes, our analysis captures basic aspects of the multifaceted burden associated with mental and behavioral health conditions in the Finnish working-age population consistent with previous Finnish studies showing strong anxiety-mood co-occurrence [53] and increasing linkage of anxiety and alcohol abuse in Finnish-register populations [54].

The right sub-branch involves patients of adults aged more than 40 years. They are characterized by a graph intersection where 11 ICD nodes are present in three largest cliques of 10 ICD nodes each. These nodes are F32.0, F32.1, F32.2, F32.3, F33.0, F33.1, F33.2, F33.3, F33.9, F34.1, F41.2, Z00.4. All of them are also present in the core set of the middle sub-branch. In this sub-branch, the diagnostic pattern is characterized by recurrent depressive disorders (F33.x), single-episode depressive disorders (F32.x), dysthymia (F34.1), and mixed anxiety–depressive disorder (F41.2), in combination with the contextual code Z00.4. This set of nodes suggests a clinical profile dominated by chronic and recurrent mood disturbances in later adulthood. The predominance of depressive-spectrum diagnoses, including both single and recurrent episodes, reflects the well-documented persistence and recurrence of affective disorders in midlife and older age, often accompanied by somatic comorbidities and functional decline. The inclusion of Z00.4 indicates sustained engagement with mental health services, which may be interpreted as a proxy for long-term monitoring or ongoing management of chronic affective conditions. These findings are consistent with epidemiological evidence showing that late-life depression is frequently recurrent, associated with anxiety symptoms, and often necessitates continued contact with healthcare providers (cf. [55, 56]).

Here, we have presented the case of disease communities with an over-expressed presence of diseases belonging to the category of mental and behavioral disorders (F class) but our results cover all ICD categories, and a similar detailed analysis can be performed for

any category of interest. In Sec. S7 of Supplementary Information, we discuss in detail a cluster of communities characterized by over-expression of category H. A brief note on SVNs of communities with C category over-expression is given in Sec. S8 of Supplementary Information.

3.4 Dismantling the comorbidity network

A classical topic in network theory is the investigation of the level of network resilience to random or targeted attacks [29]. This is interesting from two perspectives: on one side such levels of resilience are indicative of how much networks are robust against selective node removal. On the other side, looking for the best strategy to dismantle a network can give an indication of what are the nodes or group of nodes most relevant to sustain the connectivity of comorbidity networks. In the present context, identifying such peculiar nodes or groups of nodes, might give us information on the pathologies or sets of pathologies that are instrumental in the setting of comorbidities.

3.4.1 The dismantling procedure

The presence of a node in the comorbidity statistically validated network carries two types of information. On one hand it shows that comorbidity of the selected node occurs in a patient with neighboring nodes of the disease with a probability that it is not compatible with a null hypothesis based on the prevalence of each pair of diseases. The same node also signals the potential for patients having only this disease at a given time to be diagnosed with the neighboring diseases. Imagine a disease path characterized by disease A linked to disease B and B linked to disease C. Dismantling the comorbidity network (for example due to preventive medicine policies affecting disease B) by removing disease B would diminish the probability that groups of patients affected by disease A develop comorbidities with disease C (and viceversa).

The dismantling process is applied to a slightly modified version of the comorbidity networks discussed in Sect. 3.1. Specifically, for each cohort the modification we make consists in removing all nodes with ICD codes belonging to categories R and Z. These nodes concern symptoms or hospital procedures/check-ups. Thus, removing R and Z nodes from the network would have a different meaning than removing a node representing an actual disease. In this section, we want to focus on how "eradicating" a disease can alter the structure of the PROJ network or SVN, and how important each one of them is in holding the largest connected component together.

3.4.2 An illustrative example

In Fig. 8, we show results for the dismantling process for a specific cohort of patients (female, aged 60-69). More specifically, we plot the sizes of the largest and second largest connected components as the removal process unfolds. In the plot, we illustrate the dismantling process of the PROJ network (black and red lines) and of the corresponding SVN (blue and green lines). We find that the dismantling procedure of the SVN is more efficient than that for the PROJ network, in the sense that it requires the removal of only about 15% of the nodes of the SVN largest component (which contains about 57% of the nodes of the corresponding projected network). Indeed, for all cohorts, we verify that the dismantling of the SVN is more efficient than the dismantling of the PROJ network both in absolute and relative terms. The right-hand panel of Fig. 8 shows that the fraction of nodes that has

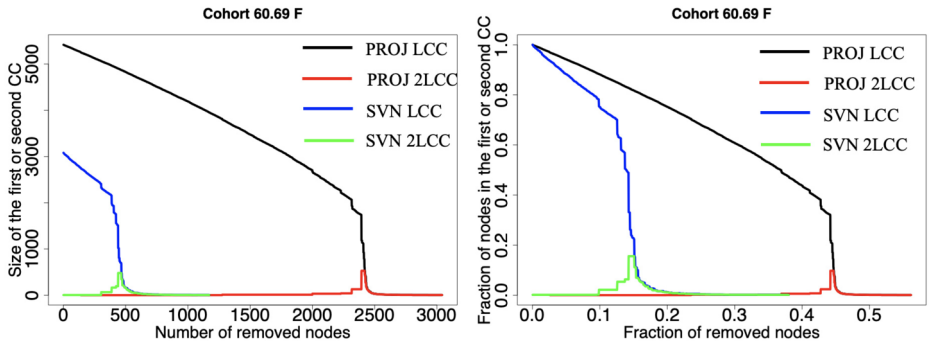

Figure 8 Dismantling process of the comorbidity networks obtained for female 60-69 cohort. In the two panels, we show the size of the largest connected component (black line for PROJ network and blue line for SVN) and the size of the second largest connected component (red line for the PROJ network and green line for the SVN) as a function of the number of removed nodes. The left panel shows the dismantling process in absolute terms, whereas the right panel in terms that are relative to the size of the LCC

to be removed from the PROJ network to cause its collapse is more than double than the fraction that needs to be removed from the SVN.

3.4.3 Categories of dismantling nodes

We now investigate the type of ICD codes that contribute most to the dismantling of the comorbidity network of a given cohort. For each cohort, we first select a set of nodes comprising the first N_{dism} ICD codes whose removal (as a whole) reduce the size of the largest connected component of the SVN to 10% of its original size. We call this set of nodes the 'dismantling set'. The number N_{dism} will be specific to each cohort. For example, in the dismantling of the SVN network of the female cohort of age 60 to 69 shown in the left panel of Fig. 8 the largest connected component has 3081 nodes after we remove R and Z nodes. During dismantling, the largest connected component crosses the 10% value of this initial size, i.e., 308 nodes when we remove N_{dism} = 481 nodes. These nodes are part of what we call the dismantling set.

For each disease category, we calculate the fraction of nodes of the category that are in the dismantling set and we compare this fraction to the average fraction observed by randomly selecting a set of ICD nodes of the same size of the dismantling set from the analyzed SVN. By repeating the random selection 1000 times we estimate the average fraction and its standard deviation. The comparison between the observed fraction and simulated value allows us to detect the relative abundance of nodes of a specific category observed in the dismantling set for each cohort. Deviations of the relative abundance of an ICD category in the dismantling set from the average value obtained through random node selection indicate that the dismantling is not merely driven by the baseline prevalence of a disease category within the cohort. Rather, such deviations reflect a specific and non-trivial contribution of that disease family to the structural robustness (or fragility) of the comorbidity network.

In Fig. 9 we show two examples of the fraction of ICD nodes in different categories among dismantling sets in the different age cohorts for females (left-hand panels) and males (right-hand panels). Specifically, we focus on categories E (endocrine, nutritional and metabolic diseases, top panels) and M (diseases of the musculoskeletal system and

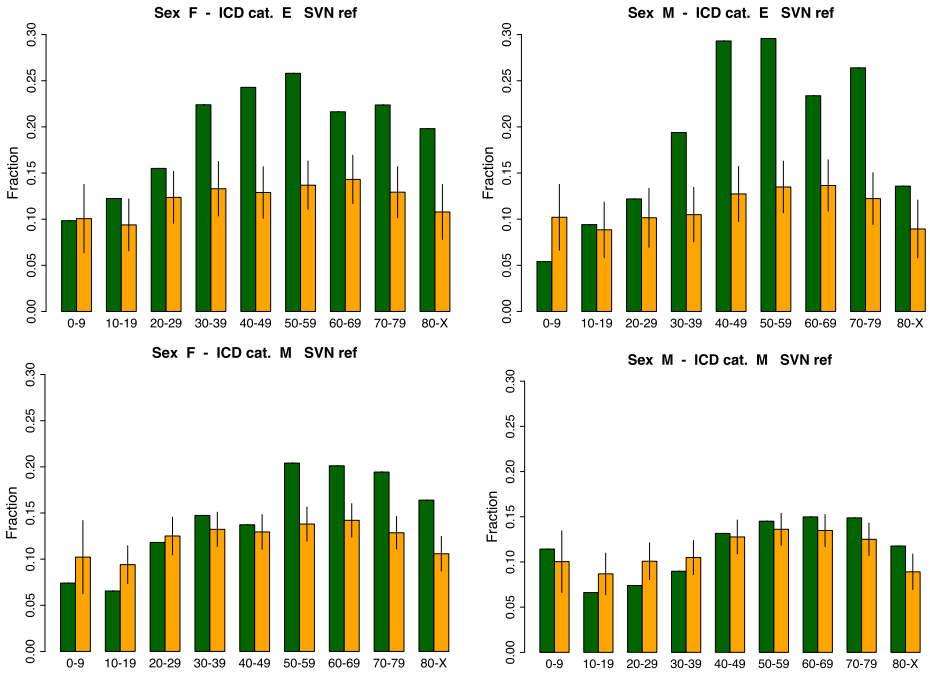

Figure 9 Dark green bars are the fraction (dark green bars) of ICD codes present in the dismantling set for the E category: Endocrine, nutritional and metabolic diseases (top panels), and for the M category: Diseases of the musculoskeletal system and connective tissue (bottom panels), for all age cohorts of females (left panels) and males (right panels). Orange bars are the corresponding fraction of category E and M obtained by random selection of the same amount of nodes for the SVN networks and for each category and cohort. Black segment on top of the orange bar displays plus and minus one standard deviation observed across 1000 realizations of the sampling procedure

connective tissue, bottom panels). Dark green bars in the figure represent the fraction observed in the dismantling set for the cohort in hand, and orange bars the average fraction of 1000 random selections of nodes of the same size of the dismantling set from the corresponding SVN. Vertical black line on top of the orange bar indicates plus and minus one standard deviation across the 1000 samples.

The two examples in Fig. 9 show two typical patterns. Specifically, the case of category E (upper panels) shows a fraction pattern for different age cohorts that is quite similar for both sexes. The fraction of the E category of diseases is compatible with the average fraction observed for an equivalent set of ICD codes randomly selected within one standard deviation (black line on top of the orange bars in the figure) for age cohorts 0-9 for females (left top panel of Fig. 9), 10-19 and 20-29 for males (right top panel of Fig. 9). For females a higher fraction than the random control is observed for the age cohorts from 10-19 to 80-XX whereas for males the cohort 0-9 shows a lower fraction than the random control and cohorts from 30-39 to 80-XX show a higher fraction that the random control.

In summary, in the age period from 30 to 80 years old, endocrine, nutritional, and metabolic diseases have a localization in the comorbidity networks that gives these conditions high values of node betweenness therefore putting these diseases on a large number of diseases paths observed in the comorbidity networks.

A quite different pattern is observed for diseases of category M, see the lower panels in Fig. 9. These are conditions related to the musculoskeletal system. For male patients,

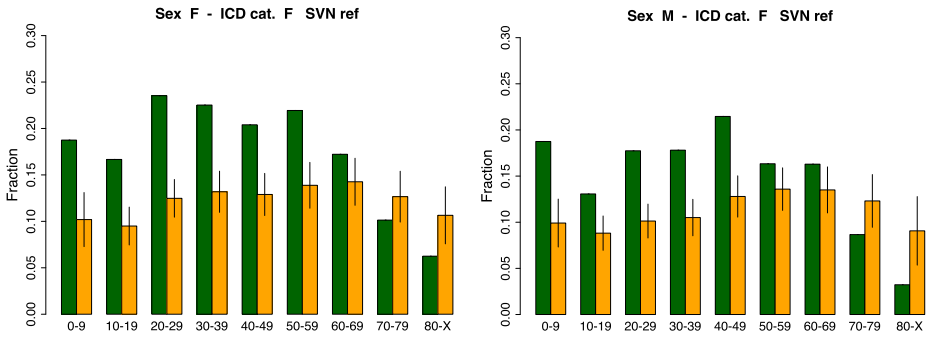

Figure 10 Dark green bars are the fraction (dark green bars) of ICD codes present in the dismantling set for the F category: Mental and behavioural disorders, for all age cohorts of females (left panel) and males (right panel). Orange bars are the corresponding fraction of category F obtained by random selection of the same amount of nodes for the SVN networks and for each category and cohort. Black segment on top of the orange bar displays plus and minus one standard deviation observed across 1000 realizations of the sampling procedure

the fraction of category M ICD codes in the dismantling set is approximately consistent within one standard deviation with the one estimated from the random sampling from the SVN network. In other words, for males, we verify that category-M diseases do not sustain resilience of the comorbidity network relative to the basic level expected by random co-occurrence. However, this observation is only valid for males. In fact, for females, we detect a different pattern. There is an abrupt fraction enhancement at the 50-59 age group. For higher ages, the enhancement is observed for all female cohorts.

As a last example of a disease category with enhanced fraction in the dismantling set of nodes across different cohorts, we comment on ICD category F. These are the mental and behavioral disorders we also discussed in Sect. 3.3. In Fig. 10, we notice that F ICD nodes show a higher fraction in dismantling nodes of younger cohorts both for females and males compared to the average value of random sampling. The presence of F code diseases is enhanced up to cohorts of age 60 to 69 in females and 50-59 in males. There are different explanation for the key role of diseases in the F group for cohorts involving children and teens (cohorts of 0-9 and 10 to 19 of age) and cohorts involving young adults (cohorts of 20-29 of age), respectively. In the former cohorts, the ICD codes of type F involve diseases associated with mental retardation, whereas starting from late teens to young adults different forms of depression play a major role. The female (left-hand panel) and male (right-hand panel) patterns are similar but not identical across different age cohorts. This suggests a certain degree of sex-specificity associated with different impacts of mental retardation and different degree of impact/resilience with respect to depression.

Previous examples show that the dismantling method is able to highlight the classes of diseases that have a localization in the comorbidity networks that gave them high values of node betweenness therefore shortening the paths of the comorbidity network, providing potential suggestions for preventive medicine policies.

4 Discussion and conclusions

In this paper, we investigate comorbidity patterns observed in a large set of patients diagnosed in southern Finland in a time period of 15 years. Evidence of comorbidity patterns is obtained by using information about diseases diagnoses that are present in the historical

medical records of patients of a large population. Starting from a bipartite network of patients and diseases we construct comorbidity networks of diseases where a link indicates an over-expression of comorbidity that is statistically validated against a null hypothesis of random co-occurrence of a pair of diseases in patients of a given age and sex.

Our study is performed at so-called level-4 of the ICD WHO codification. It is worth recalling that this level is the one primarily used in medical diagnoses. Therefore, this level of granularity is probably the one that better describes the information summarized in diagnoses by medical doctors. The redundant or inessential information associated with this high level of granularity is taken into account by choosing a statistical validation methodology that is robust with respect to the heterogeneity of the nodes (e.g., with respect to the different prevalence of diseases) and by performing a careful control of the family-wise error rate.

We have verified that comorbidity SVNs show information that is hidden in basic PROJ networks. The construction of SVNs allows the unsupervised detection of groups of diseases. ICD communities are obtained by using community detection algorithms introduced in complex network research. Disease communities obtained in this way can be used as supporting information for healthcare policy decisions. For example, mental and behavioral disorders have a strong impact in comorbidity of cohorts of young patients of both sexes. Conversely, endocrine, nutritional, and metabolic diseases present a more prominent role in sustaining comorbidity pattern in age cohorts ranging from the 30 s to the late 80 s with a similar impact for both sexes.

SVNs are also informative with respect to the category (or categories) of diseases that need attention to fragment comorbidity SVNs for different cohorts of age and sex. In fact, the nature and structure of SVNs vary for different cohorts and the relative role of different categories of diseases can be comparatively assessed.

A prominent example of different behavior is observed for the diseases of the musculoskeletal system and connective tissue where a distinct impact is seen for females of age from 50 s to late 70 s whereas no apparent role is detectable for males at any age.

A notable limitation of our study, as well as of similar research employing pairwise disease interactions to construct comorbidity networks, lies in the omission of higher-order interactions. By focusing exclusively on dyadic relationships between diseases, this approach overlooks the potentially significant role of multi-disease associations that may jointly contribute to disease patterns and progression. However, pairwise SVNs and statistical validation of higher-order sets of network nodes (e.g., statistically validated hyperlinks) address complementary aspects of comorbidity structure rather than constituting mutually exclusive or hierarchically superior descriptions. SVNs, by construction, isolate dyadic disease associations that cannot be explained by marginal prevalences alone, thereby providing a statistically controlled backbone of clinically meaningful co-occurrence relationships. This filtering step is particularly important in large-scale electronic health record data, where raw projections are extremely dense and dominated by ubiquitous or highly prevalent conditions. As a consequence, pairwise SVNs enable the identification of robust mesoscale structures—such as disease communities, central bridging conditions, and cohort-specific connectivity patterns—that are directly interpretable and amenable to comparison across age and sex strata. These features capture fundamental organizing principles of comorbidity, including pathways of disease progression

and key conditions sustaining network cohesion, which remain well-defined even when higher-order interactions are not explicitly modeled.

Higher-order approaches, in contrast, are designed to capture joint, non-reducible interactions among three or more diseases, potentially revealing synergistic or conditional dependencies that are invisible at the dyadic level. However, the presence of such higher-order effects does not invalidate the information encoded in statistically validated pairwise links. Rather, the two levels of description probe different projections of the same underlying system: pairwise SVNs identify statistically robust building blocks of comorbidity, while higher-order structures refine these patterns by highlighting multi-disease contexts in which such associations are embedded. In this sense, pairwise and higher-order analyses are complementary layers of inference, each with distinct methodological requirements and interpretative scope. The absence of higher-order modeling does not render pairwise findings unreliable; instead, it delineates a well-defined level of description that is both statistically grounded and clinically informative, and that can serve as a necessary foundation for future extensions toward higher-order network representations.

In summary, information stored in electronic health records of a large population of patients, whose history is recorded for many years, successfully contributes to highlighting the role of specific categories of diseases and of specific prominent diseases in the setting of complex comorbidity patterns observed in a large cohort of patients of specific age and sex.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-026-00651-4.

Additional file 1. (PDF 930 kB)

Acknowledgements
The project that gave rise to these results received the support of a fellowship from "la Caixa" Foundation (ID 100010434). The fellowship code is LCF/BQ/DI22/11940041. PC and TG are also grateful for partial financial support from the Agencia Estatal de Investigación and Fondo Europeo de Desarrollo Regional (FEDER, UE) under project APASOS (PID2021-122256NB-C21, PID2021-122256NB-C22), and the María de Maeztu program for Units of Excellence, CEX2021-001164-M funded by MCIN/AEI/10.13039/501100011033. RNM, SM and JP acknowledge financial support of the Italian PRIN research project P2022JAYMH "Higher-order complex systems modeling for personalized medicine" funded by NextGenerationEU.

Author contributions
P.C., A.K. , J.P., and R.N.M. conceived the study, P.C. and R.N.M. wrote scripts analysing data, P.C. and R.N.M. analysed data, P.C., T.G., A.K., S.M., J.P. and R.N.M. analysed the results. P.C., T.G., S.M., J.P. and R.N.M. wrote the manuscript. All authors reviewed the manuscript.

Funding information
Fellowship from "la Caixa" Foundation (ID 100010434). The fellowship code is LCF/BQ/DI22/11940041.
Partial financial support from the Agencia Estatal de Investigación and Fondo Europeo de Desarrollo Regional (FEDER, UE) project APASOS (PID2021-122256NB-C21, PID2021-122256NB-C22).
Partial financial support from María de Maeztu program for Units of Excellence, CEX2021-001164-M funded by MCIN/AEI/10.13039/501100011033.
Financial support of the Italian PRIN research project P2022JAYMH "Higher-order complex systems modeling for personalized medicine" funded by NextGenerationEU.

Data availability
The data investigated in this study are proprietary data of Auria Clinical Informatics which operates in connection with Varha. Data can be accessed with permission from Varha. The present study analyzes disease networks obtained with the approval of the Institutional Review Board of Turku University Hospital (license number T152/2017 [32]). Informed consent was waived due to the study's retrospective design, according to Finnish legislation

Declarations

Competing interests
The authors declare no competing interests.

Author details
[1]IFISC, Instituto de Fisica Interdisciplinar y Sistemas Complejos (CSIC-UIB), Campus Universitat de les Illes Balears, 07122 Palma de Mallorca, Spain. [2]Auria Biobank, Turku, Finland. [3]Dipartimento di Fisica e Chimica Emilio Segrè, Università degli Studi di Palermo, Palermo, Italy. [4]Department of Physics and Astronomy, University of Turku, Turku, Finland. [5]Complexity Science Hub, Vienna, Austria.

Received: 27 August 2025 Accepted: 23 March 2026 Published online: 14 April 2026

References

1. Newman M (2018) Networks. Oxford university press
2. Barabási A, Pósfai M (2016) Network science. Cambridge University Press, Cambridge
3. Latora V, Nicosia V, Russo G (2017) Complex networks: principles, methods and applications. Cambridge University Press, Cambridge
4. Barabási A, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. Nat Rev, Genet 12:56–68
5. Ivanov P (2021) The new field of network physiology: building the human physiolome. Front Netw Physiol 1, Article ID 711778
6. International Statistical Classification of Diseases and Related Health Problems 10th Revision. https://icd.who.int/browse10/2019/en
7. Lee D, Park J, Kay K, Christakis N, Oltvai Z, Barabási A (2008) The implications of human metabolic network topology for disease comorbidity. Proc Natl Acad Sci USA 105:9880–9885
8. Hidalgo C, Blumm N, Barabási A, Christakis N (2009) A dynamic network approach for the study of human phenotypes. PLoS Comput Biol 5:e1000353
9. Folino F, Pizzuti C, Ventura M (2010) A comorbidity network approach to predict disease risk. In: International conference on information technology in bio-and medical informatics, pp 102–109
10. Roque F, Jensen P, Schmock H, Dalgaard M, Andreatta M, Hansen T, Søeby K, Bredkjær S, Juul A, Werge T, et al (2011) Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS Comput Biol 7:e1002141
11. Chmiel A, Klimek P, Thurner S (2014) Spreading of diseases through comorbidity networks across life and gender. New J Phys 16, Article ID 115013
12. Jensen A, Moseley P, Oprea T, Ellesøe S, Eriksson R, Schmock H, Jensen P, Jensen L, Brunak S (2014) Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nat Commun 5:4022
13. Jeong E, Ko K, Oh S, Han H (2017) Network-based analysis of diagnosis progression patterns using claims data. Sci Rep 7:15561
14. Bao Y, Lu P, Wang M, Zhang X, Song A, Gu X, Ma T, Su S, Wang L, Shang X, et al (2023) Exploring multimorbidity profiles in middle-aged inpatients: a network-based comparative study of China and the United Kingdom. BMC Med 21:495
15. Varha is responsible for organizing social and health services and rescue operations within the county and is made up of 27 municipalities, a hospital district, special care services and rescue services. https://www.varha.fi/en
16. Auria Clinical Informatics. https://www.auria.fi/en/
17. Fotouhi B, Momeni N, Riolo M, Buckeridge D (2018) Statistical methods for constructing disease comorbidity networks from longitudinal inpatient data. Appl Netw Sci 3:1–34
18. Tumminello M, Miccichè S, Lillo F, Piilo J, Mantegna R (2011) Statistically validated networks in bipartite complex systems. PLoS ONE 6, Article ID e17994. https://doi.org/10.1371/journal.pone.0017994
19. Li M, Palchykov V, Jiang Z, Kaski K, Kertész J, Micciche S, Tumminello M, Zhou W, Mantegna R (2014) Statistically validated mobile communication networks: the evolution of motifs in European and Chinese data. New J Phys 16, Article ID 083038
20. Hatzopoulos V, Iori G, Mantegna R, Micciche S, Tumminello M (2015) Quantifying preferential trading in the e-MID interbank market. Quant Finance 15:693–710
21. Serrano M, Boguná M, Vespignani A (2009) Extracting the multiscale backbone of complex weighted networks. Proc Natl Acad Sci USA 106:6483–6488
22. Radicchi F, Ramasco J, Fortunato S (2011) Information filtering in complex weighted networks. Phys Rev E, Stat Nonlinear Soft Matter Phys 83, Article ID 046101
23. Saracco F, Di Clemente R, Gabrielli A, Squartini T (2015) Randomizing bipartite networks: the case of the world trade web. Sci Rep 5:10595
24. Saracco F, Straka M, Di Clemente R, Gabrielli A, Caldarelli G, Squartini T (2017) Inferring monopartite projections of bipartite networks: an entropy-based approach. New J Phys 19, Article ID 053022
25. Marcaccioli R, Livan G (2019) A Pólya urn approach to information filtering in complex networks. Nat Commun 10:745
26. Kobayashi T, Takaguchi T, Barrat A (2019) The structured backbone of temporal social ties. Nat Commun 10:220
27. Vallarano N, Bruno M, Marchese E, Trapani G, Saracco F, Cimini G, Zanon M, Squartini T (2021) Fast and scalable likelihood maximization for exponential random graph models with local constraints. Sci Rep 11:15227
28. Fortunato S, Hric D (2016) Community detection in networks: a user guide. Phys Rep 659:1–44
29. Albert R, Jeong H, Barabási A (2000) Error and attack tolerance of complex networks. Nature 406:378–382. https://doi.org/10.1038/35019019
30. Braunstein A, Dall'Asta L, Semerjian G, L Z (2016) Network dismantling. Proc Natl Acad Sci USA 113:12368
31. Wandelt S, Sun X, Feng D, Zanin M, Havlin S (2018) A comparative analysis of approaches to network-dismantling. Sci Rep 8:13513

32. Kivelev J, Saarenpää I, Karlsson A, Crisafulli P, Musciotto F, Piilo J, Mantegna R (2024) Complex networks approach to study comorbidities in patients with unruptured intracranial aneurysms. Sci Rep 14:4. https://doi.org/10.1038/s41598-024-59919-2
33. Rupert G (2012) Simultaneous statistical inference. Springer, Berlin
34. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc, Ser B Stat Methodol 57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
35. SValNet GitHub repository. https://github.com/complexParide/svalnet
36. Newman M, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69, Article ID 026113
37. Traag V, Waltman L, Eck N (2019) From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 9:5233
38. Karrer B, Newman M (2011) Stochastic blockmodels and community structure in networks. Phys Rev E 83, Article ID 016107
39. Rosvall M, Bergstrom C (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci USA 105:1118–1123
40. Juher D, Saldaña J (2018) Tuning the overlap and the cross-layer correlations in two-layer networks: application to a susceptible-infectious-recovered model with awareness dissemination. Phys Rev E 97, Article ID 032303. https://link.aps.org/doi/10.1103/PhysRevE.97.032303
41. Kao T, Porter M (2018) Layer communities in multiplex networks. J Stat Phys 173:1286–1302. https://doi.org/10.1007/s10955-017-1858-z
42. Kaufman L, Rousseeuw P (2009) Finding groups in data: an introduction to cluster analysis. Wiley, New York
43. Tumminello M, Micciche S, Lillo F, Varho J, Piilo J, Mantegna R (2011) Community characterization of heterogeneous complex systems. J Stat Mech Theory Exp 2011:P01019
44. Musciotto F, Miccichè S (2023) Exploring the landscape of dismantling strategies based on the community structure of networks. Sci Rep 13:14448
45. Onyike C, Diehl-Schmid J (2013) The epidemiology of frontotemporal dementia. Int Rev Psychiatry 25:130–137
46. Urso D, Giannoni-Luza S, Brayne C, Ray N, Logroscino G (2025) Incidence and prevalence of frontotemporal dementia: a systematic review and meta-analysis. JAMA Neurol
47. Gardner R, Valcour V, Yaffe K (2013) Dementia in the oldest old: a multi-factorial and growing public health issue. Alzheimer's Res Ther 5:27
48. Rodriguez F, Pabst A, Heser K, Kleineidam L, Hajek A, Eisele M, Röhr S, Löbner M, Wiese B, Angermeyer M (2021) Others disorientation in time and place in old age: longitudinal evidence from three old age cohorts in Germany (AgeDifferent. De platform). J Alzheimer's Dis 79:1589–1599
49. Hirschfeld R (2001) The comorbidity of major depression and anxiety disorders: recognition and management in primary care. Prim Care Companion J Clin Psychiat 3:244–254. https://pmc.ncbi.nlm.nih.gov/articles/PMC181193/, PMCID: PMC181193
50. Groen R, Snippe E, Jeronimus B, Jonge P, Wichers M (2020) Comorbidity between depression and anxiety: assessing the role of bridge mental states in dynamic psychological networks. BMC Med 18:308. https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-020-01738-z
51. Claussen A, Robinson Z, Ivanova M, Agha S, Caye A, Buitelaar J, Sonuga-Barke E (2022) All in the family? A systematic review and meta-analysis of parenting and family environment as risk factors for attention-deficit/hyperactivity disorder (ADHD) in children. Neurosci Biobehav Rev 133, Article ID 104518. https://pmc.ncbi.nlm.nih.gov/articles/PMC9017071/
52. Agha S, Zammit S, Thapar A, Langley K, Collishaw S, Banaschewski T, Taylor E, Asherson P, Kuntsi J, Faraone S, O'Donovan M, Thapar A, Thapar A (2013) Are parental ADHD problems associated with a more severe clinical presentation and greater family adversity in children with ADHD? Eur Child Adolesc Psychiatry 22:369–377. https://link.springer.com/article/10.1007/s00787-013-0378-x
53. Mantere O, Isometsä E, Ketokivi M, Kiviruusu O, Suominen K, Valtonen H, Arvilommi P, Leppämäki S (2010) A prospective latent analyses study of psychiatric comorbidity of DSM-IV bipolar I and II disorders. Bipolar Disord 12:271–284
54. Berg N, Kiviruusu O (2024) Trends in the co-occurrence and association between heavy episodic drinking and generalized anxiety among adolescents between 2013 and 2023 in Finland. Int J Ment Health Addict 22
55. Alexopoulos G (2005) Depression in the elderly. Lancet 365:1961–1970
56. Copeland J, Beekman A, Dewey M, Hooijer C, Jordan A, Lawlor B, Lobo A, Magnusson H, Mann A, Meller I, Prince M, Reischies F, Turrina C, DeVries M, Wilson K (2004) Depression in Europe: geographical distribution among older people. Br J Psychiatry 185:68–75

Publisher's note