Large scale statistically validated comorbidity networks

Paride Crisafulli and Tobias Galla
IFISC, Instituto de Fisica Interdisciplinar y Sistemas Complejos (CSIC-UIB), Campus Universitat de les Illes Balears, 07122 Palma de Mallorca, Spain

Antti Karlsson
Auria Biobank, Turku, Finland.

Salvatore Micciché
Dipartimento di Fisica e Chimica Emilio Segrè, Università degli Studi di Palermo, Palermo, Italy

Jyrki Piilo
Department of Physics and Astronomy, University of Turku, Turku, Finland.

Rosario N. Mantegna
Dipartimento di Fisica e Chimica Emilio Segrè, Università degli Studi di Palermo, Palermo, Italy and Complexity Science Hub, Vienna, Austria

(Dated: October 14, 2025)

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 associate 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, all 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.

I. INTRODUCTION

Network modeling is a powerful and flexible tool for the investigation of complex systems [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 Sec. II B.

In our analysis, we start from a bipartite patient-disease network in which all 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 patient in the target cohort who has been diagnosed with both conditions. We refer to this network as 'PROJ' (for projected network).

We assume 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. This network contains a subset of all co-occurrences of ICD codes in the dataset. 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 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 [28]. Specifically, we used the so-called 'control of the false discovery rate' [29].

Having constructed the PROJ networks and the SNVs for different cohorts of patients, we search communities of diseases in both types of network using community-detection methods for complex networks [30]. This provides communities of diseases characterized by pronounced intra-community connectivity. While the original PROJ networks are so dense that Infomap, a widely used community detection algorithm [31], is unsuccessful in detecting distinct clusters of diseases, the same algorithm finds several clusters of diseases in the SVNs. The 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 comorbidities for different cohorts of patients.

In fact, we verify that different cohorts of patients are characterized by distinct sets of communities in the SVNs. We provide a comparison of comorbidity communities in different patient cohorts by computing a similarity measure between them and grouping them by using agglomerative hierarchical clustering [32]. The inspection of the hierarchical tree is supported by a statistical test detecting the category or categories of diseases with an over-expressed number in each community (see Section III B 4). With this characterization, we note that communities in which the same disease classes 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.

We also provide a specific example of ICD community analysis by discussing in some detail communities with an over-expressed presence of mental and behavioral disorders. A similar detailed analysis can be performed for any disease category of interest.

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 might be characterized by a high value of node betweenness, making these diseases crucial to sustain the structure 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 PROJ networks and SVNs. This analysis would 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' [33, 34] 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. We model network dismantling by a successive removal of ICD nodes from comorbidity networks, aiming at fragmenting the network into smaller, disconnected components. 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 several cohorts characterized by age and sex.

The remainder of the paper is organized as follows. In Section II we present the data and the methodology used to obtain comorbidity networks together with some summary information about the PROJ networks and SVNs. Section III A presents briefly the basic properties of the SVNs while in Section III B we discuss the disease communities detected in the SVNs. Section III C contains two case studies on mental and behavioral disorders, and in Section III D we describe results obtained by applying a dismantling procedure to PROJ networks and SVNs of different cohorts of patients. In Section IV we present our conclusions.

II. DATA AND METHODOLOGY

A. 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 [35]). Informed consent was waived due to the study's retrospective design, according to Finnish legislation on the secondary use of health data.

B. 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. Starting from 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 we list the disease categories classified by the first letter of ICD codes. In our study, we set the granularity of our study by using leve 4 ICD codes. This defines a set of 9303 different codes.

Data 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 an anonymized patient ID, the ICD code of the diagnosis, the timestamp of the visit, and the patient's metadata, such as age at the time of the visit and sex. Data entries include each diagnosis registered in the aforementioned time window in a hospital connected to Varha, regardless of the fact that the patient had been treated or not.

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 first-time 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.

C. 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 datset, 48.1% are male, and 51.9% female. Some summary statistics for the PROJ networks and SVNs are shown in Table II 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.

A-B	Certain infectious and parasitic diseases
C	Neoplasms
D	Neoplasms, diseases of the blood and blood-forming organs and certain immune disorders
E	Endocrine, nutritional and metabolic diseases
F	Mental and behavioural disorders
G	Diseases of the nervous system
H	Diseases of the eye and adnexa, diseases of the ear and mastoid process
I	Diseases of the circulatory system
J	Diseases of the respiratory system
K	Diseases of the digestive system
L	Diseases of the skin and subcutaneous tissue
M	Diseases of the musculoskeletal system and connective tissue
N	Diseases of the genitourinary system
O	Pregnancy, childbirth and the puerperium
P	Certain conditions originating in the perinatal period
Q	Congenital malformations, deformations and chromosomal abnormalities
R	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
S-T	Injury, poisoning and certain other consequences of external causes
U	Codes for special purposes
V-X-Y	External causes of morbidity and mortality
Z	Factors influencing health status and contact with health services

TABLE I. Definition of each ICD category (labeled by a single letter). Definitions are taken from the ICD-10 website [6]

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	46387	53130	44451	45549	71057	55076	64337	50887	57487
Bipartite links	192414	247572	236009	237602	470116	236183	583306	240946	469684
PROJ nodes	3904	4121	5250	4933	5904	5163	6113	5296	6184
PROJ links	215096	266002	407770	381156	711053	398215	911158	467127	964727
SVN nodes	1531	1810	2401	2276	3374	2437	3503	2549	3632
SVN links	9192	11629	14861	13652	31306	13178	37204	16624	37187

Number of	**40-49 M**	**50-59 F**	**50-59 M**	**60-69 F**	**60-69 M**	**70-79 F**	**70-79 M**	**80+ F**	**80+ M**
Patients	52821	66699	62776	65587	63156	55947	48046	41505	24463
Bipartite links	292536	506556	418254	568945	541195	578651	530509	539798	345711
PROJ nodes	5561	6349	5910	6130	6047	5860	5713	5429	4878
PROJ links	613123	1014863	832433	1083676	1020775	1051192	973948	859456	682662
SVN nodes	2897	3805	3438	3709	3632	3405	3103	2574	2122
SVN links	22068	38539	32072	39181	37627	34911	29397	20500	12379

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

As shown in Fig. 1, for all cohorts, the degree of ICD nodes in the bipartite network is highly heterogeneous while the degree of patient nodes is only moderately heterogeneous. The broader nature of the ICD distribution shows the need of a statistical validation.

D. Validating the networks

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

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 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

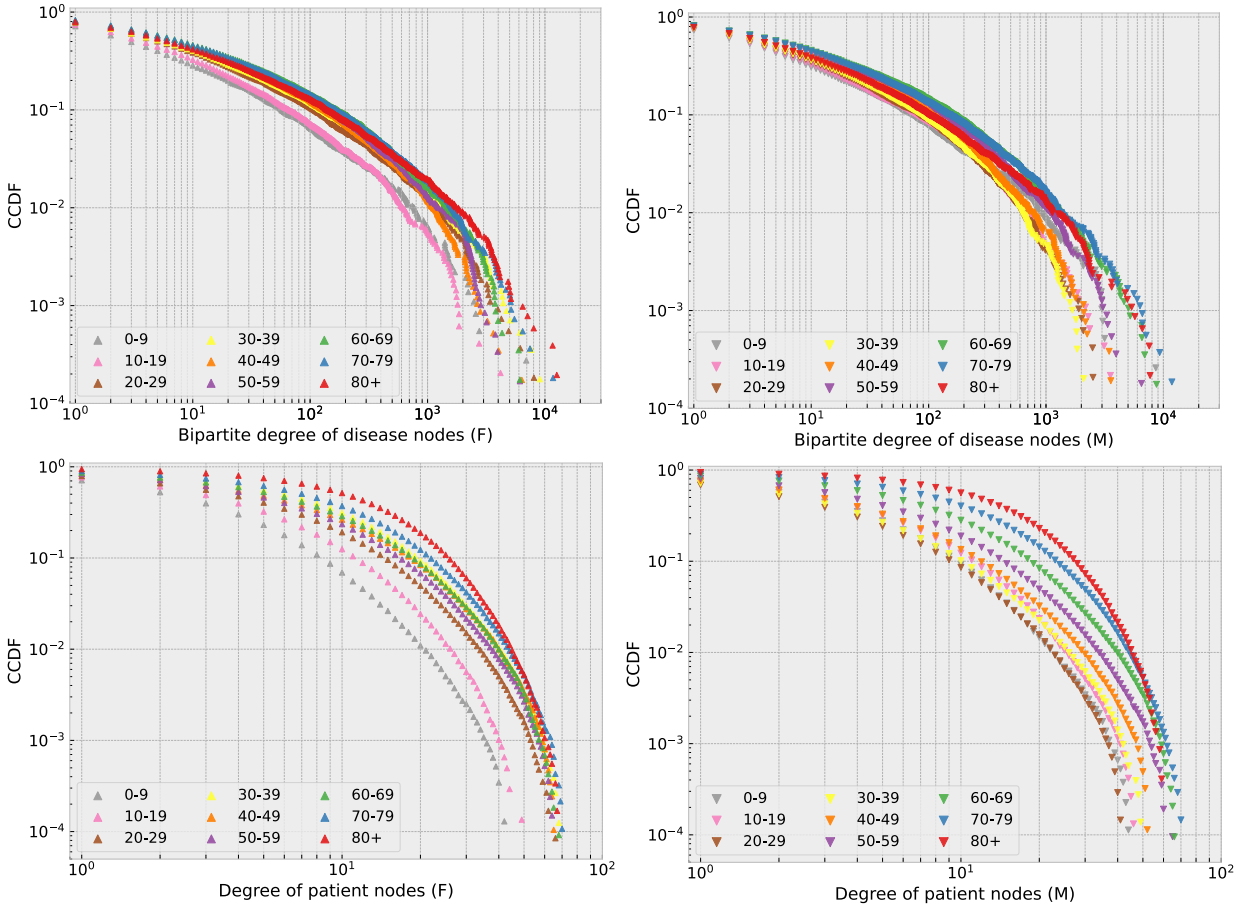

FIG. 1. Complementary cumulative distribution function of the degree of ICD nodes (top panels) and of patient nodes (bottom panels) in the bipartite network for all patient cohorts of females (left) and males (right).

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]:

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

Eq. 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 the hypothesis of random co-occurrence the probability of observing N_{AB} or more co-occurrences is:

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

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. To perform the hypothesis test, we set the univariate statistical threshold at 0.01 and then apply false discovery rate (FDR) correction [29] required when performing multiple hypothesis tests. If a link in the PROJ network leads to a p-value lower than expected for the FDR correction, 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

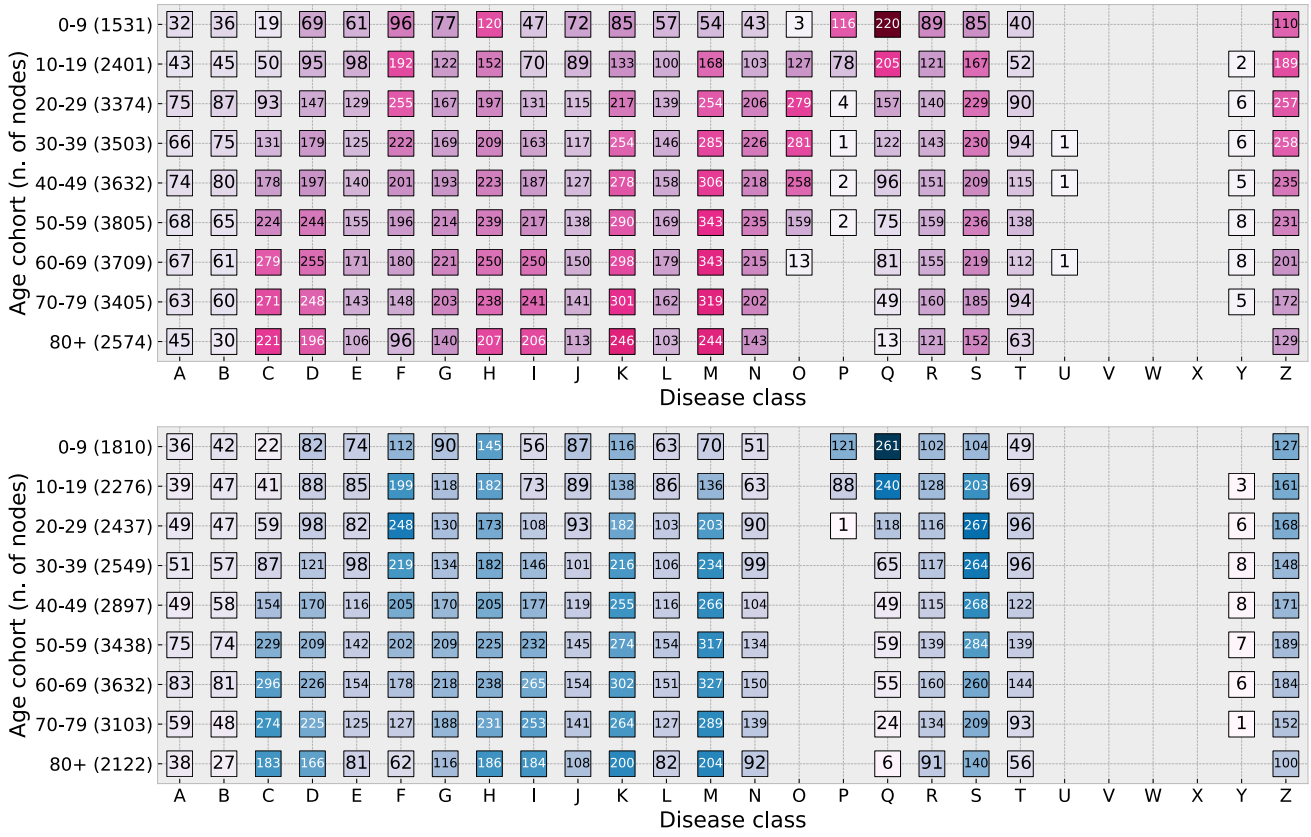

FIG. 2. Statistics of ICD category composition of SVNs for all cohorts of patients. ICD categories (columns) for different age cohorts (rows) and sexes (females in the upper panel and males in the lower). The number in brackets next to the age limits of the cohort is the total number of nodes in the 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

comorbidity link in the SVN. Finally, isolated nodes are not included in the SVN. The validation was performed by using the Python module accessible at the `svalnet` github repository [36].

III. RESULTS: COMORBIDITY NETWORKS

A. Properties of the validated networks

Table II 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 percent of retained nodes is ranging from about 40% to 60% therefore providing information on a large number of ICD lewvel 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. 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 observed between sexes are trivially the categories O (pregnancy, childbirth and the puerperiuas), and further N (diseases of the genitourinary system). A higher number of nodes of this category are seen for each age-specific female network (Fig. 2) compared to the male network for the same age bracket. Categories H, K, and M

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
Oct-19 F	2401	178	42	20	233
Oct-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

TABLE III. 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.

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.

B. Disease categories and community structure

1. *Community detection in comorbidity networks*

We now analyze the structure of comorbidity networks by looking at subgroups of ICD codes highly connected within each other. 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 [30]. We chose the Infomap algorithm [31] 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.

2. *Number of disease communities in different cohorts*

In Table III, 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 fact that the number of communities is bigger than one in each SVN shows how our method highlights more informative subnetworks than the PROJ ones, where Infomap fails to infer distinct communities and returns a single one for each connected component of each cohort. The difference in the number of communities between male and female SVNs does not reflect the difference in the number of nodes. This hints to the fact that communities are more influenced by disease class rather 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 to communities larger than 25 nodes.

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 decided to test this hypothesis by computing the similarity between pairs of SVN communities in different cohorts.

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:

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

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 sets of links).

Across all cohorts in Table III 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 Jaccard similarity, we define a distance for each pair of communities as $d(C_{k,a}, C_{l,b}) = 1 - J(C_{k,a}, C_{l,b})$. Using this distance we perform an agglomerative hierarchical clustering procedure. Agglomerative Hierarchical Clustering is an unsupervised bottom-up clustering method where each element starts in its own cluster, and clusters are iteratively merged based on a decreasing similarity (i.e., increasing distance) measure until all objects form a single cluster. 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 by 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 [32]. This leads to the 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.

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 in Fig. 16 in Appendix B. The rows and columns of the matrix are arranged in the order as they appear in the tree in Fig. 16 to efficiently visualize the clustering of groups of communities.

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 analogous to the one performed to detect statistically validated networks (details can be found in reference [37]). 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.

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. These two communities are located in the portion of hierarchical tree shown in Fig. 4. The selected branch of hierarchical tree shows that the clustering of communities involves cohorts of both sex and of different age intervals. This branch of communities shows an over-expression if

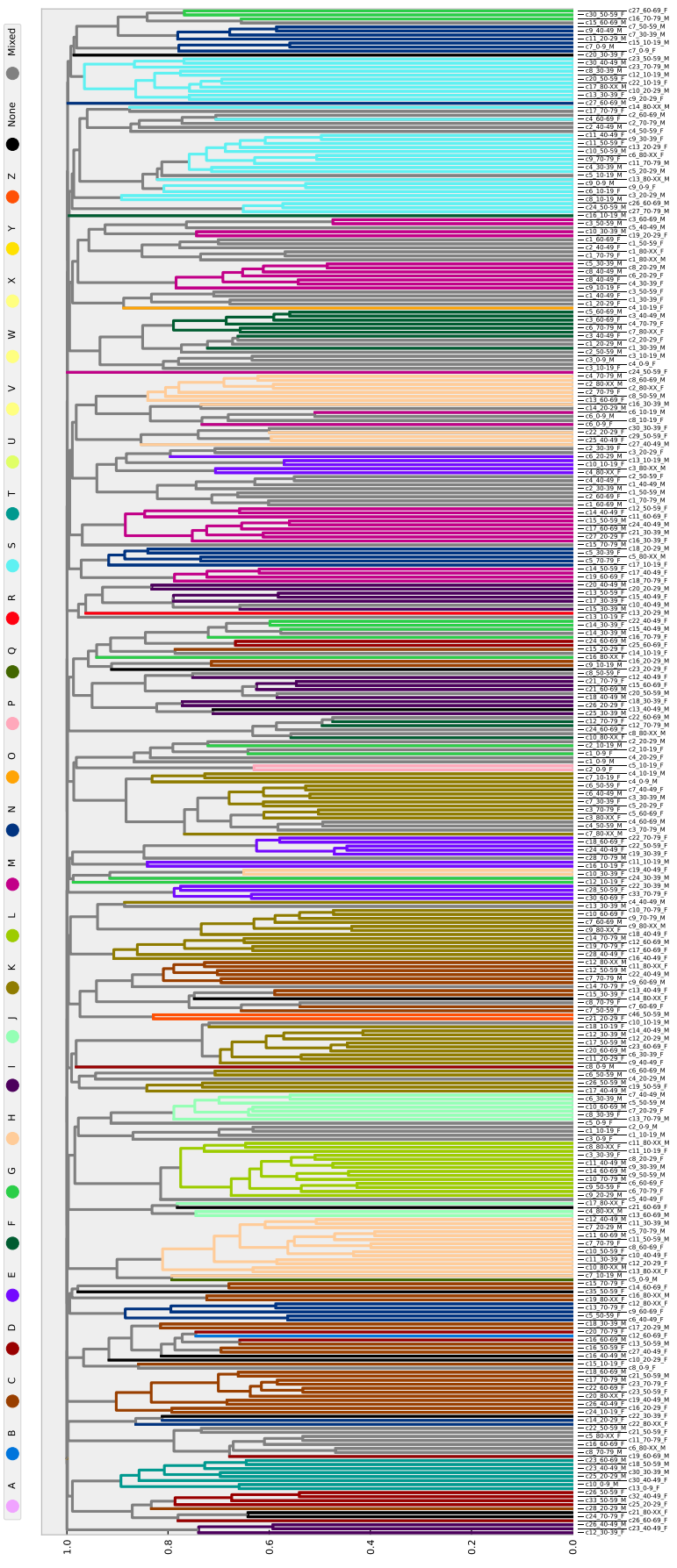

FIG. 3. Average linkage hierarchical tree of the 380 ICD communities with more than 25 nodes detected in SVNs of different cohorts. The average linkage tree is obtained by using $d = 1 - J$ as a dissimilarity measure. The color of each leaf of the tree indicates over-expression of an ICD category as 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 in black lines.

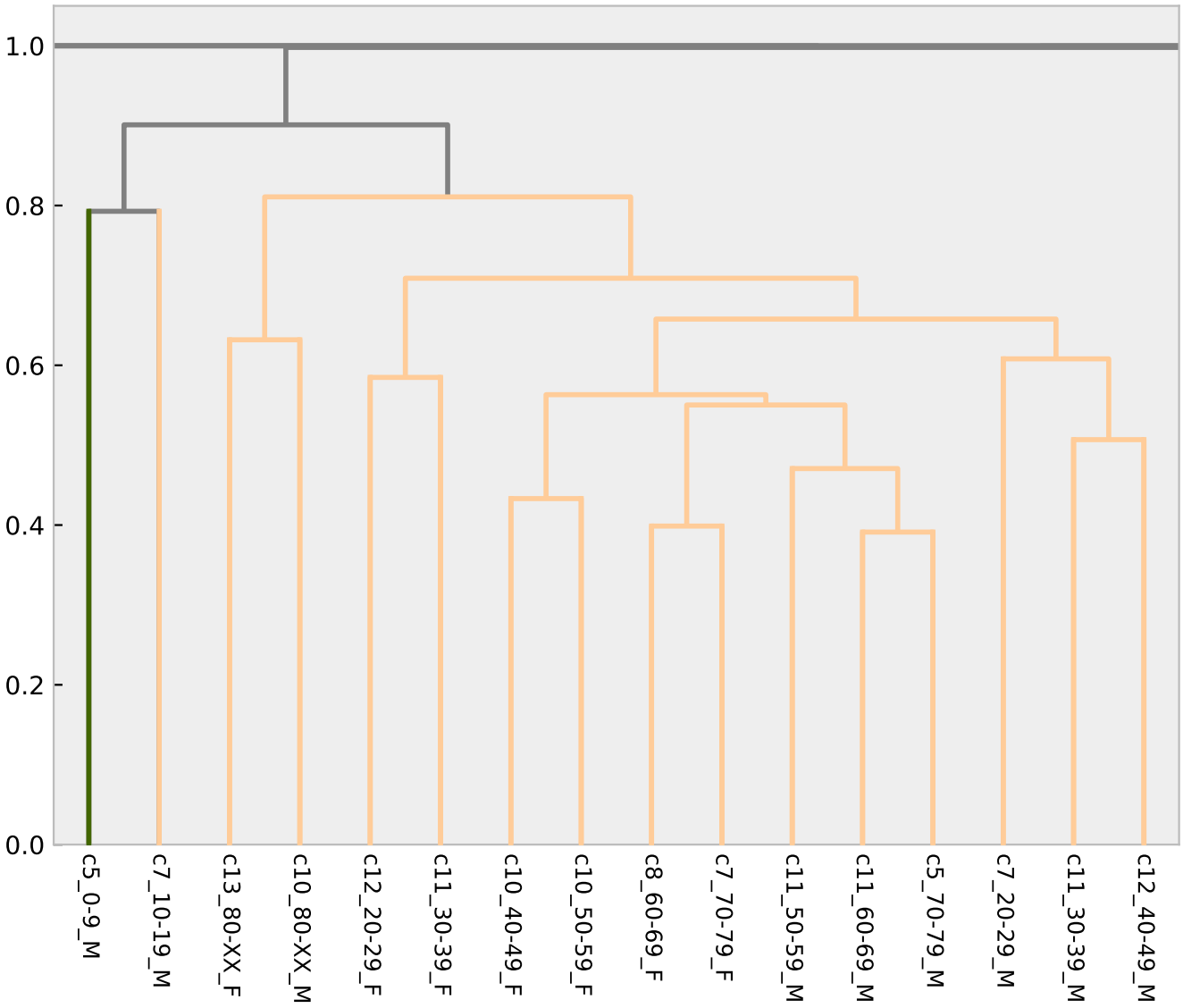

FIG. 4. Selected region of the average linkage hierarchical tree of the set of SVN communities with number of nodes larger than 25. The selected region contains the pair of ICD communities with the highest Jaccard similarity (c5_70-79_M and c11_60-69_M). All communities except one (c5_0-9_M) in this cluster present an over-expression of the ICD category H.

ICD codes in category H, and with numerical codes ranging from 60 to 90 therefore referring to diseases involving the ear and hearing.

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 the tables in Appendix A, 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 high level of homogeneity of over-expression of the same ICD category in subregions of the hierarchical tree shows that detected clusters are, in most cases, associated with a specific category of diseases. The list of over-expression categories also highlights the presence of a few large clusters characterized by over-expression of more than one ICD category. 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.

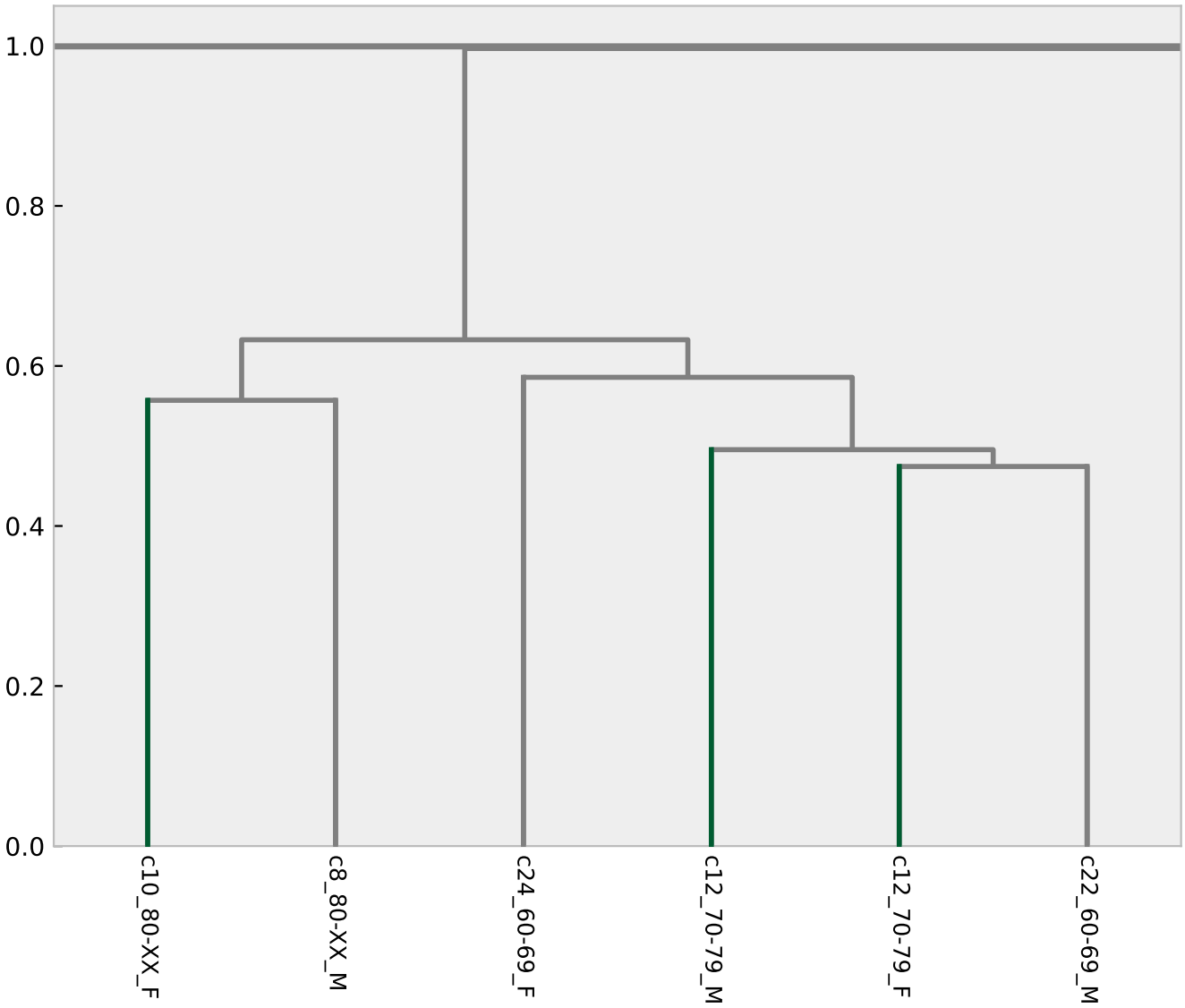

FIG. 5. Selected region of the average linkage hierarchical tree with communities located at positions from 199 to 204. The over-expression of disease category for each community starting from left is as follows F, FG, FG, F, F, FG.

C. 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 Appendix A). They are the cluster comprising communities located in the average linkage hierarchical tree from position 199 to 204 (see Fig. 5) and the cluster of communities located from position 291 to 305 (see Fig. 6).

The first group (Fig. 5) consists of 6 communities of ICD codes from cohorts with age intervals from 60-69 to 80+ for females and males. Four of the communities in this cluster are shown in Fig. 7. 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. 8, we show information about the 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

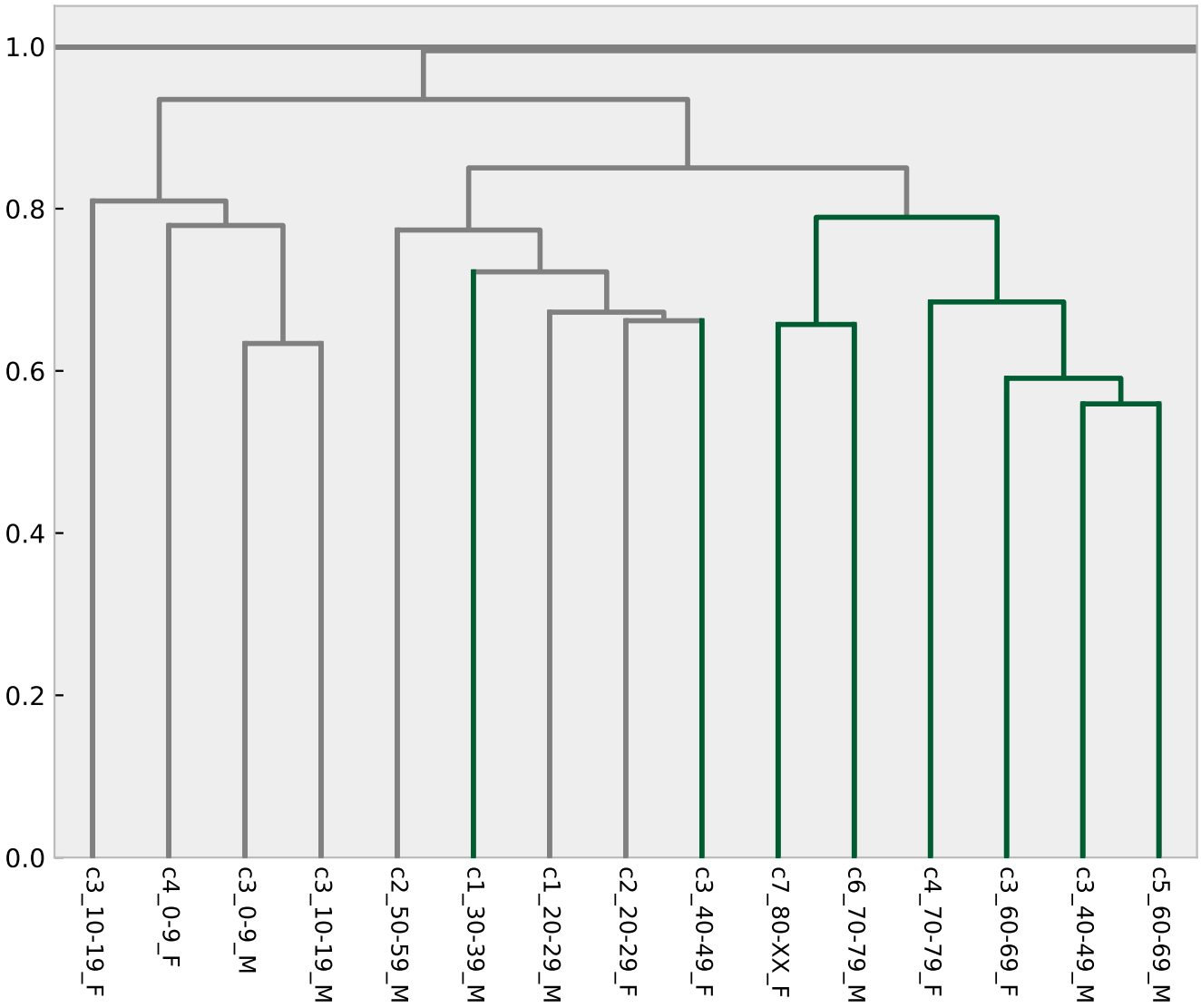

FIG. 6. Selected region of the average linkage hierarchical tree with communities located at positions from 291 to 305. The over-expression of categories for each community, starting from left, is as follows FYZ, FZ, FZ, FZ, FS, F, FZ, FZ, F, F, F, F, F, F, F.

specific for degenerative disease of the nervous system affecting both females and males starting from the age of sixties. ICD codes R concern primarily symptoms, and ICD codes Z primarily describe anamnesis and/or specific health procedures or circumstances (see Table I for the full definitions of all categories). The most prominent ICD code indicating symptoms we note is R41.8 ("other and unspecified symptoms and signs involving cognitive functions and awareness"), i.e., the most basic evaluation accompanying degenerative diseases of the nervous system. For the community c8_80.XX_M we also note "Other and unspecified convulsions" that in this community is connected to several diseases of the group G40.* used for epilepsy.

The second group of ICD communities (see Fig. 6) consists of three distinct subclusters. These subclusters are different in terms of age intervals and prominence of certain types of mental and behavioral disorders. The subcluster composed of c3_10.19_F, c4_0.9_F, c3_0.9_M, and c3_10.19_M (first branch on the left region of Fig. 6) covers the age cohorts of childhood and adolescence and, in fact, the ICD codes are primarily in the range F9*. This group of codes corresponds to "behavioral and emotional disorders with onset occurring in childhood and adolescence". In addition to the prominent F and Z categories, these communities also show some ICD codes of categories P (certain conditions originating in the perinatal period) and Q (congenital malformations deformations and chromosomal abnormalities).

The second branch (middle branch in Fig. 6) covers the age period from twenty to sixty. It comprises the five communities c2_50.59_M, c1_30.39_M, c1_20.29_M, c2_20.29_F, and c4_40.49_F. These are communities with a large

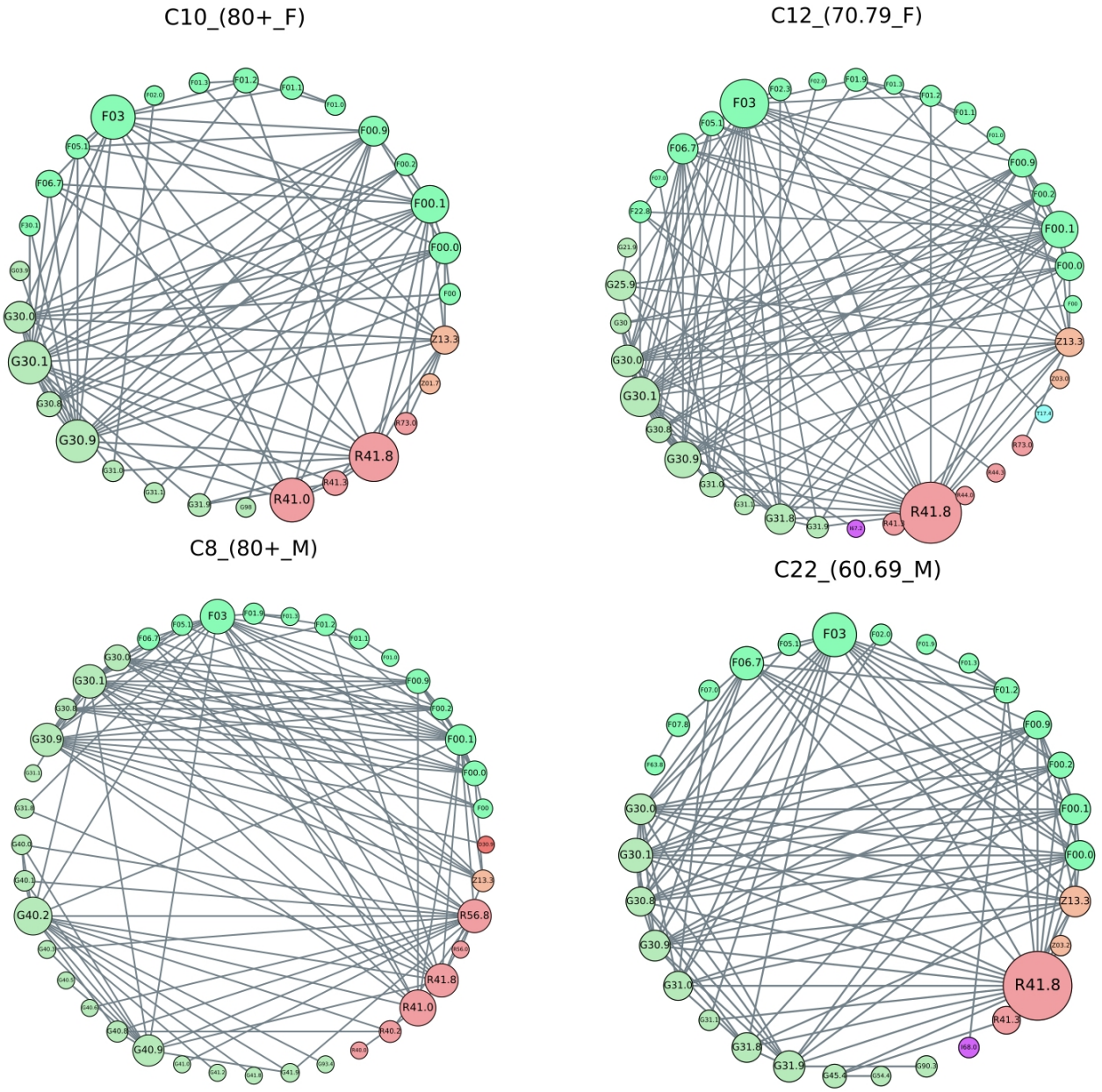

FIG. 7. Four communities c10_80.XX_F, c12_70.79_F, c8_80.XX_M, and c22_60.69_M of the community cluster shown in Fig. 5. Different node colors indicate different ICD categories. The size of each node is proportional to the degree of the node in the SVN.

number of nodes and links, including some nodes belonging to many different disease categories (see top panel of Fig. 9). However, in spite of the presence of nodes of several different categories, the over-expressed categories are F and Z, showing that the backbone of these communities consists of mental and behavioral disorders.

The third branch (right branch in Fig. 6) has primarily communities of older patients. The six communities in this branch are c7_80.XX_F, c6_70.79_M, c4_70.79_F, c3_60.69_F, c3_40.49_M, and c5_60.69_F. These communities contain nodes from multiple categories of ICD codes (see top panel of Fig. 9). Specifically, in addition to the over-expressed F category, we observe the G, R, T, and Z categories. The presence of the G category reflects the role of diseases of the nervous system whereas the S and T categories represent "injury, poisoning and certain other consequences of external causes". For this subcluster, the only over-expressed category is F. For these age cohorts, the differences between males and females decrease when age increases.

In summary, the comparative analysis of ICD code communities obtained from statistically validated comorbidity networks gives us interesting insights about how disease classes are distributed across different cohorts of patients of

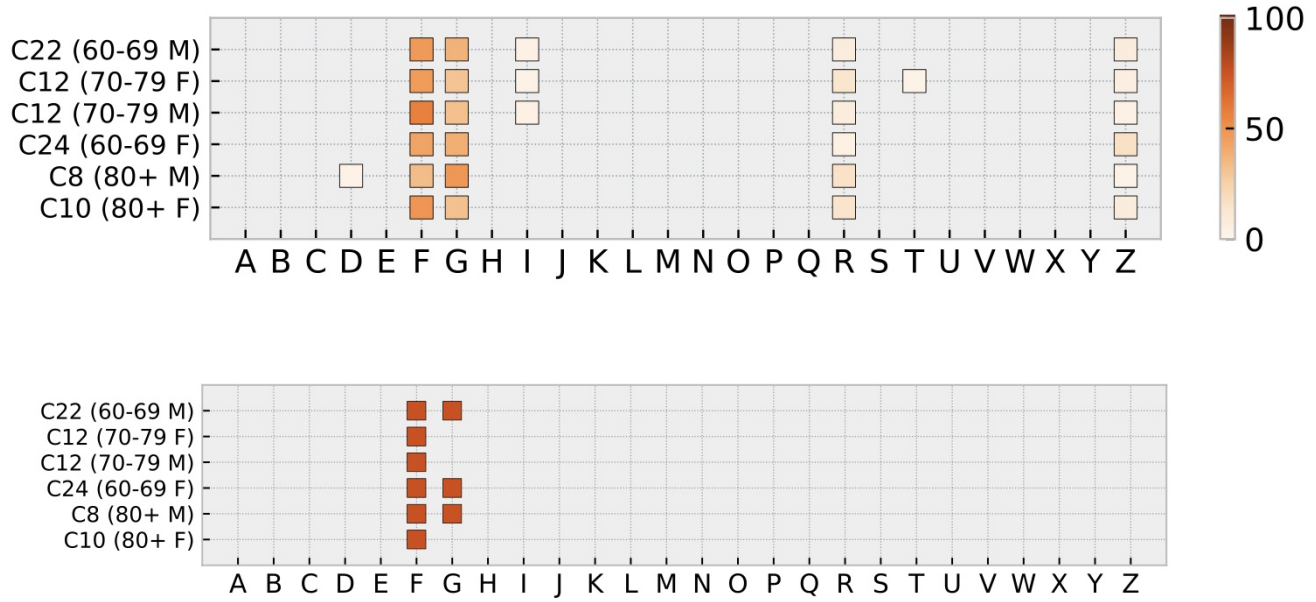

FIG. 8. Occurrence (upper panel) and over-expression (lower panel) of the different ICD letter categories in the communities in the cluster of Fig. 5. The color scale in the upper graph represents the percentage of nodes belonging to a particular disease class (letter) over all the classes. In the lower graph, the color is fixed since over-expressions are either present or not.

different ages and sexes. 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.

For example, when we consider communities involving diseases primarily related to categories C and D (i.e., Neoplasms) we note that the node of highest degree of each community is informative with respect to the branch of hierarchical tree the community belongs to. In fact, for each cluster observed in the hierarchical tree with over-expression of category C, communities of a given branch of the hierarchical tree are rich in information about comorbidity associated with a specific type of neoplasm. Below, we report the ICD node of the highest degree that is present in each branch with C over-expression. Specifically, at different positions of the hierarchical tree, we observe: for positions 4-11 (D48 - Neoplasm of uncertain or unknown behavior of other and unspecified sites), for 20-27 (C43/44 - Malignant melanoma of skin/Other malignant neoplasms of skin), for 31-41 (C83 - Non-follicular lymphoma), 42-53 (C90 - Multiple myeloma and malignant plasma cell neoplasms), for 61-63 (C56 - Malignant neoplasm of ovary), 131-136 (C50 - Malignant neoplasm of breast), for 137-143 (C79 - Secondary malignant neoplasm of other and unspecified sites), and for 216-228 (C71 - Malignant neoplasm of brain).

D. Dismantling the comorbidity network

A classical topic in network theory is the investigation of the level of network resilience to random or targeted attacks [38]. This is relevant under two different view points: on one side such levels of resilience are indicative of how much networks are robust against attacks. On the other side, however, looking for the best strategy to dismantle a network can give an indication of what are the nodes or group of nodes most relevant in guarentiing the network cohesivity. 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 central in the setting of comorbidities.

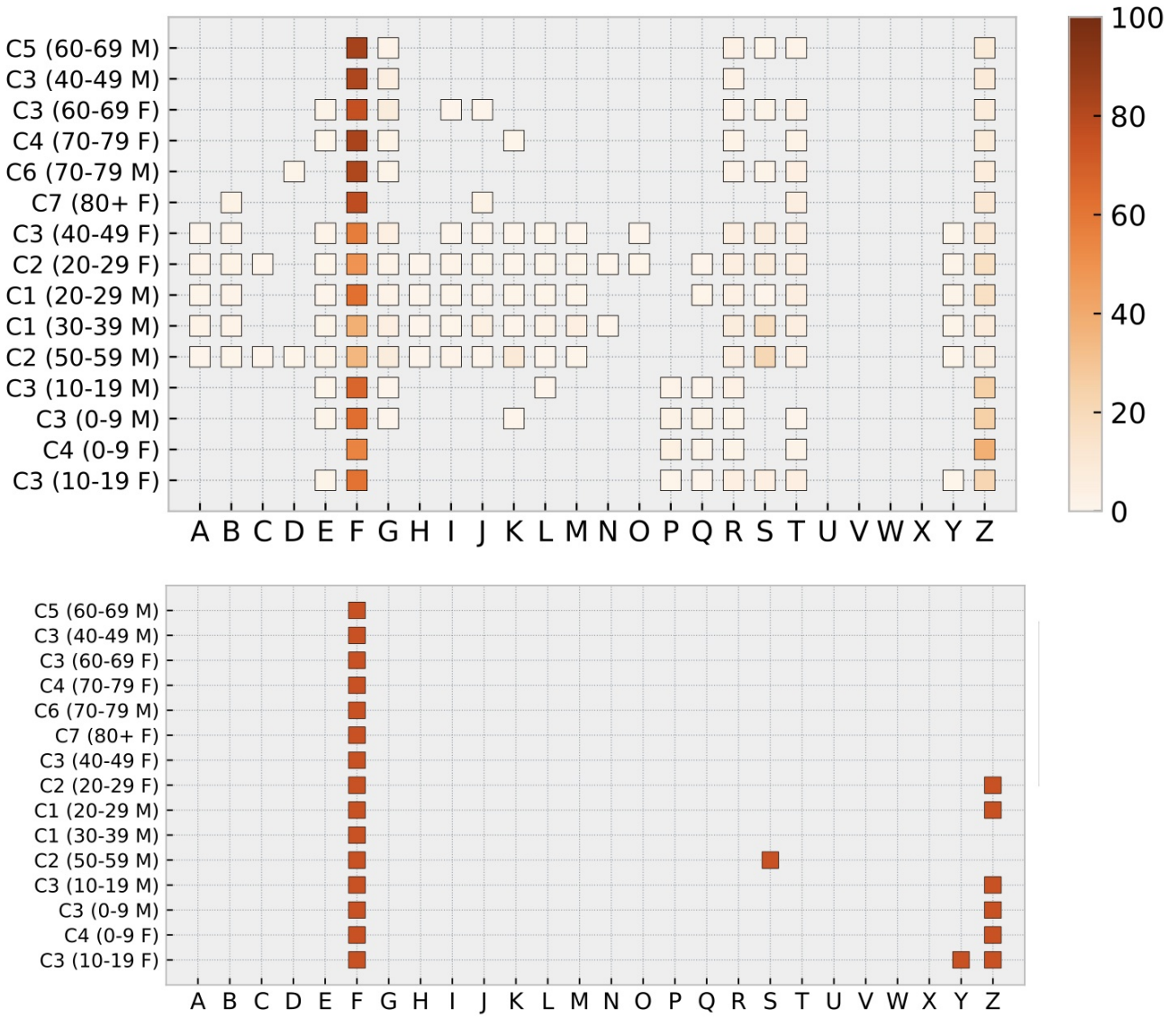

FIG. 9. Occurrence and over-expression of ICD letter category in the communities in the cluster of Fig. 6.

1. *The dismantling procedure*

The presence of a link in the comorbidity network is carrying two types of information. On one hand it is showing that the two diseases occurs in a patient with a probability that it is not compatible with a null hypothesis based on the prevalence of the two diseases. The same link is also signaling the potential for patients having only one of these diseases at a given time to get the second one. This means that developping comormidities over time for a specific patient can be seen as a walk performed on the comorbidity network. Imagine a disease path characterized by diseases A, B and C. Dismantling the comorbidity network (for example due to preventive medicine policies affecting disease B) therefore would diminish groups of patients affected by disease A to develop comorbidities with disease C (and viceversa) thanks to the hipothesized dismantling action.

Let us now consider a simulated "dismantling" procedure [34] of the PROJ and SVN comorbidity networks to highlight the categories of ICD codes which are most important for bridging different regions of the comorbidity network. For each age and sex cohort, we remove one node (i.e., one ICD code) at a time from the largest connected component of the comorbidity network, iterating the removal process until percolation across all remaining nodes of the comorbidity network breaks down. We monitor the presence of percolation among the remaining nodes of the

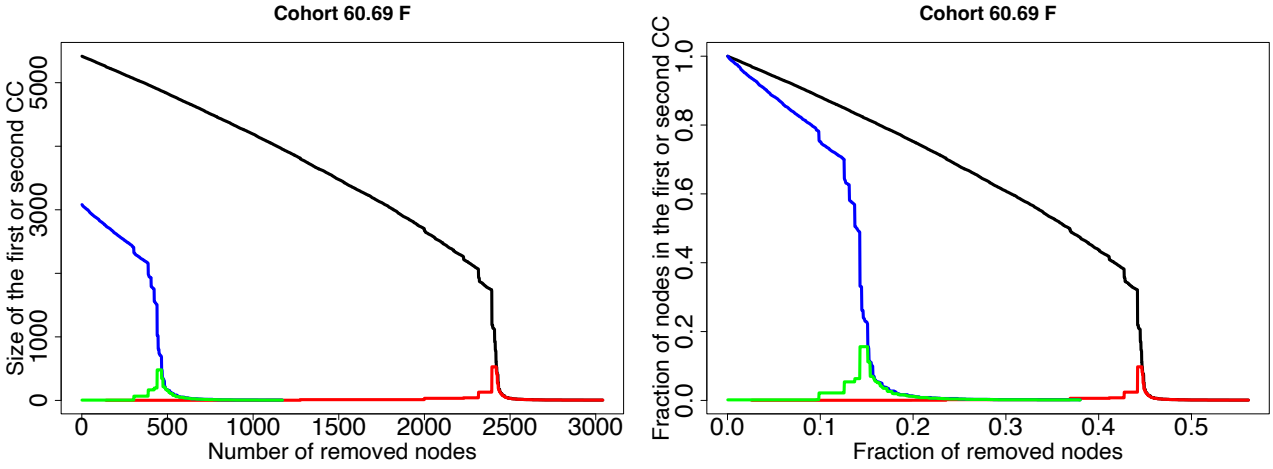

FIG. 10. Dismantling process of the comorbidity networks obtained for female 60-69 cohort. In the two panels, we show the size of the largest connected component (black line for PROJ network and blue line for SVN) and the size of the second largest connected component (red line for the PROJ network and 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.

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 [39].

The sequence of node removal of ICD codes is obtained using the following procedure: at each step, we compute the betweenness of all nodes, and we remove the node with the highest betweenness [34]. The dismantling process is applied to a slightly modified version of the comorbidity networks discussed in Sect. III A. Specifically, for each cohort the modification we make consists in removing all nodes with ICD codes belonging to categories R and Z. These ICD categories either represent codes for symptoms or are used by physicians to document the type of interactions patients had with the health services. These nodes connect diseases through common symptoms or hospital procedures/check-ups. Thus, removing these nodes from the network has 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.

2. An illustrative example

In Fig. 10, 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. 10 shows that the fraction of nodes that has to be removed from the

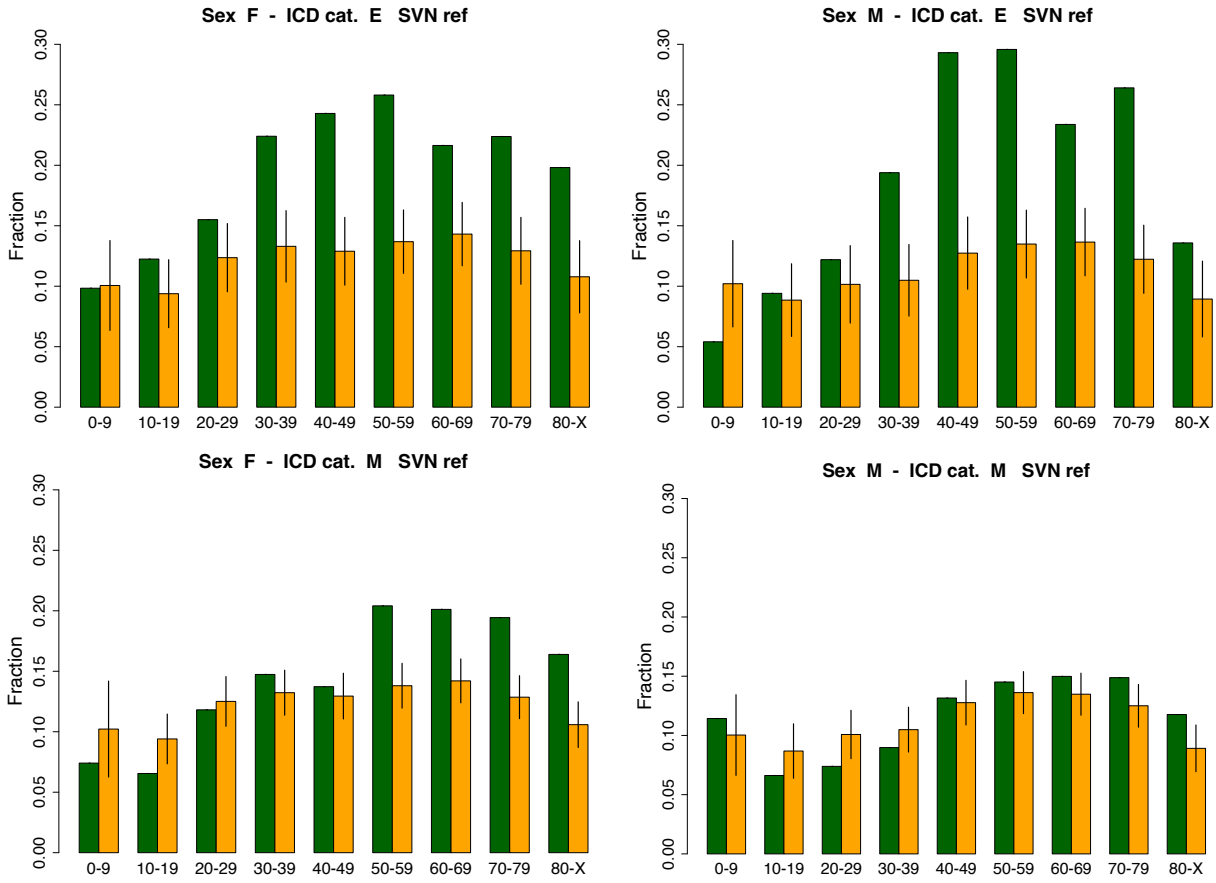

FIG. 11. Top panels: Fraction (dark green bars) of removed ICD codes reducing the size of the largest connected component of SVN to 10% of its original size for the category E: Endocrine, nutritional and metabolic diseases (top panels), and for the category M: Diseases of the musculoskeletal system and connective tissue (bottom panels), for all age cohorts of females (left panel) and males (right panel). Orange bars are the corresponding fraction of category E obtained by random selection of nodes of the SVN networks, obtained by selecting a number of nodes equal to the size of "dismantling" nodes detected for each cohort. Black segments on top of the orange bars display plus and minus one standard deviation observed across 1000 realizations of the sampling procedure. Bottom panels: the same quantities for category M (Diseases of the musculoskeletal system and connective tissue).

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

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 the 'dismantling set'. The number N_{dism} will be specific to each cohort. For each disease category, we calculate the fraction of nodes that are in the dismantling set and we compare this fraction to the fraction observed by randomly sampling a 1000 sets of ICD nodes of the same size of the dismantling set from the corresponding SVN network. This comparison allows us to detect the relative abundance of nodes of a specific category in the dismantling set for each cohort.

In Fig. 11 we show two examples of the frequencies of ICD codes 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

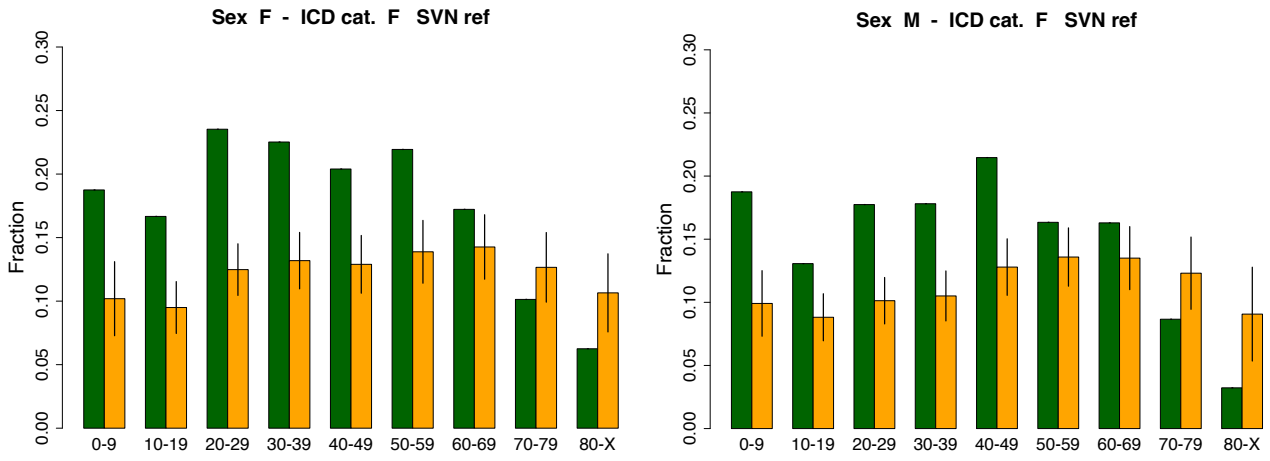

FIG. 12. Fraction (dark green bars) of removed ICD codes reducing the size of the largest connected component of SVN to 10% of its original size for the category F (Mental and behavioural disorders) for all age cohorts of females (left panel) and males (right panel). Orange bars are the corresponding fraction of category F obtained by randomly selecting in the SVNs a number of nodes equal to the size of "dismantling" nodes detected for each cohort. Black segments on top of the orange bars display plus and minus one standard deviation observed in 1000 random realizations.

connective tissue, bottom panels). Dark green bars in the figure represent the frequency observed in the dismantling set for the cohort in hand, and orange bars the average frequency of 1000 simulations of random selection from the corresponding SVN network. Vertical black lines on top of the orange bars indicates plus and minus one standard deviation across the 1000 samples.

The two examples in Fig. 11 show two typical patterns. Specifically, the case of category E (upper panels) shows a frequency pattern for different age cohorts that is quite similar for both sexes. The frequency of the E category of diseases is compatible, or almost compatible, with the one expected for a random inclusion in the dismantling set for age cohorts 0-9, 10-19, 20-29 and 80-XX. On the other hand, a higher frequency is observed for the age cohorts 30-39, 40-49, 50-59, 60-69 and 70-79. 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 gave them 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. 11. These are conditions related to the musculoskeletal system. For male patients, the frequency of category-M ICD codes in the dismantling set is consistent 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. Absence of enhancement of frequency of M diseaases is observed for younger age cohorts (until age 49) but an abrupt frequency enhancement is observed starting from 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 frequency 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 Sec. III C. In Fig. 12, we notice that F ICD nodes show a higher frequency in dismantling nodes of younger cohorts both for females and males compared to random sampling. The presence of these codes 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 Finvolve 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.

For each disease category, 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 suggestions for preventive medicine policies.

IV. 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 code. It is worth to recall that this level is the one primarily used in medical diagnoses. Therefore, this level of granularity is probably the one that better describe the information stored 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 projected disease networks (the networks that we called 'PROJ'). SVN allows unsupervised detection of groups of diseases (labeled by us as 'diseases communities'). ICD communities are obtained by using community detection algorithms introduced in complex network research. Disease communities obtained in this way are rich of medical information that 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 30s to the late 70s 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 musculosketal system and connective tissue where a distinct impact is seen for females of age from 50s to late 70s whereas no apparent role is detectable for males at any age.

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.

V. DECLARATIONS

A. Availability of data and 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 [35]). Informed consent was waived due to the study's retrospective design, according to Finnish legislation on the secondary use of health data.

B. Competing interests

The authors have no competing interests as defined by Springer, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

C. Funding

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.

D. Authors' 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.

E. Acknowledgments

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.

[1] M. Newman, Networks (Oxford university press, 2018).
[2] A.-L. Barabási and M. Pósfai, Network Science (Cambridge University Press, 2016).
[3] V. Latora, V. Nicosia, and G. Russo, Complex networks: principles, methods and applications (Cambridge University Press, 2017).
[4] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, Network medicine: a network-based approach to human disease, Nature reviews genetics **12**, 56 (2011).
[5] P. C. Ivanov, The new field of network physiology: building the human physiolome, Frontiers in network physiology **1**, 711778 (2021).
[6] International statistical classification of diseases and related health problems 10th revision, `https://icd.who.int/browse10/2019/en`.
[7] D. Lee, J. Park, K. Kay, N. Christakis, Z. Oltvai, and A. Barabási, The implications of human metabolic network topology for disease comorbidity, Proceedings of the National Academy of Sciences **105**, 9880 (2008).
[8] 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 **5**, e1000353 (2009).
[9] F. Folino, C. Pizzuti, and M. Ventura, A comorbidity network approach to predict disease risk, in International Conference on Information Technology in Bio-and Medical Informatics (Springer, 2010) pp. 102–109.
[10] F. S. Roque, P. B. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, K. Søeby, S. Bredkjær, A. Juul, T. Werge, et al., Using electronic patient records to discover disease correlations and stratify patient cohorts, PLoS computational biology **7**, e1002141 (2011).
[11] A. Chmiel, P. Klimek, and S. Thurner, Spreading of diseases through comorbidity networks across life and gender, New Journal of Physics **16**, 115013 (2014).

[12] A. B. Jensen, P. L. Moseley, T. I. Oprea, S. G. Ellesøe, R. Eriksson, H. Schmock, P. B. Jensen, L. J. Jensen, and S. Brunak, Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients, Nature communications **5**, 4022 (2014).

[13] E. Jeong, K. Ko, S. Oh, and H. W. Han, Network-based analysis of diagnosis progression patterns using claims data, Scientific reports **7**, 15561 (2017).

[14] Y. Bao, P. Lu, M. Wang, X. Zhang, A. Song, X. Gu, T. Ma, S. Su, L. Wang, X. Shang, et al., Exploring multimorbidity profiles in middle-aged inpatients: a network-based comparative study of china and the united kingdom, BMC medicine **21**, 495 (2023).

[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/`.

[17] B. Fotouhi, N. Momeni, M. A. Riolo, and D. L. Buckeridge, Statistical methods for constructing disease comorbidity networks from longitudinal inpatient data, Applied network science **3**, 1 (2018).

[18] M. Tumminello, S. Miccichè, F. Lillo, J. Piilo, and R. N. Mantegna, Statistically validated networks in bipartite complex systems, PLoS ONE **6**, e17994 (2011).

[19] M.-X. Li, V. Palchykov, Z.-Q. Jiang, K. Kaski, J. Kertész, S. Micciche, M. Tumminello, W.-X. Zhou, and R. N. Mantegna, Statistically validated mobile communication networks: the evolution of motifs in european and chinese data, New Journal of Physics **16**, 083038 (2014).

[20] V. Hatzopoulos, G. Iori, R. N. Mantegna, S. Micciche, and M. Tumminello, Quantifying preferential trading in the e-mid interbank market, Quantitative Finance **15**, 693 (2015).

[21] M. Á. Serrano, M. Boguná, and A. Vespignani, Extracting the multiscale backbone of complex weighted networks, Proceedings of the national academy of sciences **106**, 6483 (2009).

[22] F. Radicchi, J. J. Ramasco, and S. Fortunato, Information filtering in complex weighted networks, Physical Review E—Statistical, Nonlinear, and Soft Matter Physics **83**, 046101 (2011).

[23] F. Saracco, R. Di Clemente, A. Gabrielli, and T. Squartini, Randomizing bipartite networks: the case of the world trade web, Scientific reports **5**, 10595 (2015).

[24] F. Saracco, M. J. Straka, R. Di Clemente, A. Gabrielli, G. Caldarelli, and T. Squartini, Inferring monopartite projections of bipartite networks: an entropy-based approach, New Journal of Physics **19**, 053022 (2017).

[25] R. Marcaccioli and G. Livan, A pólya urn approach to information filtering in complex networks, Nature communications **10**, 745 (2019).

[26] T. Kobayashi, T. Takaguchi, and A. Barrat, The structured backbone of temporal social ties, Nature communications **10**, 220 (2019).

[27] N. Vallarano, M. Bruno, E. Marchese, G. Trapani, F. Saracco, G. Cimini, M. Zanon, and T. Squartini, Fast and scalable likelihood maximization for exponential random graph models with local constraints, Scientific Reports **11**, 15227 (2021).

[28] G. M. J. Rupert, Simultaneous statistical inference (Springer Science & Business Media, 2012).

[29] Y. Benjamini and Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing, Journal of the Royal Statistical Society Series B: Statistical Methodology **57**, 289–300 (1995).

[30] S. Fortunato and D. Hric, Community detection in networks: A user guide, Physics reports **659**, 1 (2016).

[31] M. Rosvall and C. T. Bergstrom, Maps of random walks on complex networks reveal community structure, Proceedings of the national academy of sciences **105**, 1118 (2008).

[32] L. Kaufman and P. J. Rousseeuw, Finding groups in data: an introduction to cluster analysis (John Wiley & Sons, 2009).

[33] A. Braunstein, L. Dall'Asta, G. Semerjian, and Z. L, Network dismantling, Proceedings of the National Academy of Sciences **113**, 12368 (2016).

[34] S. Wandelt, X. Sun, D. Feng, M. Zanin, and S. Havlin, A comparative analysis of approaches to network-dismantling, Scientific reports **8**, 13513 (2018).

[35] J. Kivelev, I. Saarenpää, A. Karlsson, P. Crisafulli, F. Musciotto, J. Piilo, and R. N. Mantegna, Complex networks approach to study comorbidities in patients with unruptured intracranial aneurysms, Scientific Reports **14**, 10.1038/s41598-024-59919-2 (2024).

[36] Svalnet github repository, `https://github.com/complexParide/svalnet`.

[37] M. Tumminello, S. Micciche, F. Lillo, J. Varho, J. Piilo, and R. N. Mantegna, Community characterization of heterogeneous complex systems, Journal of Statistical Mechanics: Theory and Experiment **2011**, P01019 (2011).

[38] R. Albert, H. Jeong, and A.-L. Barabási, Error and attack tolerance of complex networks, Nature **406**, 378–382 (2000).

[39] F. Musciotto and S. Miccichè, Exploring the landscape of dismantling strategies based on the community structure of networks, Scientific Reports **13**, 14448 (2023).

Appendix A: Over-expression of disease categories in diseases' communities

This appendix shows the full list of over-expression of disease categories in diseases' communities with more than 25 ICD codes. The 380 communities are split into three figures (Fig 13, 14, and 15). Each figure contains three main columns, each listing a series of communities. Each column contains: (i) an integer number (the order of the disease community in the Average Linkage hierarchical tree, then, (ii) the label the community, and (iii) the over-expressed categories of ICD codes in the community. In the absence of over-expression Na is reported.

Appendix B: Jaccard similarity matrix for SVN communities

This appendix contains the Jaccard similarity matrix among all pairs of 380 SVN communities (Fig. 16).

ALHC order	Disease Community	OE category	ALHC order	Disease Community	OE category	ALHC order	Disease Community	OE category
1	c12_30.39_F	I	51	c20_70.79_F	D	101	c1_10.19_F	JHR
2	c23_40.49_F	I	52	c17_20.29_M	C	102	c2_0.9_M	JLHB
3	c26_40.49_M	I	53	c18_30.39_M	C	103	c5_0.9_F	JL
4	c26_60.69_F	D	54	c6_40.49_F	N	104	c13_70.79_M	J
5	c24_70.79_F	Na	55	c5_50.59_F	N	105	c8_30.39_F	J
6	c21_80.XX_F	Na	56	c9_60.69_F	N	106	c7_20.29_F	J
7	c28_20.29_M	C	57	c13_70.79_F	N	107	c10_60.69_M	J
8	c25_20.29_F	D	58	c12_80.XX_F	N	108	c5_50.59_M	J
9	c33_50.59_M	D	59	c19_80.XX_F	C	109	c6_30.39_M	J
10	c32_40.49_F	D	60	c16_80.XX_M	C	110	c7_40.49_M	J
11	c26_50.59_F	D	61	c35_50.59_F	Na	111	c17_40.49_M	K
12	c13_0.9_F	T	62	c14_60.69_F	C	112	c19_50.59_F	K
13	c10_0.9_M	T	63	c15_70.79_F	C	113	c26_50.59_M	K
14	c30_40.49_F	T	64	c5_0.9_M	Q	114	c4_20.29_M	AKJ
15	c25_20.29_M	T	65	c7_10.19_M	H	115	c6_50.59_M	K
16	c30_30.39_M	T	66	c13_80.XX_F	H	116	c6_60.69_M	K
17	c23_40.49_M	T	67	c10_80.XX_M	H	117	c8_0.9_M	D
18	c18_50.59_M	T	68	c12_20.29_F	H	118	c9_40.49_F	K
19	c23_60.69_M	T	69	c11_30.39_F	H	119	c11_20.29_F	K
20	c19_60.69_M	D	70	c10_40.49_F	H	120	c6_30.39_F	K
21	c8_70.79_M	DC	71	c10_50.59_F	H	121	c20_60.69_M	K
22	c6_80.XX_M	DC	72	c8_60.69_F	H	122	c23_60.69_F	K
23	c16_60.69_F	DC	73	c7_70.79_F	H	123	c17_50.59_M	K
24	c11_70.79_F	DC	74	c11_50.59_M	H	124	c12_20.29_M	K
25	c5_80.XX_F	DC	75	c11_60.69_M	H	125	c12_30.39_M	K
26	c21_50.59_F	DC	76	c5_70.79_M	H	126	c14_40.49_M	K
27	c22_50.59_M	DC	77	c7_20.29_M	H	127	c18_10.19_F	K
28	c22_80.XX_F	N	78	c11_30.39_M	H	128	c10_10.19_M	KD
29	c14_20.29_F	N	79	c12_40.49_M	H	129	c21_20.29_F	Z
30	c22_30.39_F	Na	80	c13_60.69_M	J	130	c46_50.59_M	Z
31	c24_10.19_F	C	81	c4_80.XX_M	J	131	c7_50.59_F	C
32	c16_20.29_F	C	82	c21_60.69_F	Na	132	c7_60.69_F	C
33	c26_40.49_F	C	83	c17_80.XX_F	J	133	c8_70.79_F	CZ
34	c19_40.49_M	C	84	c5_40.49_F	JL	134	c14_80.XX_F	Na
35	c20_80.XX_F	C	85	c9_20.29_M	L	135	c15_30.39_F	C
36	c23_50.59_F	C	86	c6_70.79_F	L	136	c13_40.49_F	C
37	c22_60.69_F	C	87	c9_50.59_F	L	137	c14_70.79_F	CD
38	c23_70.79_F	C	88	c6_60.69_F	L	138	c9_60.69_M	C
39	c17_70.79_M	C	89	c10_70.79_M	L	139	c7_70.79_M	C
40	c21_50.59_M	C	90	c9_50.59_M	L	140	c22_40.49_M	C
41	c18_60.69_M	C	91	c14_60.69_M	L	141	c12_50.59_M	C
42	c8_0.9_F	CA	92	c9_30.39_M	L	142	c11_80.XX_F	C
43	c15_10.19_F	C	93	c11_40.49_M	L	143	c12_80.XX_M	C
44	c10_20.29_F	Na	94	c8_20.29_F	L	144	c16_40.49_F	K
45	c16_40.49_M	Na	95	c3_30.39_F	L	145	c28_40.49_F	K
46	c27_40.49_F	C	96	c11_10.19_F	L	146	c17_60.69_F	K
47	c16_50.59_F	C	97	c8_80.XX_F	L	147	c19_70.79_F	K
48	c13_50.59_M	C	98	c11_80.XX_M	L	148	c12_60.69_M	K
49	c16_60.69_M	D	99	c3_0.9_F	JH	149	c14_70.79_M	K
50	c12_60.69_F	B	100	c1_10.19_M	PJ	150	c18_40.49_F	K

FIG. 13. Over-expression of disease categories in diseases communities with more than 25 ICD codes. For each group of three columns, the first is the order of the disease community in the Average Linkage hierarchical tree, the second is the label of the disease community and the third is the over-expressed category (or categories in case of multiple letters). In. the absence of over-expression Na is reported. Average linkage hierarchical clustering (ALHC) order from 1 to 150.

ALHC order	Disease Community	OE category	ALHC order	Disease Community	OE category	ALHC order	Disease Community	OE category
151	c9_80.XX_F	K	201	c24_60.69_F	FG	251	c17_60.69_M	M
152	c9_80.XX_M	K	202	c12_70.79_M	F	252	c24_40.49_M	M
153	c7_60.69_M	K	203	c12_70.79_F	F	253	c15_50.59_M	M
154	c9_70.79_M	K	204	c22_60.69_M	FG	254	c11_60.69_F	M
155	c10_60.69_F	K	205	c25_30.39_M	I	255	c14_40.49_F	M
156	c10_70.79_F	K	206	c13_40.49_M	Na	256	c12_50.59_F	M
157	c13_30.39_M	KE	207	c26_20.29_F	I	257	c1_60.69_M	IEN
158	c4_40.49_M	K	208	c18_30.39_F	I	258	c1_70.79_M	IRMEJN
159	c30_60.69_F	E	209	c18_40.49_M	I	259	c2_60.69_F	IEN
160	c33_70.79_F	E	210	c20_50.59_M	IG	260	c1_50.59_M	IEN
161	c28_50.59_F	E	211	c21_60.69_M	I	261	c2_30.39_M	EN
162	c22_30.39_M	E	212	c15_60.69_F	I	262	c1_40.49_M	EHN
163	c12_10.19_F	G	213	c21_70.79_F	I	263	c4_40.49_F	EH
164	c24_30.39_M	G	214	c12_40.49_F	I	264	c2_50.59_F	HEI
165	c10_30.39_F	H	215	c8_50.59_F	GI	265	c4_80.XX_F	E
166	c19_40.49_F	H	216	c23_20.29_F	Na	266	c3_80.XX_M	E
167	c16_10.19_F	E	217	c9_10.19_M	C	267	c10_10.19_F	E
168	c11_10.19_M	E	218	c16_20.29_M	C	268	c13_10.19_M	E
169	c28_70.79_M	EH	219	c16_80.XX_F	G	269	c6_20.29_M	E
170	c19_30.39_F	E	220	c14_10.19_F	EC	270	c3_20.29_F	EH
171	c24_40.49_F	E	221	c15_20.29_F	C	271	c2_30.39_F	EH
172	c22_50.59_F	E	222	c25_60.69_F	D	272	c27_40.49_M	H
173	c18_60.69_F	E	223	c24_60.69_M	D	273	c25_40.49_F	H
174	c22_70.79_F	E	224	c16_70.79_F	G	274	c29_50.59_F	H
175	c7_80.XX_M	K	225	c14_30.39_M	GC	275	c22_20.29_F	H
176	c3_70.79_M	KC	226	c15_40.49_M	G	276	c30_30.39_F	HB
177	c4_50.59_M	KC	227	c14_30.39_F	G	277	c6_0.9_F	M
178	c4_60.69_M	KC	228	c22_40.49_F	G	278	c8_10.19_F	MH
179	c3_80.XX_F	K	229	c13_10.19_F	QI	279	c6_0.9_M	MH
180	c5_60.69_F	K	230	c13_20.29_M	R	280	c6_10.19_M	M
181	c3_70.79_F	K	231	c15_30.39_M	I	281	c14_20.29_M	HM
182	c5_20.29_F	K	232	c10_40.49_M	IR	282	c16_30.39_M	H
183	c7_30.39_F	K	233	c17_30.39_F	I	283	c13_60.69_F	H
184	c3_30.39_M	K	234	c15_40.49_F	I	284	c8_50.59_M	H
185	c6_40.49_M	K	235	c13_50.59_F	I	285	c2_70.79_F	H
186	c7_40.49_F	K	236	c20_20.29_M	I	286	c2_80.XX_F	H
187	c6_50.59_F	K	237	c20_40.49_M	I	287	c2_80.XX_M	H
188	c4_0.9_M	K	238	c18_70.79_F	M	288	c8_60.69_M	H
189	c7_10.19_F	K	239	c19_60.69_F	M	289	c4_70.79_M	H
190	c4_10.19_M	K	240	c17_40.49_F	M	290	c24_50.59_F	M
191	c2_0.9_F	P	241	c14_50.59_F	M	291	c3_10.19_F	YFZ
192	c5_10.19_F	P	242	c17_10.19_F	N	292	c4_0.9_F	FZ
193	c1_0.9_M	PG	243	c5_70.79_F	N	293	c3_0.9_M	FZ
194	c4_20.29_F	GF	244	c5_80.XX_M	N	294	c3_10.19_M	FZ
195	c1_0.9_F	G	245	c5_30.39_F	N	295	c2_50.59_M	FS
196	c2_10.19_F	GF	246	c18_20.29_M	N	296	c1_30.39_M	F
197	c2_10.19_M	G	247	c15_70.79_M	TM	297	c1_20.29_M	FZ
198	c2_20.29_M	KG	248	c16_30.39_F	M	298	c2_20.29_F	FZ
199	c10_80.XX_F	F	249	c27_20.29_F	M	299	c3_40.49_F	F
200	c8_80.XX_M	GF	250	c21_30.39_M	M	300	c7_80.XX_F	F

FIG. 14. Over-expression of disease categories in disease communities with more than 25 ICD codes. For each group of three columns, the first is the order of the disease community in the Average Linkage hierarchical tree, the second is the label of the disease community and the third is the over-expressed category (or categories in case of multiple letters). In. the absence of over-expression Na is reported. ALHC order from 151 to 300.

ALHC order	Disease Community	OE category	ALHC order	Disease Community	OE category
301	c6_70.79_M	F	351	c2_40.49_M	SFY
302	c4_70.79_F	F	352	c2_70.79_M	YSG
303	c3_60.69_F	F	353	c4_60.69_F	S
304	c3_40.49_M	F	354	c2_60.69_M	YS
305	c5_60.69_M	F	355	c17_70.79_F	FS
306	c4_10.19_F	O	356	c14_80.XX_M	S
307	c1_20.29_F	ON	357	c27_60.69_M	N
308	c1_30.39_F	POFZ	358	c9_20.29_F	S
309	c1_40.49_F	OZ	359	c13_30.39_F	S
310	c3_50.59_F	OZ	360	c10_20.29_M	S
311	c9_10.19_F	M	361	c17_80.XX_M	S
312	c4_30.39_F	M	362	c22_10.19_F	S
313	c8_40.49_F	M	363	c20_50.59_F	S
314	c6_20.29_F	M	364	c12_10.19_M	S
315	c8_40.49_M	M	365	c8_30.39_M	S
316	c8_20.29_M	M	366	c23_70.79_M	S
317	c5_30.39_M	M	367	c30_40.49_M	S
318	c1_80.XX_M	IR	368	c23_50.59_M	S
319	c1_70.79_F	MIR	369	c20_30.39_F	Na
320	c1_80.XX_F	IMR	370	c7_0.9_F	N
321	c2_40.49_F	RM	371	c7_0.9_M	N
322	c1_50.59_F	FJMR	372	c15_10.19_M	N
323	c1_60.69_F	MRJ	373	c11_20.29_M	N
324	c19_20.29_F	M	374	c7_30.39_M	N
325	c10_30.39_M	M	375	c9_40.49_M	N
326	c5_40.49_M	MG	376	c7_50.59_M	N
327	c3_50.59_M	M	377	c15_60.69_M	GN
328	c3_60.69_M	M	378	c16_70.79_M	G
329	c16_10.19_M	F	379	c30_50.59_F	G
330	c27_70.79_M	S	380	c27_60.69_F	G
331	c24_50.59_M	S			
332	c26_60.69_M	S			
333	c8_10.19_M	S			
334	c3_20.29_M	S			
335	c6_10.19_F	S			
336	c9_0.9_F	S			
337	c9_0.9_M	S			
338	c13_80.XX_M	S			
339	c5_10.19_M	SM			
340	c5_20.29_M	S			
341	c4_30.39_M	S			
342	c11_70.79_M	S			
343	c9_70.79_F	S			
344	c6_80.XX_F	S			
345	c10_50.59_M	S			
346	c13_20.29_F	S			
347	c11_50.59_F	S			
348	c9_30.39_F	S			
349	c11_40.49_F	S			
350	c4_50.59_F	SK			

FIG. 15. Over-expression of disease categories in diseases communities with more than 25 ICD codes. For each group of three columns, the first is the order of the disease community in the Average Linkage hierarchical tree, the second is the label of the disease community and the third is the over-expressed category (or categories in case of multiple letters). In. the absence of over-expression Na is reported. ALHC order from 301 to 380.

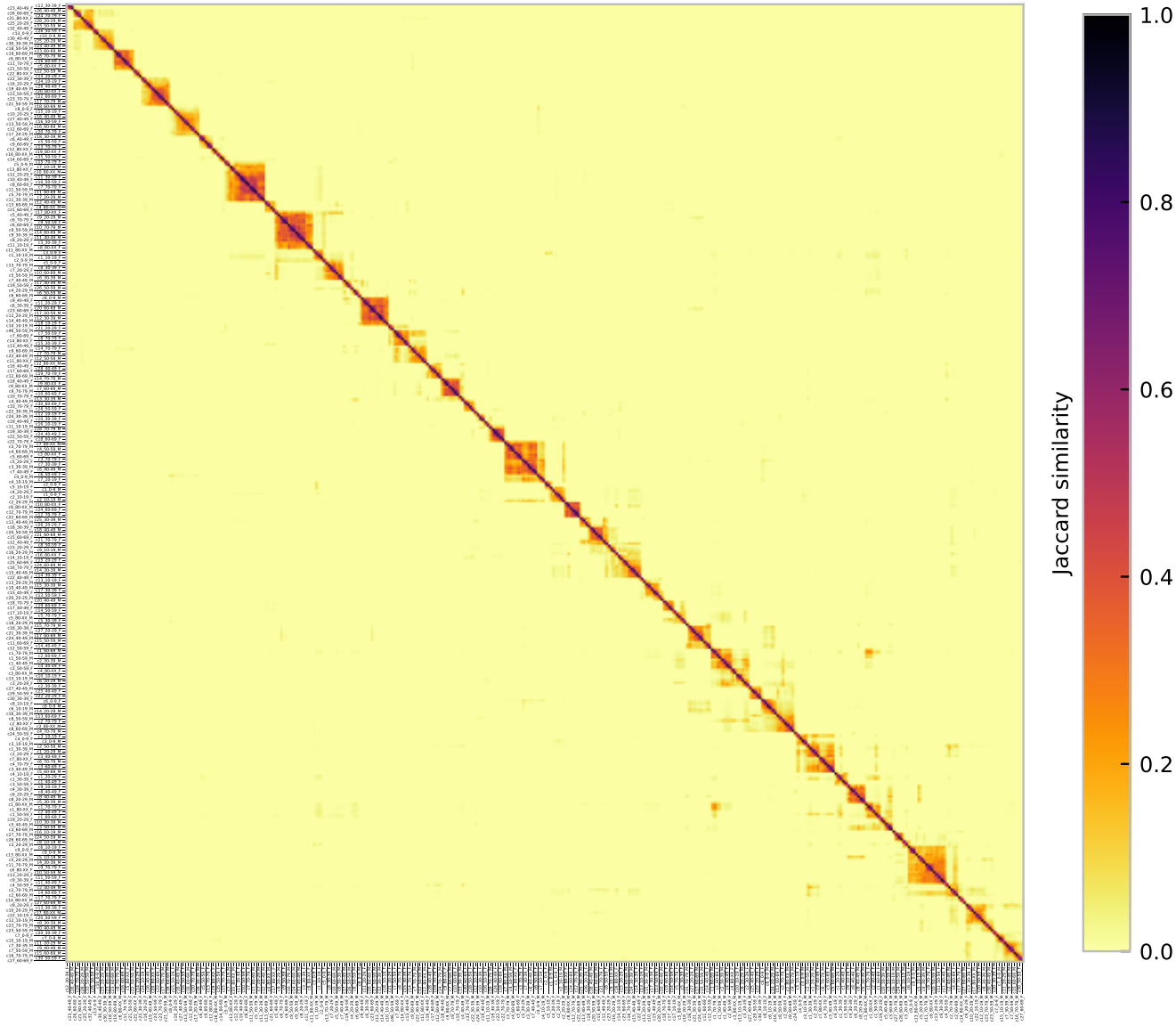

FIG. 16. Jaccard similarity matrix among all 380 × 380 pairs SVN communities. The order of communities in the rows and columns of the matrix is as in the hierarchical tree in Fig. 3.