Data mining and analysis of comorbidity networks from practitioner prescriptions

Il trattamento e l'analisi delle reti di comorbilitá delle prescrizioni dei medici generici

Giancarlo Ragozini, Giuseppe Giordano, Sergio Pagano, Mario De Santis, Pierpaolo Cavallo

Abstract For the present paper, the administrative databases of general practitioners were mined for healthcare systems analysis. A sample of 14,958 patients along with their 1,728,736 prescriptions were considered over a time span of eleven years. From this database, we derived a set of comorbidity networks by considering pathologies as nodes and the co-occurrences of two pathologies to the same prescription as links. The aim is to mine the complexity of this information by using network analysis techniques. Specifically, Islands algorithm method is well-suited to extract the most relevant and connected parts of large and dense networks, as in the present case. Main comorbidity patterns are discussed, along with future lines of research.

Abstract *In questo lavoro si mostra come un database amministrativo di un insieme di medici di base può essere considerato per un'analisi del sistema sanitario. Il database contiene un campione di 14,958 pazienti con le loro 1,728,736 prescrizioni mediche registrate in un arco temporale di 11 anni. Utilizzando i codici delle patologie contenute nelle prescrizioni, è possibile costruire la rete di comorbidità. Tale rete, densa di legami e con un numero elevato di nod, richiede metodologie di analisi appropriate. L'uso dell'algoritmo delle Islands è proposto per estrarre le principali strutture di associazione fra coppie di patologie.*

Key words: Community detection, Islands, Large network, Comorbidity pattern

---

Giancarlo Ragozini
Department of Political Science, University of Naples Federico II e-mail: giragoz@unina.it

Giuseppe Giordano
Department of Economics and Statistics, University of Salerno e-mail: ggiordan@unisa.it

Sergio Pagano
Department of Physics E.R. Caianiello, University of Salerno e-mail: spagano@unisa.it

Mario De Santis
Cooperativa Medi service Salerno e-mail: mariodesantis@osservatoriosanitario.it

Pierpaolo Cavallo
Department of Physics E.R. Caianiello, University of Salerno e-mail: pcavallo@unisa.it

1 Introduction

Nowadays, the presence of patients affected by many different diseases at the same time is becoming a major health and societal issue. In the United States, for instance, 80% of the health budget is spent on patients with four or more diseases (8). In clinical literature, this phenomenon is known as comorbidity.

The idea of comorbidity has been around since 1921 (9) used in a positive or negative sense, with the terms ”syntropy” and ”dystropy” (10). The former is the mutual disposition, or the appearance of two or more diseases in the same individual, while the latter indicates those pathologies that are rarely found in the same patient at the same time. In addition, the literature differentiates two core concepts, comorbidity and multimorbidity: the latter is defined as the coexistence of two or more long-term conditions in an individual not biologically or functionally linked (8), while the former concept refers to a coexistence of conditions that are linked, either biologically or functionally (12).

Simply considering this taxonomy, the intrinsically complex nature of comorbidity can be easily understood. On this basis, the burden of morbidity and comorbidity is considered to be influenced by a number of factors –namely, health-related, socioeconomic, cultural, environmental, and behavioral characteristics– but there is a lack of agreement (4) on how to understand the complex interdependent relationships between diseases due to: 1) a large number of variables (many of which are latent), 2) a lack of accuracy in measurements, and 3) technological limitations in generating data.

In this paper, we propose an indirect approach for a large-scale study of comorbidity patterns based on the administrative databases of prescription data from general practitioners (GPs), without the necessity of a complex clinical study. This methodology could be easily replicated. Access to prescription data from GPs is relatively simple, as these data are used for administrative purposes by the national health system. Given this kind of data, a morbidity state is associated with a patient and such a state is considered both over time for the same subject and within different categorized subjects. A recent literature review (13) has shown that comorbidities can be studied from GP databases based on both diagnoses, using the International Classification of Diseases –ICD– codes, and medications, using pharmacy data.

The paper is organized as follows. Section 2 presents the data and how they are used to define the comorbidity networks, along with an overview of the methodology we used. In Section 3 we report the first results, while the final section is devoted to a brief discussion with the future lines of research.

2 Data and methodology

2.1 Data

The Electronic Health Recordings (EHR) of the prescriptions made by a group of ten GPs belonging to the Cooperative Medi Service and operating in a town in Southern Italy were analyzed. A total number of 14,958 patients, covering a time interval of eleven years from 2002 to 2013, was considered. The total number of analyzed prescriptions was 1,728,736. The data were provided in anonymous form, for both patients and GPs, in accordance with Italian law on privacy and the guidelines of the Declaration of Helsinki. This retrospective observational study involved data mining and analysis of comorbidity networks from practitioner prescriptions, with data analyzed in aggregate form, and the relevant Ethics Committee granted approval (Comitato Etico Campania Sud, document number 59, released on 2016-06-08).

After a patients visit to a GP, there is generally a prescription containing a series of items of various types: drugs, laboratory tests, imaging tests, etc. For each patient visit, the GP administrative prescription data provided the following information:

- patient ID: a unique random number assigned to the patient;
- demographic data: age and sex;
- prescription date;
- prescription type: drug, laboratory test, imaging, specialist referral, hospitalization;
- prescription code: a specific code for each prescription type;
- associated ICD diagnostic code: the pathology connected to the specific prescription.

In accordance with the Italian National Health System rules, each item present in a GP prescription has an associated possible disease, encoded using the International Classification of Diseases, Ninth Revision, Clinical Modification –ICD-9-CM.[1] The ICD is the standard diagnostic tool for epidemiology, health management, and clinical purposes and is maintained by the World Health Organization. This includes the analysis of the general health situation of population groups. It is used to monitor the incidence and prevalence of diseases and other health problems, proving a picture of the general health situation of countries and populations. Each code has the general form xxx.yy, where xxx is the general disease and yy is a specific occurrence. For example, 250 is the code for ”Diabetes”, and 250.91 is the code for ”Diabetes Type 1 (juvenile) with unspecified complications, not stated as uncontrolled”.

[1] For details see http://www.salute.gov.it/portale/home.html.

2.2 Network data definition

To analyze the comorbidity patterns, we defined a set of networks. In the first step, a two-mode network was derived by considering the ICD9CM diagnostic codes and the prescriptions as the two disjoint sets of nodes. They are linked if corresponding codes appear in prescriptions made to the same patient on the same day. The sex and age of patients, and type and time of prescriptions can be considered as attributes of a given prescription.

Formally, the two-mode network can be represented as a bipartite graph $\mathscr{B}$ consisting of the two sets of relationally connected nodes and can be represented by a triple $\mathscr{B}(\mathscr{V}_1, \mathscr{V}_2, \mathscr{L})$, with $\mathscr{V}_1$ denoting the set of ICD-9-CM codes, $\mathscr{V}_2$ the set of prescriptions, and $\mathscr{L} \subseteq \mathscr{V}_1 \times \mathscr{V}_2$ the set of ties.

Being interested in the association of pathologies, we derived the one-mode network of the ICD-9-CM codes by projecting the two-mode network. The corresponding graph will be represented by $\mathscr{G}(\mathscr{V}_1, \mathscr{E}, \mathscr{W})$, with $\mathscr{V}_1$ the set of ICD-9-CM codes, $\mathscr{E} \subseteq \mathscr{V}_1 \times \mathscr{V}_1$ the set of edges, and $\mathscr{W}$ the set of weights, $w : \mathscr{E} \to \mathscr{N}$, $w(v_{1i}, v_{1j}) =$ the number of times that two ICD-9-CM codes appear in the same prescription.

In addition, different networks were generated by selecting a subset of the original dataset according to the patients sex, age intervals (0-15, 15-30, 30-45, 45-60, 60-75, 75-105), and type of prescription.

2.3 The methodology

The one-mode comorbidity network $\mathscr{G}$ is a large and dense network, and it is sometimes difficult to extract the main relevant comorbidity patterns. In order to identify the most relevant and connected parts of the network, we used the Islands approach (3), which is an algorithm useful for finding important parts in large networks with respect to a given property of nodes or lines (edges). By representing a given or computed value of nodes/lines as a height of nodes/lines and by immersing the network into water up to selected level, the islands are derived (2). Given a threshold value t, it is possible to obtain a cut of the network $\mathscr{G}(t)$. By varying this level t, different islands are identified. For our purposes, we decided to use the line island approach, which looks for a connected subnetwork with several nodes in a specified interval, such that the edges inside the island have a higher weight than the edges connecting nodes in the island with their neighbors. Formally, a set of nodes $\mathscr{I} \subseteq \mathscr{V}_1$ is a regular line island in the network $\mathscr{G}$ if:

$$\max_{(v_{1i},v_{1j})\in\mathscr{E}, v_{1i}\notin\mathscr{I}, v_{1j}\in\mathscr{I}} w(v_{1i}, v_{1j}) \leq \min_{(v_{1i},v_{1j})\in\mathscr{T}} w(v_{1i}, v_{1j}),$$

where $\mathscr{T}$ is the spanning tree over $\mathscr{I}$.

In the following, we first have deleted all nodes with a degree lower than two (representing rare diseases) and then we have computed the islands with a minimum size of 2 and maximum size of 20.

3 First results

By using the island algorithm implemented in Pajek software (6) on the full one-mode network including all epidemiological codes and all patients, 28 islands are obtained (Figure 1): one large group consisting of 20 ICD-9-CM codes and 27 islands of size 2-4 nodes. All the islands are included in one large component. As for comorbidity, the largest island includes cardiovascular disease, hypertension, osteoarthritis and osteoporosis, diabetes, prostatic hypertrophy, atherosclerosis, carotid artery stenosis, kidney disease, cystitis, and thyroiditis. In the smaller islands, we found the check up and treatment for pregnancy and infertility, as well as links among different kinds of cancer. It seems that there is a strong comorbidity structure that includes a set of diseases already linked in the clinical studies (e.g. cardiovascular disease and hypertension, diabetes, and so on). Others are probably related to the patient's age (diseases of the elderly such as arthritis, osteoporosis, prostatic hypertrophy) or to the patient's gender (e.g., pregnancy and infertility).

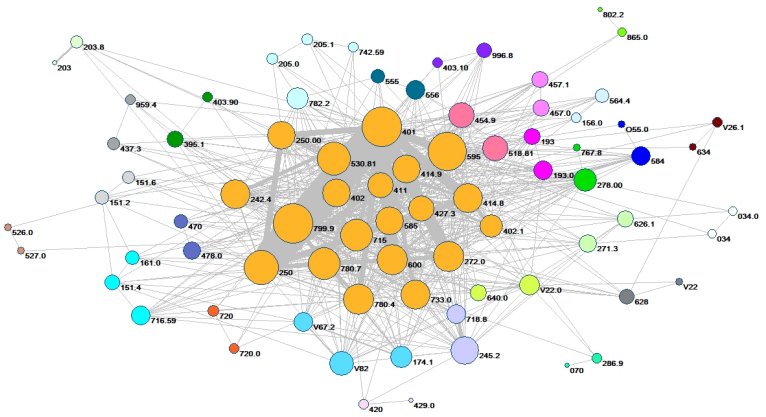
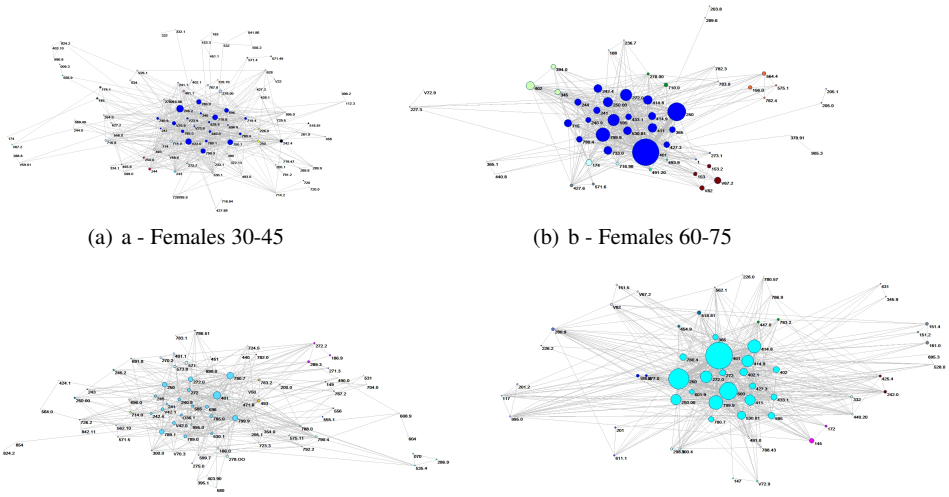


Fig. 1 Islands of size 2-20 in the comorbidity network. Node color= islands' partition; node size= nodes' degree.

Based on the data available, it is also possible to explore the effect of sex and age on the comorbidity network structure. To do that, the two-mode network can be divided into subnetworks according to these attributes associated with the prescriptions, and then specific one-mode networks can be derived and analyzed. By comparing the results of the islands algorithm for males and females divided by different age groups, different comorbidity patterns appear. For example, in the core of the main island of young females (Figure 2a), we found thyroiditis, gynecological problems, pregnancy, menstrual cramps, and cystitis; while on the periphery of the

island, obesity, lipidosis, breast and thyroid cancer, arthritis and osteoarthritis were found. For older men (Figure 2d), in the core of the largest island we found arterial hypertension, prostatic hypertrophy, diabetes, heart disease, renal colic, and bronchitis; while, on the periphery, we found periodic check up after cancer, psychosis and depression, glaucoma, prostate cancer, and diverticula.



(a) a - Females 30-45

(b) b - Females 60-75

(c) c- Males 30-45

(d) d - Males 60-75

Fig. 2 Islands of size 2-20 in the comorbidity networks of females and males in the two age intervals. Node color= islands' partition; node size= nodes' degree.

4 Further developments

As the first results have shown, the prescription database is a very rich source of information about people, diagnosis prevalence, and temporal emergence of diagnosis as a proxy of incidence index. The possibility of handling this kind of data in terms of networks seems very promising (5). The graph representation allows many algorithmic tools from the network domain. Graphs can be built on patients, prescriptions, or diagnosis, according to different aims and kinds of investigation. For instance, a graph whose nodes are patients and has links established as common diagnosis or prescriptions could be used to predict diagnosis as new links between nodes. Alternatively, building the graph on prescriptions tied by common diagnoses allows us to highlight specific occurrences of comorbidity. In both cases, network techniques can be exploited to extract relevant information. Looking for cohesive sets of patients and diagnoses calls for community detection methods. Such com-

munities are dense parts of a graph that could reveal important features like frequent patterns of comorbidity.

Beyond the Islands algorithm method here proposed for the treatment of large and dense comorbidity networks, another promising technique we are investigating for this case study is the extraction of the association rules for bipartite network data (1). This technique is strictly related to the aim of finding frequent item sets in a large dataset and commonly applied in transactional data for marketing strategy. Applying association rules in medical diagnosis can be used to assist physicians in making a diagnosis. Even if reliable diagnostic rules are difficult and may result in hypotheses with unsatisfactory predictions, which are too unreliable for critical medical applications (7). Serban et al. (11) has proposed a technique based on relational association rules and supervised learning methods. It helps to identify the probability of illness in a certain disease.

In the case of medical prescriptions, the collection of prescriptions per diagnosis can be arranged, as defined above, in the bipartite graph $\mathscr{B}$ and arranged in a transaction matrix where each prescription is a transaction defined on the set $\mathscr{V}_2$, so that $\mathscr{V}_1$ is the universal set of items and each prescription in $\mathscr{V}_2$ is a subset of $\mathscr{V}_1$. In this context, the aim stated in the association rules setting is to find set of diagnosis that are strongly correlated in the prescription database and is coherent with the notion of close neighbors in bipartite graphs. In this context, the aim stated in the association rules setting is to find a set of diagnoses that is strongly correlated in the prescription database and is coherent with the notion of close neighbors in bipartite graphs.

The usual metrics derived from the association rules context –*Support*, *Confidence*, and *Lift*– will be used to characterize the graph representation of the diagnosis item set, defined as the projection of the bipartite graph induced by the prescription database. For the sake of computational efficiency, the analysis could concentrate on the subset of diagnosis having a minimum Support S. On the other hand, if the graph is partitioned in k disjoint communities (islands or blocks), it makes sense to apply the association rules search on the separate subgraph induced by the partition. The more meaningful rules are usually sorted by decreasing the value of *Lift*. The first important rules among diagnosis will help to uncover association between frequent patterns of diagnosis co-occurrence in the whole set of prescriptions.

References

[1] Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207-216. ACM, New York (1993)

[2] Batagelj, V.: Social network analysis, large-scale. In: Encyclopedia of Complexity and Systems Science, pp. 8245-8265. Springer, New York. (2009)

[3] Batagelj, V., Doreian, P., Ferligoj, A., Kejzar, N., : Understanding large temporal networks and spatial networks: Exploration, pattern searching, visualization and network evolution (Vol. 2). John Wiley & Sons, United Kingdom (2014)

[4] Capobianco, E., Lio, P.: Comorbidity: a multidimensional approach. Trends Mol Med **19**, 515–521 (2013)

[5] Cavallo, P., Pagano, S., Boccia, G., De Caro, F., De Santis, M., Capunzo, M.: Network analysis of drug prescriptions. Pharmacoepidemiol Drug Saf **22**, 130–137 (2013)

[6] De Nooy, W., Mrvar, A., Batagelj, V.: Exploratory social network analysis with Pajek (Vol. 27). Cambridge University Press, Cambridge (2011)

[7] Gamberger, D., Lavrac, N., Jovanoski, V.: High confidence association rules for medical diagnosis, In: Proceedings of IDAMAP99, pp. 42–51 (1999)

[8] Mercer, S.W., Smith, S.M., Wyke, S., O'dowd, T., Watt, G.C.: Multimorbidity in primary care: developing the research agenda. Family Practice **26**, 7980 (2009) Available via DIALOG. https://academic.oup.com/fampra/article/26/2/79/2367540. Cited 07 May 2018

[9] Pfaundler, M. and von Seht, L.: Uber Syntropie von Krankheitszustanden, Z. Kinderheilk. vol. **30**, 298–313 (1921)

[10] Puzyrev, V.P.: Genetic Bases of Human Comorbidity. Genetika **51**, 491–502 (2015)

[11] Serban, G., Czibula, I. G., Campan A.: A Programming Interface for Medical diagnosis Prediction. Studia Univ. Babes-Bolyai, Informatica **LI**, 21–30 (2006)

[12] Valderas, J.M., Starfield, B., Sibbald, C., Salisbury, M. Roland: Defining comorbidity: implications for understanding health and health services. Ann Fam Med, **7**, 357–63 (2009)

[13] Yurkovich, M., Avina-Zubieta J.A., Thomas J., Gorenchtein M., Lacaille D.: A systematic review identifies valid co-morbidity indices derived from administrative health data. J Clin Epidemiol **68**, 3–14 (2015)