Association rule and network analysis for exploring comorbidity patterns

Giuseppe Giordano[1], *Pierpaolo Cavallo*[2], *Sergio Pagano*[2], *Giancarlo Ragozini*[3], *Maria Prosperina Vitale*[1]

[1]Department of Economics and Statistics, University of Salerno (Italy)
[2]Department of Physics E.R. Caianiello, University of Salerno (Italy)
[3]Department of Political Science, University of Naples Federico II (Italy)

ASA2018, September 12-14, 2018 - Pescara
Session *Health Risk Management*

Talk Outline

The framework

Comorbidity

- Comorbidity can be defined as the presence of different diseases at the same time in the same person (Pfaundler & von Seht, 1921)

Complex nature of comorbidity

- Syntropy –appearance of two or more diseases in the same individual– and Dystropy (Puzyrev 2015) –pathologies that are rarely found in the same patient at the same time.
- Comorbidity (Valderas et al., 2009) –coexistence of conditions that are linked, either biologically or functionally– and multimorbidity (Mercer et al., 2018) –coexistence of two or more long-term conditions in an individual not biologically or functionally linked.

Why comorbidity

- The presence of patients affected by many different diseases 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 (Mercer et al., 2018)

Measurement and analysis issues

Measurement issues

Lack of agreement (Capobianco, Lio 2013) on how to understand the complex interdependent relationships between diseases due to:

- a large number of variables;
- a lack of accuracy in measurements;
- technological limitations in generating data.

Analysis Issues

- The study of comorbidity usually requires complex clinical studies
- Indirect approach using health system databases can be adopted

Indirect approach

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.

- Access to prescription data from GPs is relatively simple $\longrightarrow$ data 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

The strategy

- Italian National Health System rules $\rightarrow$ 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
- ICD standard diagnostic tool for epidemiology, health management, and clinical purposes (World Health Organization) $\rightarrow$ to monitor the incidence and prevalence of diseases and other population health problems
- Comorbidities (Yurkovich et al., 2015) $\rightarrow$ studied from GP databases based on both diagnoses, using the International Classification of Diseases –ICD– codes, and medications, using pharmacy data

Our approach

We propose to handle this kind of data in terms of networks

We define and analyze the comorbidity networks

The Case study

- Electronic Health Recordings (EHR) of the prescriptions made by a group of GPs belonging to the Cooperative Medi Service and operating in a town in Southern Italy
  - 14,958 patients
  - 1,728,736 prescriptions covering a time interval of eleven years from 2002 to 2013

Information in GPs administrative prescription data

- 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.

Comorbidity networks

Comorbidity patterns described as a network

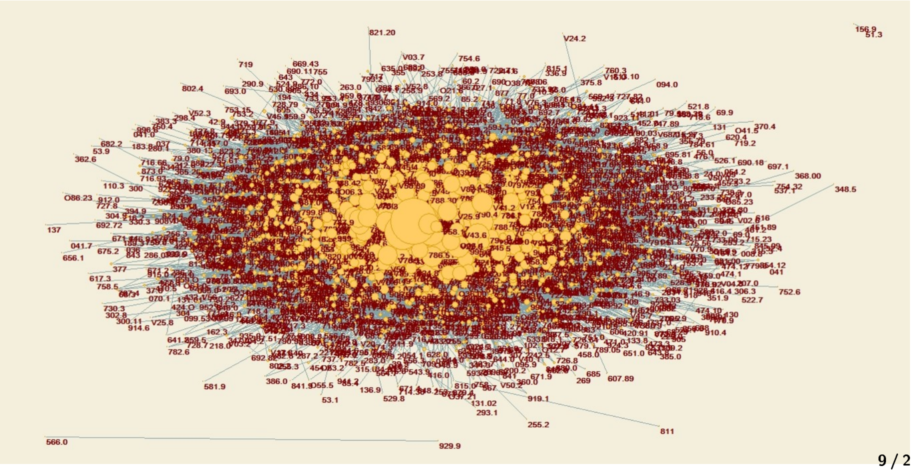
- Two-mode network $\rightarrow$ ICD9CM diagnostic codes and prescriptions (two disjoint sets of nodes)
  $\rightarrow$ Links: if corresponding codes appear in prescriptions made to the same patient on the same day
  - Bipartite graph $\mathcal{B}$ consisting of the two sets of relationally connected nodes and can be represented by a triple $\mathcal{B}(\mathcal{V}_1, \mathcal{V}_2, \mathcal{L})$, with $\mathcal{V}_1$ denoting the set of ICD-9-CM codes, $\mathcal{V}_2$ the set of prescriptions, and $\mathcal{L} \subseteq \mathcal{V}_1 \times \mathcal{V}_2$ the set of ties
- One-mode network of the ICD-9-CM codes by projecting the two-mode network
  - Graph $\mathcal{G}(\mathcal{V}_1, \mathcal{E}, \mathcal{W})$, with $\mathcal{V}_1$ the set of ICD-9-CM codes, $\mathcal{E} \subseteq \mathcal{V}_1 \times \mathcal{V}_1$ the set of edges, and $\mathcal{W}$ the set of weights, $w : \mathcal{E} \rightarrow \mathcal{N}$, $w(v_{1i}, v_{1j}) =$ the number of times that two ICD-9-CM codes appear in the same prescription

The sex and age of patients, and type and time of prescriptions can be considered as attributes of a given prescription

Comorbidity networks

Comorbidity patterns described as a network

The resulting networks are very large, dense and complex.
Proper statistical tools are need to mine such a complexity.



Approaches for the analysis of co-morbidity network

We propose to use:

Association rules

- Association rules extraction (Agrawal, 1993) for two-mode comorbidity networks $\rightarrow$ technique useful for finding frequent itemsets in a large dataset helping to identify the probability of illness in a certain disease

Association rules: Concept and notation

Association Rules Definition

Let

- $J = i_1, i_2, \ldots, i_m$ be a set of items. (i.e. Diagnoses)
- $D =$ be a set of transactions where each transaction T is a set of items such that $T \subseteq J$. (i.e. Set of Prescriptions or Patients)

An association rule extracted from D is an implication of the form $A \Rightarrow B$, where $A \subset J, B \subset J$, and $A \cap B = \varnothing$.

A is the left-hand side of the Rule (LHS)

B is the right-hand side of the Rule (RHS)

Association rules: Concept and notation

Association Rules Metrics

The **Support** of an itemset A is the percentage of transaction in D that contain A.

Given a rule $A \Rightarrow B$ in the transaction set D

- the **Confidence** is the percentage of transaction in D containing A that also contain B, i.e., $conf(A \Rightarrow B) = [supp(A \cup B)/supp(A)]$.
- the **Lift** is the ratio of the observed support of a rule to that expected if A and B were independent, i.e. $lift(A \Rightarrow B) = \dfrac{supp(A \cup B)}{supp(A) \times supp(B)}$

Association rules: Interpretation

Association Rules Metrics

- *Support* is an indication of how frequently the itemset appears in the dataset.
- *Confidence* is an indication of how often the rule has been found to be true.
- *Lift* If the lift is > 1, that lets us know the degree to which those two occurrences are dependent on one another

Used to characterize the graph representation of the diagnosis item set $\rightarrow$ rules sorted by decreasing the value of *Lift* to uncover association between frequent patterns of diagnosis co-occurrence in the whole set of prescriptions toward patients

Back to the Case study

- Electronic Health Recordings (EHR) of the prescriptions made by a group of GPs belonging to the Cooperative Medi Service and operating in a town in Southern Italy
  - 14,958 patients
  - 1,728,736 prescriptions covering a time interval of eleven years from 2002 to 2013

Information in GPs administrative prescription data

- 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.

First Analysis

Network analysis for males and females divided by different age groups

- core of young females $\rightarrow$ thyroiditis, gynecological problems, pregnancy, menstrual cramps, and cystitis
- periphery of young females $\rightarrow$ obesity, lipidosis, breast and thyroid cancer, arthritis and osteoarthritis
- core of older men $\rightarrow$ arterial hypertension, prostatic hypertrophy, diabetes, heart disease, renal colic, and bronchitis
- periphery of older men $\rightarrow$ periodic check up after cancer, psychosis and depression, glaucoma, prostate cancer, and diverticula

Organizing Data for Association rules

To carry out Association Rules we reduced the original database to peculiar prescriptions and patients:

- Type = DRUG: that are commonly related to actual diagnosis
- Not relevant ICD9CM codes: Pregnancy, Congenit, Newborn, Ill-defined, etc.
- Ages range: from 35 to 110 years
- Gender: Male and Female (two databases)

Total Unique Drugs: 627,924 Total Prescriptions: 405,323 (each Prescription is a subset of the Drugs) Total Patients: 9,845 (each Patient is a subset of the Prescriptions)

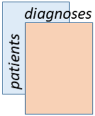
Association Rules

From Prescription Data to Transaction Matrix

405,323 Prescriptions
Female: 220,469
Male: 184,854

627,924 Unique Drugs
Female: 341,368
Male: 286,556

9,845 Patients
Female: 5,252
Male: 4,593

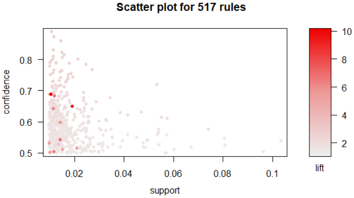


«Transactions» Matrices

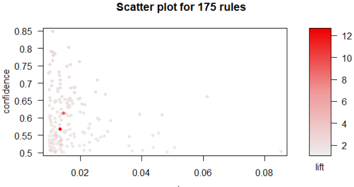
2,387 Diagnoses
Female: 1,214
Male: 1,173

"*Apriori*" (Agrawal R. and Srikant R., 1994) Association rules results - Females Vs/ Males

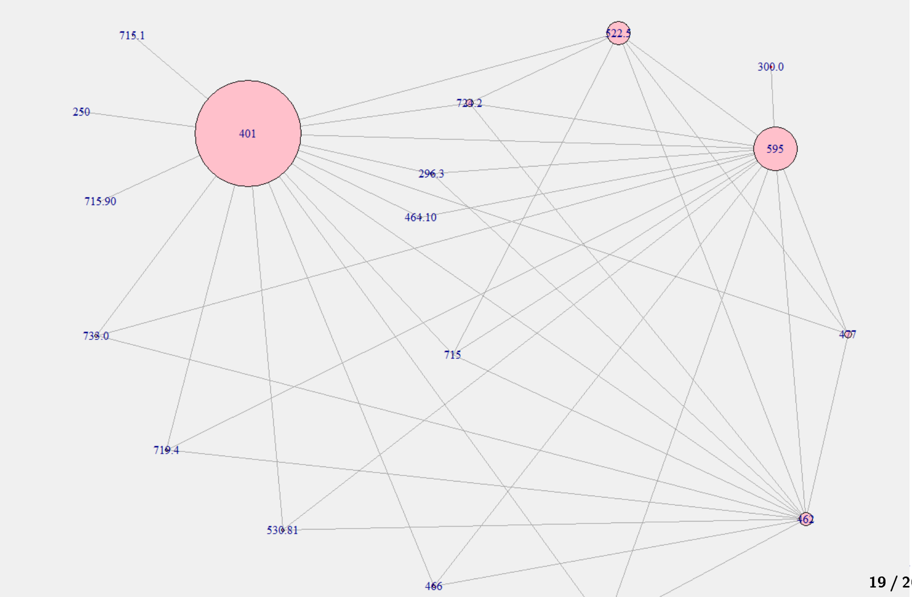
Females



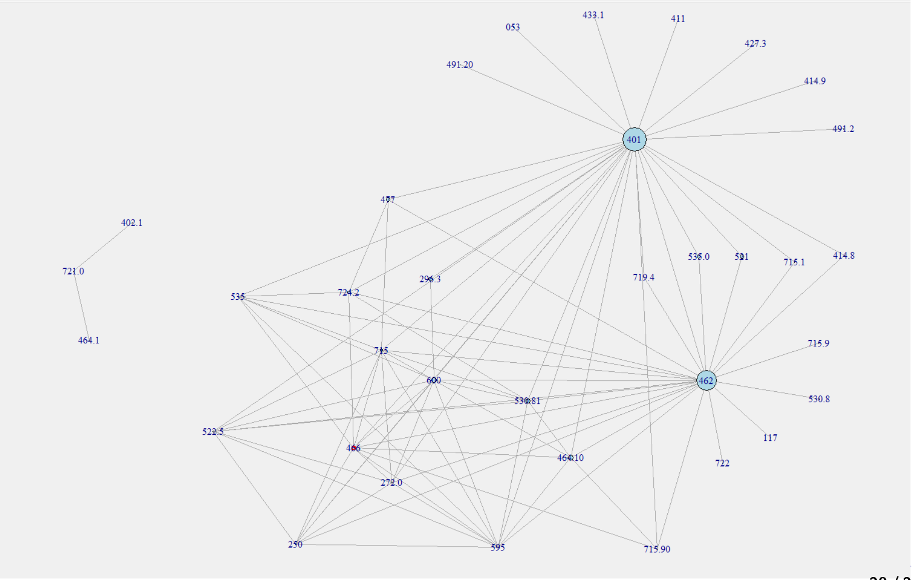
Males



Network Association rules results - Female



Network Association rules results - Male



Network Analysis of Comorbidity
Main Females' Diagnoses sorted by betweenness

code	label	degree	betweenness	eigenvector	closeness
462	Faringitis	0.88	0.527	0.915	0.455
595	Cystites	0.76	0.186	1.000	0.373
401	Hypertension	0.76	0.185	0.993	0.309
466	Bronchitis	0.20	0.136	0.094	0.385
715	Arthrosis	0.36	0.117	0.473	0.333
724.2	Back pain	0.20	0.080	0.095	0.385
477	Allergic rhinitis	0.16	0.076	0.073	0.342
522.5	Periapical abscess	0.36	0.066	0.274	0.333
296.3	Depression	0.16	0.056	0.077	0.385
715.1	Arthrosis	0.12	0.020	0.070	0.352

Table: Network Statistics for Diagnoses Graph

Network Analysis of Comorbidity
Main Males' Diagnoses sorted by betweenness

code	label	degree	betweenness	eigenvector	closeness
401	Hypertension	0.941	0.535	0.896	0.370
595	Cystites	0.824	0.219	1.000	0.304
522.5	Periapical abscess	0.353	0.119	0.509	0.230
462	Faringitis	0.706	0.068	0.958	0.254
477	Allergic rhinitis	0.235	0.035	0.216	0.283
724.2	Back pain	0.235	0.035	0.216	0.283
719.4	Arthralgia	0.176	0.013	0.183	0.258
733.0	Osteoporosis	0.176	0.013	0.183	0.258
530.81	Refluxo esofageo	0.176	0.013	0.183	0.258
296.3	Depression	0.176	0.013	0.183	0.258

Association rules Mining

Mining Association Rules by graphical interface

Show 10 entries Search:

	LHS	RHS	support	confidence	lift	count
	All	All	All	All	All	All
[1]	{477.0}	{464.1}	0.011	0.687	10.131	57.000
[2]	{411}	{401}	0.011	0.811	2.466	60.000
[3]	{112.9}	{462}	0.011	0.670	1.771	59.000
[4]	{715.5}	{401}	0.011	0.756	2.300	59.000
[5]	{490.0}	{462}	0.012	0.619	1.635	65.000
[6]	{703.1}	{462}	0.010	0.529	1.399	54.000
[7]	{285}	{595}	0.010	0.520	1.344	53.000
[8]	{626.4}	{462}	0.012	0.537	1.419	65.000
[9]	{491}	{401}	0.011	0.617	1.876	58.000
[10]	{715.89}	{462}	0.012	0.571	1.510	64.000

Showing 1 to 10 of 517 entries Previous 1 2 3 4 5 … 52 Next

Figure: Screen shot of the navigation tool of association rules.

Concluding

Concluding remarks

1. Strategy to analyze comorbidity networks based on data mining techniques
2. Comorbidity data have been read as network data exploiting both the graph visualization and the analytical tools of SNA
3. Explicative power of the proposed strategy on real data-set of diagnoses in prescriptions

Future lines of research

1. to exploit the summarizing properties of community detection algorithm as a tool to enhance association rules
2. to assess prediction rules both with internal and external validation methods, i.e. cross validation and applying rules to different medical data set

Main references

- 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)
- Batagelj, V.: Social network analysis, large-scale. In: Encyclopedia of Complexity and Systems Science, pp. 8245-8265. Springer, New York. (2009)
- 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)
- Capobianco, E., Lio, P.: Comorbidity: a multidimensional approach. Trends Mol Med 19, 515–521 (2013)
- 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, 79–80 (2009)
- Pfaundler, M. and von Seht, L.: Uber Syntropie von Krankheitszustanden, Z. Kinderheilk. vol. **30**, 298–313 (1921)
- Puzyrev, V.P.: Genetic Bases of Human Comorbidity. Genetika 51, 491–502 (2015)
- 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)

Any questions?

Thank you for your attention