Network analysis of comorbidity patterns in heart failure patients using administrative data

Francesca Ieva[1], *Daniele Bitonti*[1]

(1) MOX– Modeling and Scientific Computing Department of Mathematics, Politecnico di Milano

CORRESPONDING AUTHOR: *Francesca Ieva, Department of Mathematics, 6th floor, Politecnico di Milano, via Bonardi 9, 20133 Milano (IT). Phone: 02 2399 4578. email: francesca.ieva@polimi.it*

DOI: ***10.2427/12779***
Accepted on February 15, 2018

ABSTRACT

Background: Congestive Heart Failure (HF) is a widespread chronic disease characterized by a very high incidence in elder people. The high mortality and readmission rate of HF strongly depends on the complicated morbidity scenario often characterising it. The aim of this paper is to show the potential and the usefulness of Network models when applied to the analysis of comorbidity patterns in HF, as a new methodological tool to be considered within the epidemiological investigation of this complex disease.
Methods: Data were retrieved from the healthcare administrative datawarehouse of Lombardy, the most populated regional district in Italy. Network analysis techniques and community detection algorithms are applied to comorbidities registered in hospital discharge papers of HF patients, in 7 cohorts between 2006 and 2012.
Results: The relevance network indexes applied to the 7 cohorts identified, hypertension, arrythmia, renal and pulmonary diseases as the most relevant nodes related to death, in terms of prevalence and closeness/strength of the relationship. Moreover, some relevant clusters of nodes have been identified in all the cohorts, i.e. those related to cancer, lung diseases liver diseases and heart/circulation related problems. It seems that such patterns do not evolve along time (i.e., nor indexes of relevance computed on the nodes of the networks neither communities change significantly from one year/cohort to another), featuring HF comorbidity burden as stable over the years.
Conclusions: Network analysis can be a useful tool in epidemiologic framework when relational data are the objective of the investigation, since it allows to visualize and make inference on patterns of association among nodes (here HF comorbidities) by means of both qualitative indexes and clustering techniques.

Key words: Network Analysis; Administrative databases; Heart Failure; Comorbidities

INTRODUCTION

Congestive Heart Failure (HF in the following) is a widespread chronic disease characterized by a very high incidence in elder people [1]. HF prevalence steeply increases with aging [2]. One year mortality ranges from 35-40% and more than 50% of patients are readmitted to hospital between 6 months and 1 year after the diagnosis, due to a complicated morbidity scenario, among others. In this epidemiological setting, elders with

HF are representative of a growing segment living longer with chronic conditions prone to multiple transitions from hospital to home and vice versa. This unavoidably affects their quality of life, and turns in an important healthcare management and costs issue. Last but not least, in such a context it is pretty unreasonable to consider the health status of a patient as due to a "main" disease surrounded by other possible minor diseases. It is more often the case that more than one condition contributes to determine the health need and consumption.

Another issue related to HF and related healthcare practice and management is the following: it is more and more common nowadays to make use of secondary databases to conduct epidemiological enquires concerning HF. In fact, patients with HF randomized in controlled trials are generally selected and do not fully represent the "real world" [3].

For all these reasons, the objective of our study is to show the potential, the usefulness and the advantages of applying Network analysis [4,5,6] and in general a relational approach in the study of the comorbidities recorded in hospitalizations charts of HF patients [7]. Specifically, we wish to highlight if the same pattern of relationships/connection among comorbidities is maintained over the time window of interest (we analyse 7 cohorts, one per year from 2006 to 2012, as specified in Section 2), possibly quantifying the strength of the connection among different comorbidities and death. Moreover, we would like to detect groups/communities of comorbidities which are more strongly connected among each other. Last but not least, we aim at doing this for the first time in literature using administrative data [8,9].

The article is organized as follow: after an introduction to the basics of network analysis and a brief description of data, we illustrate the applications of network analysis to our data and finally the results' discussion.

METHODS

Network analysis in a nutshell

A network is a graph with N nodes (or vertices) and L links (or edges) that can be weighted or unweighted, directed or not. An unweighted network is completely represented by its N x N adjacency matrix A such that A_{ij} = 1 if node i points to node j, A_{ij} = 0 otherwise.

Let G = (V; E) be a graph, where V is the set of its vertices such that |V| = N and E is the set of its edges such that |E| = L. Edges may denote just the connection among two nodes or being labeled with a number indicating weights assigned to them. In the latter case, we graph is called *weighted*.

There are many important properties through which a network can be described [4,6], providing interesting insight of the phenomenon the network is representing (in our case, the connection among comorbidities in HF patients). Some of the most relevant, among others, are:

- **Degree**: it is he simplest way to measure the importance of a node, consisting of the count o the number of neighbors. A vertex can be considered as more important than the others in the network if it has a greater degree with respect to the others. In the current case, the degree of a node measures the number of pathologies connected to that node.
- **Strength**: in a weighted network, the strength is the sum of the weights on the links connected to a given node. In the current case, it measures the strength of the connection of a given pathology with other pathologies witin the network.
- **Weighted local transitivity or closeness centrality**: it quantifies how many vertices are connected to each other among the neighbors of a given node. In the current case, it measures the proximity of a given pathology to other pathologies.

It can be also of interest to group nodes together according to their level of similarity. Community detection algorithms [10,11,12] are used to reach this goal. For further details and mathematical definition of the aforementioned indexes, as well as for deeper explanation of community detection algorithms, see [5] and references therein.

Setting

Data were retrieved from e the healthcare system of Lombardy, Italy, a region of Italy which accounts for about 16% (almost ten million) of its population.Hospital discharge forms with HF-related diagnosis codes were the basis for identifying HF hospitalizations as clinical events, or episodesWith the aim of identifying hospitalizations for HF, data on hospitalizations in Major Diagnostic Categories (MDC) 1, 4, 5 and 11 in the years from 2000 to 2012 have been extracted. Data on hospital admissions of Lombardy residents in other regions for the same MDC were also requested. In-hospital deaths were collected from hospital discharge forms database, while data on out of hospital deaths were retrieved from vital statistics regional dataset. The presence of an ID (identification) code was used to identify the patient over the years and across the different data sources. The ID code was made anonymous to respect privacy. After a comprehensive literature review and an open discussion between epidemiologists, statisticians and clinicians, two criteria were chosen to obtain a complete and accurate selection of HF cases: indicators proposed by the Agency for Healthcare Research and Quality (AHRQ) [13] and HF codes as identified by the Center for Medicare and Medicaid Services (CMS) [14]. Figure 1 and Table 1 in [16] provide a detailed list of the codes used for the cohort identification. Data from 2000 to 2005 have been

TABLE 1. Patients in each cohort from 2006 to 2012.

Year	2006	2007	2008	2009	2010	2011	2012
N° of pts.	4,813	8,627	12,082	15,769	21,619	29,933	49,744

used to identify the incident cases. Comorbidities were evaluated with the method proposed in [16]. Appendix A reports a legend of the comorbidities arising from the algorithm detailed in the authors website. One important detail concerning the recognition of comorbidities is the so-called "look-back period", i.e., the time prior to the hospitalization that represents the index event. This period must be analyzed to intercept comorbidities that may not be reported within the diagnosis list of the current hospitalization event. It is suggested from literature that a period of 1 year should be sufficient for identifying comorbidities that influence the patient' probability of survival. Therefore, a period of 1 year prior to the incident hospitalization was considered for recovering information about patient's comorbidities at that time. Full details about the dataset and selection criteria of the cohort are reported in [15] and [17].

The final dataset considered for this work is a representative subset of 142,587 patients, distributed over the years as presented in Table 1.

Each patient appears only in the cohort (i.e., in the network) related to the year of his/her last discharge.

Data analysis

Analyses are carried out with R software [18,19] and network dedicated packages, like igraph [20].

We consider only the last hospitalization of each patient in the period 2006-2012, since it is assumed to describe his/her most compromised clinical condition. In doing so, 7 cohorts (networks) were established, one per year of the period 2006-2012,d, where each patient contributes only to the year his/her last hospitalization happens within. Originally we deal with bipartite networks, i.e., a network whose vertices can be divided into two disjoint and independent sets (say U and V) such that every edge connects a vertex in U to one in V. In our case, patients and comorbidities act as the two disjoint sets.We then get the networks used for the analysis projecting the bipartite network "patients-comorbidity" on the "comorbidity" dimension.. Therefore, nodes are represented by comorbidities (death is a node of the comorbidity network, since we want to identify which pathologies are most connected to it). Two nodes are connected by an edge, weighted according to the amount of patients presenting that couple of comorbidities. The strength of the association between two nodes is measured in terms of ϕ-correlation [21]. For each patient, in addiction to the comorbidities and death/survival indicators, information about age [years] and gender are available.

From the procedure described above, we get a dense network [4], i.e., a network in which each node is linked to almost all other nodes, which is odd to treat both from a modelling and computational point of view. Therefore, a thresholding [5] is needed, and we adopted the following criterion: let G be the undirected network (i.e., a network where all the edges are bidirectional) under study, and t a prescribed or desired density for the network. Then the network density (defined as r = L/[N(N-1)/2], where L and N the number of links and nodes of the network G, respectively) can be tuned in order to maintain edges only if they fulfill the requirement $\phi > t$.

For each node in each network, an index of relevance is computed. The index is composed by *degree centrality, strength, weighted local transitivity or closeness centrality* and *prevalence* of that node. The index is then constituted by 4 components, and a node is relevant if it presents high values in each component. This allows to identify which nodes are more relevant within each network and within each year.

Finally, a community detection algorithm based on modularity maximization [10,11,12] is applied in order to find relevant communities of nodes within the networks.

The current methodology may help the analysis and detection of possible evolution of morbidity patterns accompanying HF and their relationship with death over the years in a twofold way: first, this kind of approach moves the attention from the outcome-covariates relationship to the relationship among variables themselves (here comorbidities); secondly, it provides quantitative indexes describing the network which might be monitored over time.

RESULTS

The procedure described in the last Section results in 7 networks to be analyzed. We reduced the density of the graphs considering only links that had a ϕ-correlation greater than $\tau = 0.02$. This is a reasonable trade off between the necessity of reducing the density of the networks, and the ability of capturing the relevant connections among nodes.

Figure 1 shows networks concerning the years 2006 and 2012, . The shape of the nodes (comorbidities) are defined according to the presence of men (higher if the node is square shaped) or women (higher if the node is circle shaped) presenting that pathology, and the colours are related to the corresponding prevalence (the higher the prevalence, the darker the colour). The thickness of the edge is proportional to the number of patients presenting both the pathologies.

FIGURE 1. Representations of the 2006 (left panel) and 2012 (right panel) networks.

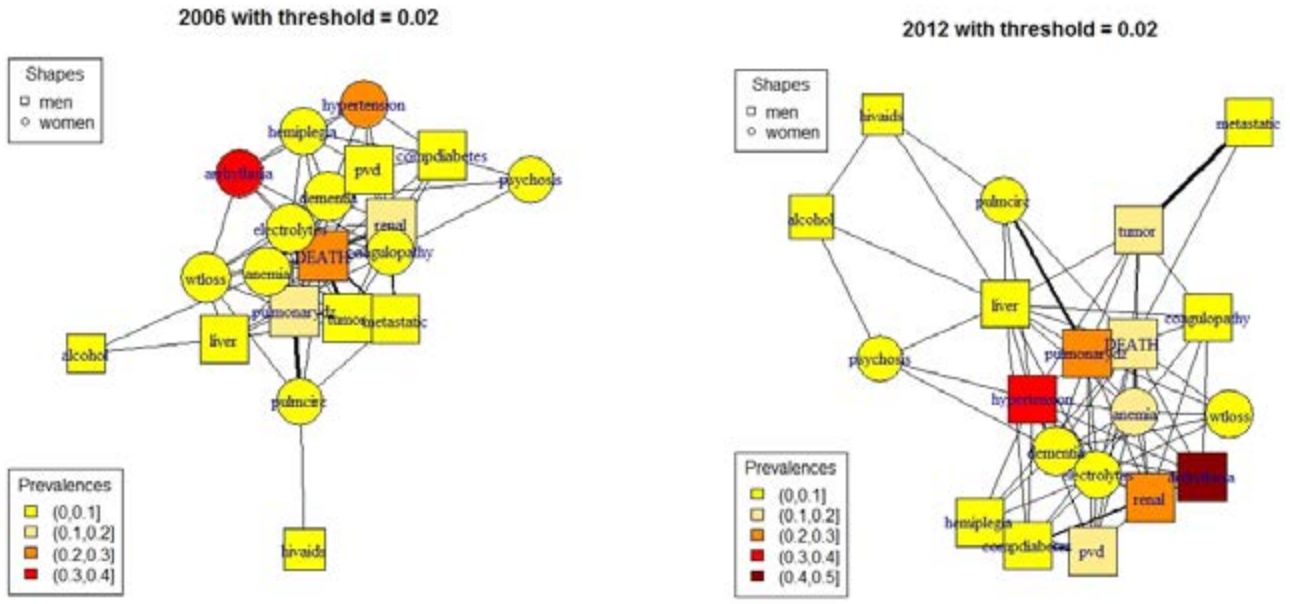
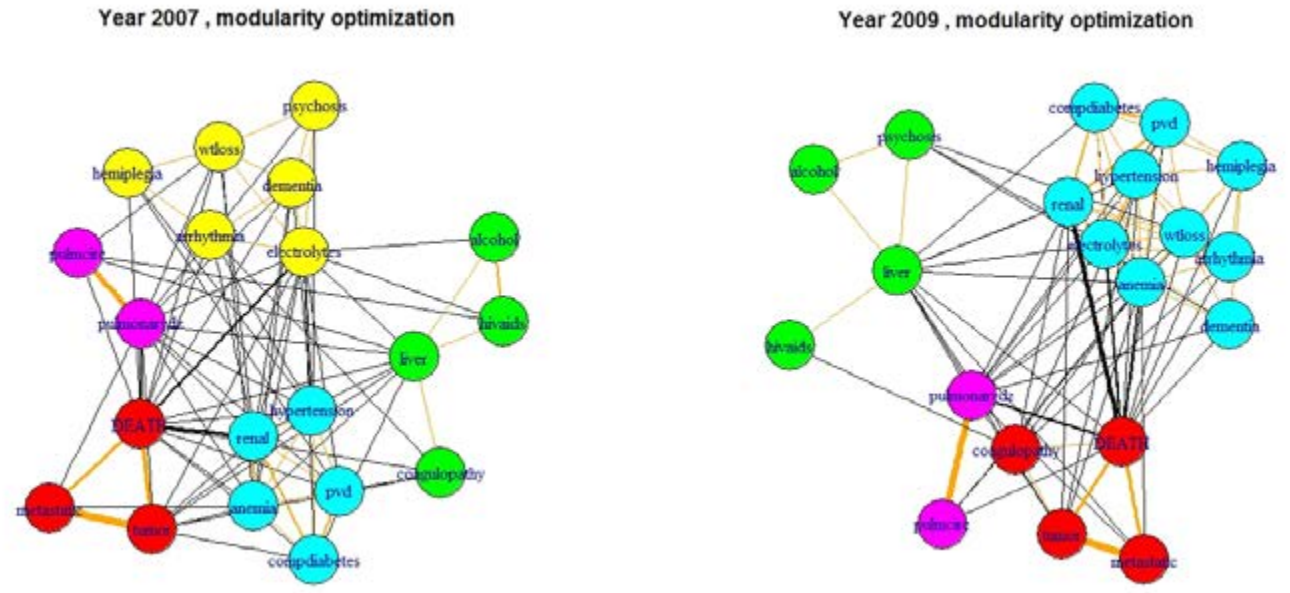


FIGURE 2. Communities of nodes (i.e., comorbidities) detected in the 2007 (left panel) and 2009 (right panel) networks.



In order to investigate if the relationships among comorbidities in HF (and among comorbidities and death) remain the same over the years, we compared the patterns presented by each network both in terms of indexes and communities detected.

The relevance indexes described in the previous Section and applied to each network identified hypertension, arrhythmia, renal and pulmonary diseases as the most relevant nodes related to death. This means that their prevalence and closeness centrality result higher than the others. They are also the most strongly connected among each other.

Figure 2 shows the communities detected in 2007 and 2009 cohorts, which are present in almost all the cohorts in the same configuration. The communities are those related to cancer, lung diseases, liver diseases and heart/circulation related problems. Each community is identified by a different color. .

These results show that even in a simple example like the one proposed, patterns of connections among comorbidities related to HF may be discovered and monitored in their relationships with death over time, given proper definition of the cohorts. From these preliminary

results, it seems that such patterns do not evolve along time (i.e., nor indexes neither communities change significantly from one year/cohort to another), featuring HF comorbidity burden as stable over the years. Further investigations are needed to consider potential risk profiles of patients to be monitored in dedicated programs.

DISCUSSION AND FURTHER DEVELOPMENTS

In this work we showed a promising approach to the analysis of comorbidity patterns in patients affected by HF using networks. It represents an innovative and flexible method that can be adopted for many different kind of epidemiological investigations.

The main novelty introduced by the network modeling approach is the idea of exploiting the relational aspect of comorbidity patterns within the epidemiological analysis of a given disease (here HF). To the best of our knowledge, there is not a wide literature treating the analysis of comorbidities in HF from a relational point of view. In fact, all the regression/survival based methods focus on correlations of a given set of independent variables with an outcome of interest. Here the interest lies in the relations existing among variables (morbidities), and the focus is on the determinants of the presence of a given relationship, instead of the correlation between such variables and the final outcome. This makes it unfruitful and unfair the comparison with techniques like survival analysis of regression analysis, which are aimed at different goals with respect to network analysis. Investigations on HF based on these techniques using the same data may be found in [17], [22] and [23].

Anyway, some features emerged thanks to the network approach we adopted might be exploited in subsequent analyses based on more classical statistical methods. For example, survival and/or (logistic) regression models may be implemented, building suitable (possibly dynamic) comorbidity indexes to be inserted among the covariates.

There are no distributional assumptions that data are required to fulfill in order to carry out the proposed analysis analysis, and this is another advantage of the network approach. Weaknesses, if any, consist of the amount of choices (projections, thresholding values and so on) which are needed to practically build the networks from administrative data, since they come out from not from a relational analysis context. In general, despite the limitations induced by the nature of administrative data (e.g., limited epidemiological contents), network analysis can be considered a useful tool in epidemiologic framework when relational data are the objective of the investigation, since it allows to visualize and make inference on patterns of association among nodes (here HF comorbidities) by means of both quantitative indexes and clustering techniques. This is particularly relevant when the size of the network (i.e., the number of nodes) becomes high.

Future developments of the present work may regard:

I. To increase the size of the network, using DRGs instead of comorbidities.
II. To consider bipartite networks of patients and comorbidities (or diagnoses) directly, without projecting and thresholding. T;
III. To define an univariate index that takes the prevalence, degree, strength and closeness into account, properly weighting their contributes (possibly according to clinicians' suggestions);
IV. To refine the community detection, exploiting techniques like stochastic block models (SBM) [24] or latent class models for bipartite networks.

Using DRG codes (point (I)) associated to the (possibly) six diagnosis fields of the electronic health record would allow for the construction of networks with a larger number of nodes (one for each DRG mentioned for the patient) with respect to the actual one based on comorbidities. This would enable a wider investigation of the pathology the patient is affected by.

On the other hand, suggestion (II) and (IV) go the direction of the application of suitable clustering and community detection algorithms directly on the original network, avoiding conceptual and computational problems (and related methodological choices) induced by projection.

Extension (III) is intended as a clinical refinement that might be used to summarize the results in a more effective way.

Acknowledgements

The work has been developed within the HEAD project "research on Health and Education systems Assessment using administrative Data", funded by Politecnico di Milano. The authors acknowledge the Project "Utilization of Regional Health Service databases for evaluating epidemiology, short- and medium-term outcome, and process indexes in patients hospitalized for heart failure", for providing data.

References

1. Bui, A.L., Horwich, T.B., Fonarow, G.C. (2011) Epidemiology and risk profile of heart failure. Nature Reviews Cardiology, 8: 30-41.
2. Curtis, L.H., Whellan, D.J., Hammill, B.G., et al. (2008) Incidence and prevalence of heart failure in elderly persons, 1994- 2003. Archives of Internal Medicine, 168(4): 418-424.
3. Maggioni, A.P., Orso, F., Calabria, S., et al. (2016) The real-world evidence of heart failure: findings from 41 413 patients of the ARNO database. European Journal of Heart Failure, 18(4): 402-10.
4. Barabasi, A.L. (2016) Network Science. Cambridge University Press; first edition, August 5, 2016.
5. Kolaczyk, E.D. (2009) Statistical Analysis of Network Data - Methods and Models. Springer, New York, 2009.

6. Newman, M.E.J. (2010) Networks - An Introduction. Oxford University Press, New York, 2010.
7. Van Deursen, V.M., Urso, R., Laroche, C., et al. (2014) Co-morbidities in heart failure. Heart Failure Review, 19(2): 163-72
8. Hoover, K.W., Tao, G., Kent, C.K., Aral, S.O. (2011) Epidemiologic research using administrative databases: garbage in, garbage out. Obstetrics and Gynecology, 117(3): 729; author reply 729-30.
9. Schneeweiss, S., Avorn, J. (2005) A review of uses of health care utilization databases for epidemiologic research on therapeutics. Journal of Clinical Epidemiology, 58(4): 323-37.
10. Brandes, U., Delling, D., Gaertler, M., et al. (2008) On Modularity Clustering. IEEE Transactions on Knowledge and Data Engineering, 20(2):172-188.
11. Newman, M.E.J. (2004) Fast algorithm for detecting community structure in networks. Physical review, E69(6), 066133.
12. Girvan, M., Newman, M.E.J. (2002) Community structure in social and biological networks. Proceedings of the national academy of sciences, 99(12): 7821-7826.
13. AHRQ Quality Indicators. Heart Failure Mortality Rate, Technical Specifications, version 5.0, March 2015. Available at: http://www.qualityindicators.ahrq.gov/Downloads/Modules/IQI/V50/Tech Specs/IQI_16_ Heart_Failure_Mortality_Rate.pdf. Accessed on May 19, 2015.
14. Pope GC, Kautter J, Ingber MJ., Freeman S, Sekar R, Newhart C Evaluation of the CMS-HCC Risk Adjustment Model, Final Report, Centers for Medicare and Medicaid Services, march 2011.
15. Available at:https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/ downloads/evaluation_risk_adj_model_2011.pdf, Accessed on May 19, 2015.
16. Mazzali, C., Paganoni, A.M., Ieva, F., et al. (2016) Methodological issues on the use of administrative data in healthcare research: the case of heart failure hospitalizations in Lombardy Region, 2000 to 2012. BMC Health Services Research, 16(1): 234 doi: 10.1186/s12913-016-1489
17. Gagne JJ, Glynn RJ, Avorn J, Levin R, Schneeweiss S. (2011) A combined comorbidity score predicted mortality in elderly patients better than existing scores. J Clin Epidemiol.;64:749–59.
18. Frigerio, M., Mazzali, C., Paganoni, A.M., et al. (2017) Trends in heart failure hospitalizations, patient characteristics, in-hospital and 1-year mortality: a population study, from 2000 to 2012 in Lombardy. Accepted for publications on International Journal of Cardiology. To appear
19. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, (2015). URL http://www.R-project.org/.
20. Kolaczyk, E.D., Csardi, G (2014) Statistical Analysis of Network Data with R. Springer, New York, 2014.
21. Csardi, G., Nepusz, T. (2006) The igraph software package for complex network research. Inter. Journal Complex Systems RIVEDERE REF.
22. Hidalgo, C.A., Blumm, N., Barabasi, A.L., Christakis, N.A. (2009) A Dynamic Network Approach for the Study of Human Phenotypes. PLoS Computational Biology, 5(4): e1000353.
23. Ieva, F., Jackson, C.H., Sharples, L.D. (2017, online first 2015) Multi-State modelling of repeated hospitalisation and death in patients with Heart Failure: the use of large administrative databases in clinical epidemiology. Statistical Methods in Medical Research, 26 (3): 1350-1372
24. Grossetti, F., Ieva, F., Paganoni, A.M. (2017) A Multi-state Approach to Patients Affected by Chronic Heart Failure The Value Added by Administrative Data. Health Care Management of Science. doi: 10.1007/s10729-017-9400-z
25. Holland, P.W., Blackmond Laskey, K., Leinhardt, S. (1983) Stochastic blockmodels: First steps. Social Networks, 5: 109-137.

APPENDIX A

Legend of acronyms for comorbidities

The following table reports the legend of the acronyms used for labeling networks nodes according to the comorbidity arising from the algorithm of Gagne [17]. A detailed algorithm showing the correspondence between such denominations and the underlying ICD-9-CM codes can be found at the following website:

https://scholar.harvard.edu/files/gagne/files/jjg-comorbidity-sas-program.txt

Acronym	Comorbidity coded in Gagne algorithm	Acronym	Comorbidity coded in Gagne algorithm
metastatic	Metastatic Cancer	compdiabetes	Complicated diabetes
dementia	Dementia	anemia	Deficiency anemias
renal	Renal Failure	elecrtolytes	Fluid and electrolyte disorders
wtloss	Weight loss	liver	Liver diseases
hemiplegia	Hemiplegia (stroke)	pvd	Peripheral vascular disorders
alcohol	Alcohol abuse	psycosis	Psycosis
tumor	Any tumor	pulmcirc	Pulmonary circulation disorders
arrhythmia	Cardiac Arrhythmyas	hivaids	HIV/AIDS
pulmonarydz	Chronic Pulmonary disease	hypertension	Hypertension
coagulopathy	Coagulopathy		