

University of Reading

Department of Computer Science

**Mining Co-Morbidity Patterns and Associations with Health Outcomes from an Intensive Care Unit Registry**

Leah Gourley

*Supervisor*: Yevgeniya Kovalchuk

A report submitted in partial fulfilment of the requirements of

the University of Reading for the degree of

Bachelor of Science in *Computer Science*

DATE

**Declaration**

I, Leah Gourley, of the Department of Computer Science, University of Reading, confirm that this is my own work and figures, tables, equations, code snippets, artworks, and illustrations in this report are original and have not been taken from any other person’s work, except where the works of others have been explicitly acknowledged, quoted and referenced. I understand that if failing to do so will be considered a case of plagiarism. Plagiarism is a form of academic misconduct and will be penalised accordingly.

I give consent to a copy of my report being shared with future students as an exemplar.

I give consent for my work to be made available more widely to members of UoR and public with interest in teaching, learning and research.

Leah Gourley

DATE

**Abstract**

…

**Keywords**: Python, clustering analysis, partitioning algorithms, data extraction

**Report’s total word count:** - words (excluding references and appendices)

**Acknowledgements**

Thank you to Yevgeniya for your support and guidance, and helping me quickly adapt to unexpected changes in plan. Thank you to Pat for lending me your ear and for your confidence in me. Thanks to Mum and Dad for always being a phone call away. And lastly, thanks to my housemates for not complaining about my laptop fans running at all hours of the day and night!

**Contents**

**List of Figures**

**List of Tables**

**List of Abbreviations**

CITI – Collaborative International Training Institute

CSV file – Comma-Separated Value file

HIPAA – Health Insurance Portability and Accountability Act

ICD-9 – International Classification of Diseases, ninth revision

MIMIC-III Database – Medical Information Mart for Intensive Care

PCA – Principal Component Analysis

SSE – Sum of Squared Errors

**Chapter 1**

**Introduction**

* 1. **Background**

Describe comorbidity; purpose of clustering analysis;

types of partitioning algorithms? Maybe this comes in methodology instead

* 1. **Problem statement**

Describe the approach? And the main goal?

Describe why this research needs to be done

Outline the issues that need addressing (eg how to quantify success of algorithm, why the algorithms need comparing, how to handle data)

Discuss difference between comorbidity and multimorbidity and the selection of comorbidity for this project

* 1. **Aims and objectives [162]**

The aim of this project is to develop a working program that performs clustering analysis on a large dataset. The dataset is the MIMIC-III healthcare database. The resulting clusters should be able to be analysed to consider comorbidities between different diagnoses, and their impact on health outcomes. (Has this particular area been done yet?)

The project will utilise multiple clustering algorithms, such as K-Means, mini-batch K-Means, and \_. Additionally, an interpretation of the M-algorithm, created by S. Sieranoja and P. Fränti in ‘Adapting k-means for graph clustering’, will be implemented to handle the given data as opposed to graph data, as in their original research. The code should take selected tables from the database, extract the relevant features (primary/secondary diagnoses, and outcome (death/discharge)), format it in an appropriate manner for the clustering algorithms.

The goal is to compare the results of these different algorithms using different evaluation measures, to determine the appropriate number of clusters for this data, and the optimal clustering.

* 1. **Solution Approach [113]**

Explain how the algorithms/metrics/measures were picked…

The completed program should be able to accept multiple inputs of the tables from the MIMIC III database. Utilising the pandas library in Python, it should manipulate these tables into one DataFrame or array, which will be passed to the various clustering algorithms. The algorithms will be run in a loop(s) for a pre-defined set of variable values, such as k (number of clusters) and b (batch size). After fitting each algorithm with the data, intrinsic evaluation metrics such as silhouette score, calinski-harabasz/davies-bouldin index should be calculated, along with SSE and \_. As well as this, the clusters should be plotted on a scatter plot, coloured by cluster or by event flag (diseased/discharged).

* 1. **Summary of contributions and achievements**

Implemented an adapted M-algorithm (explain how adapted) and implemented \_ measure.

* 1. **Organisation of the report**

Section 2 describes existing research in the area. In section 3, the methodology is explained. Section 4 details the results, with discussions made in Section 5, drawing to a conclusion in Section 6. An additional Appendix chapter is provided with the graphical results of each clustering algorithm for each value of k/b.

**Chapter 2**

**Literature Review**

**2.1 …**

Broad issues:

* Explain briefly (as in, don’t need to write more than a paragraph) the broad issues related to the investigation, like why even look into this in the first place? Make it easy to segway into discussions about current research.

Studies overlapping with my research area:

* Here talk about areas similar to the research area but save those final two/three studies I’ll be drawing from. Anything that helps for context or isn’t directly relevant but worth noting goes here.

Studies directly related to my investigation:

* Now, discuss directly relevant studies. This should proportionally be the largest part of the discussion.

Discussion points:

* Review current research in the area. Discuss how it is relevant to your problem but compare it to what you’ve done and pick out its weaknesses or downfalls that mean my project needs to be implemented (e.g.: no one has used ICD 9 codes, or whatever).

Ng, S. K., Holden, L., Sun, J. Identifying comorbidity patterns of health conditions via cluster analysis of pairwise concordance statistics. *Statistics in Medicine* volume 31, issue 27, p.p. 3393-3405. (Jun 19, 2012). <https://doi.org/10.1002/sim.5426>

* “Identification of comorbidity patterns of health conditions is critical for evidence-based practice to improve the prevention, treatment and health care of relevant diseases…In this paper, we propose an asymmetric version of Somers’ D statistic to provide a quantitative measure of comorbidity that accounts for co-occurrence of conditions by chance, and develop a unified clustering algorithm to identify comorbidity patterns…” in an attempt to combat unreliable existing research into qualitative/descriptive measures of comorbidity.
* “For the first issue regarding distance measures, measures of association such as odds ratios or relative risks were often used in comorbidity research to express how strongly health conditions are related to one another. However, it has been shown that odds ratios and relative risks do not adequately adjust for coincidental comorbidity by chance when non-random comorbidity exists (the next link below is the reference for this).”
* “…hierarchical clustering methods may have serious limitations including the lack of robustness, the nonuniqueness of clustering results, and complicated interpretation of the hierarchy in terms of the number of clusters. Furthermore, there is no reason for imposing the nested structure of the dendrogram and the nonoverlapping constraint for clusters of health conditions.”
* This paper presents an asymmetric Somers’ D statistic for “quantifying nonrandom comorbidity” as well as a clustering method that allows for overlapping clusters. My method will not account for overlapping clusters, this should be mentioned in the discussion.

Batstra, L., Bos, E. H., Neeleman, J. Quantifying psychiatric comorbidity: Lessons from chronic disease epidemiology. *Social Psychiatry and Psychiatric Epidemiology*, volume 27, p.p 105-111 (2002) <https://doi.org/10.1007/s001270200001>

* “This article compares measures of association and clustering.”
* “Results: Odds and risk ratios, but the former more than the latter, confound clustering with coincidental comorbidity…can express comorbidity between no more than two disorders, whilst clustering coefficients, although computationally laboursome, can capture multimorbidity of any number of disorders.”
* “Conclusion: odds and risk ratios are well suited for comorbidity research which focuses on which sets of disorders or syndromes tend to occur in combination and the implications of this for, for instance, nosological classification…the cluster coefficient is to be preferred if the interest is more aetiological, addressing for example why certain individuals are prone to multiple health problems.”

Garcia-Olmos, L., Salvador, C. H., Alberquilla, A., *et al*. Comorbidity Patterns in Patients with Chronic Diseases in General Practice. (Feb 16, 2012). <https://journals.plos.org/plosone/article/metrics?id=10.1371/journal.pone.0032141>

* “Healthcare management is oriented toward single diseases, yet multimorbidity is nevertheless the rule and there is a tendency for certain diseases to occur in clusters.”
* This study tried to seek comorbidity patterns for four diseases across 198,670 individuals aged 14 or older. They found one with a high comorbidity rate, another with a low rate, and two with intermediate rates.
* They performed clustering by dichotomizing each of their selected chronic diseases into (presence, absence) and splitting all characteristics into various numbers of categories. The analysis involved explaining all 61 categories for the 28 variables in a total of 33 dimensions, and drawing summaries from these explanations.

Ghosh, D., Cabrera, J., Adam, T. N., *et al*. Comorbidity Patterns and Its Impact on Health Outcomes: Two-Way Clustering Analysis. *IEEE Transactions on Big Data* volume 6, issue 2, p.p. 359-368. (Jun 1, 2020). doi: 10.1109/TBDATA.2016.2623323

* This paper creates the K and M algorithms, “combining model-based and weighted K-means clustering methods for characterizing and summarizing a patient’s comorbid conditions.”
* The clustering is a two-way approach to reduce dimensionality.
* Data used is adapted into a graph kind of model, with weighted connections, drawing inspiration from comorbidity networks.
* Performance of their clustering analysis is evaluated using linear regression and SVMs.
* Highlights difference between comorbidity (one or more medical/psychiatric conditions in addition to an index disease disease) and multimorbidity (co-occurrence of multiple medical/psychiatric conditions without any reference to an index disease).
* “Identifying the most common patterns of comorbidity can help in targeting specific interventions for the specific subgroups and monitoring the impact of those interventions.”
* “A widely used method for assessing a patients comorbidity is the CCI…several limitations, including equal scoring of all diagnoses without accounting for the impact of different diseases severity on patient health outcomes; ignoring potentially important relationships among diseases that might differ from their simple sum…furthermore, as numerical indices do not account for multimorbidity by chance they often require clinical judgement for gathering information on each medical condition.”
* “The model based clustering is based on correlation estimates among comorbidities that take into account the occurrence by chance of coexisting conditions, controlling for the false discovery rate, thus avoiding spurious correlation among comorbid conditions.”
* They predict outcomes from these clusters like length of hospital stay.
* Method: calculate correlations among comorbid conditions; cluster comorbidities based on these coefficients; determine patient sub-clusters by appling weighted K-means. “The result is a set of clusters; each includes patients who have similar pattern of comorbid conditions.” Then they perform logistic regression and SVM model.

Wartelle, A., Mourad-Chehade, F., Yalaoui, F., *et al*. Clustering of a Health Dataset Using Diagnosis Co-Occurrences. *Applied Sciences* volume 11, issue 5, p.p. 2373. (Mar, 2021). DOI: <http://dx.doi.org/10.3390/app11052373>

* Uses hierarchical agglomerative clustering based on multimorbidity analysis. It constructs the dendrogram based on relative risk of co-occurrences, “[detecting] the multimorbidity patterns by merging similar patient profiles according to their common diagnoses.”

Franti, P., Sieranoja, S., Wikstrom, K., Laatikainen, T. Clustering Diagnoses From 58 Million Patient Visits in Finland Between 2015 and 2018. *JMR Medical Informatics*, volume 10, issue 5, pp. ?. (Apr 4, 2022). DOI: <https://doi.org/10.2196/35422>

* This is the study utilising the K/M algorithms.
* “On the basis of the co-occurrences, we calculated the relative risk of each pair of diagnoses and clustered the data by using a graph-based clustering algorithm called the M-algorithm…”
* This study evaluated the identified clusters with a money cost “The annual cost of all clusters was 10 billion Euros, and the costliest cluster…costing 2.3 billion Euros.”

Srinivasan, K., Currim, F., Ram, S. Predicting High-Cost Patients at Point of Admission Using Network Science. *IEEE J Biomed Health Inform.* Volume 22, issue 6, p.p. 1970-1977. (Dec 13, 2017). DOI:  10.1109/JBHI.2017.2783049

* This is a data mining model for prediction. A disease co-occurrence network is made, and tree-based data mining models trained on sets of features are used to traverse the network. As well as this, the network is explored to produce HPEPP models for community formation and structural properties, which the models are also trained on.

Ng, S. K. A two-way clustering framework to identify disparities in multimorbidity patterns of mental and physical health conditions among Australians. *Statistics in Medicine*, volume 34, issue 26, p.p. 3444-3460. (May 21, 2015). DOI: <https://doi.org/10.1002/sim.6542>

* Two-way clustering model for identifying significant comorbidity clusters and disparities in multimorbidity patterns among individuals.
* It uses the binary indicator of present or not present, and “provides additional information on the heterogeneity of multimorbidity among individuals”
* Uses a “clumping” clustering algorithm (like in the first link) and mixture model-based approach.
* Uses the Somers’ D statistic, and assesses their significance using the Benjamini-Hochberg procedure. The strength of multimorbidity is given by an averaged pairwise Somers’ D statistic.

Dey, S., Simon, G., Westra, B., *et al*. Mining Interpretable and Predictive Diagnosis Codes from Multi-source Electronic Health Records, booktitle = {Proceedings of the 2014 SIAM International Conference on Data Mining (SDM)}, chapter = {}, pages = {1055-1063}, doi = {10.1137/1.9781611973440.120}

* “We aim to find the groups of ICD-9 diagnosis codes from EHRs that can predict the improvement of urinary incontinence of home health care patients…”
* A LASSO based regularised predictive model was used.
* Grouping ICD9 codes by considering knowledge from the clinical classification system (CCS) and then adding additional demographic, behavioural, psycho-social and physiological information from patients, then using the predictive model.

Cornell, J., Pugh, J., Williams, J. W. Jr., *et al*. Multimorbidity Clusters: Clustering Binary Data From Multimorbidity Clusters: Clustering Binary Data From a Large Administrative Medical Database. *Applied Multivariate Research*, 12(3), p.p. 163-182 (Jan, 2007) . DOI: DOI:[10.22329/amr.v12i3.658](http://dx.doi.org/10.22329/amr.v12i3.658)

* “Our purpose…is to describe and illustrate the application of cluster analysis to identify clinically relevant multimorbidity groups.”
* “Six clinically useful multimorbidity clusters were identified: a Metabolic Cluster, an Obesity Cluster, a Liver Cluster, a Neurovascular Cluster, a Stress Cluster and a Dual Diagnosis Cluster.”
* “A small but growing body of epidemiological research has focused specifically on multimorbidity, but typically only describes the average number of conditions, or the proportion of a population with a certain number of conditions.”
* “[Limited research into clustering of diseases] is due in part to the astronomically high number of theoretically possible combinations for any set of diseases. For example, given p diseases, there are [p(p-1)]/2 possible 2-disease pairs; [p(p-1)(p-2)]/[3\*2] possible disease triads; [p(p-1)(p-2)(p-3)]/[4\*3\*2] possible disease quartets, and so on.”
* This study talks about clustering being based on some index of proximity, which measures the closeness of two objects, defining closeness to do with multivariate distance (Euclidean distance, quantitative) or similarity coefficients (Jaccard coefficient, binary).
* The study uses agglomerative hierarchical clustering.

Vasilopoulos, T., Kotwal, A., Huisingh-Scheetz, M. J., *et al*. Comorbidity and Chronic Conditions in the National Social Life, Health and Aging Project (NSHAP), Wave 2. *The Journals of Gerontology: Series B*, volume 9, issue 2, p.p. 154-165. (Oct 29, 2014). DOI: <https://doi.org/10.1093/geronb/gbu025>

* Chronic health conditions (measured in NSHAP Wave 2) were grouped in several “health domains”.
* Two comorbidity indices were created:
  + A Modified Charlson Comorbidity Index (CCI) including conditions associated with mortality risk
  + NSHAP Comorbidity Index (NCI) including conditions from CCI and additional ones related to overall health and function.
* Most prevalent conditions were identified (hypertension, incontinence, arthritis, heart conditions, cancer, diabetes).
* “This paper details the chronic condition measures of Wave 2 of NSHAP that are the leading causes of morbidity and mortality in older Americans.”
* “An important feature of the CCI is that it assigns weighted scores to conditions based on their relationship to mortality risk. Each condition receives a score of 1, 2, 3, or 6, with higher scores assigned to conditions associated with a higher mortality…Scores are then summed to produce an index score.”

**2.2 Critique of the review**

Describe the main findings and an evaluation as a conclusion of the literature.

**2.3 Summary**

https://healthcaredelivery.cancer.gov/seermedicare/considerations/comorbidity.html - NCI Comorbidity Index Overview

**Chapter 3**

**Methodology**

**3.1 The MIMIC-III Clinical Database [415]**

The MIMIC-III Clinical Database [] is a large, free-use database containing data taken from the Beth Israel Deaconess Medical Center in Boston, MA. Data was collected between the years 2001 and 2012 for 46,520 patients and 58,976 admissions to the critical care units of the hospital.

In line with ethical guidelines set by HIPAA standards, all personal information in the database has been deidentified. This involves shifting dates (such as date of birth, date and time of admission, etc) at a random offset while preserving time of day or time of year; and removing any personally identifiable information like names, addresses and phone numbers. As such, all records in the database are stated between the years 2100 and 2200, and all patients with an age greater than 89 years appear to have an age greater than 300 years.

The database consists of 26 tables. It encompasses a wide range of data, from patient demographics, discharge/mortality information, laboratory results and reports, medications and vital signs. Among the tables are dictionary tables, denoted by the prefix ‘D\_’, which contain definitions for identifiers in the related table. For instance, the ‘DIAGNOSES\_ICD’ table has a corresponding ‘D\_DIAGNOSES\_ICD’ table containing a dictionary of all ICD9 code meanings present in the first table.

This database was selected because of its use of ICD-9 diagnostic codes [] in documenting patient diagnosis for each admission. As the problem relates to identifying patterns in diagnosis, the use of ICD-9 codes provides ease in data handling as the data has already been categorised and tokenised. Additionally, the database is provided as a collection of CSV files, meaning the data will be easy to import into Postgres and Python.

In order to access and use this data, the individual must first complete a course in HIPAA requirements, the Stage 1 Data or Specimens Only Research qualification provided by CITI []. As well as this, an individual must sign the data use agreement agreeing to data use and security standards.

Version 1.4 of the database was used, with it being the most recently released version at the time of this report. The tables within the database relevant to the problem are the ‘PATIENTS’ table, containing basic patient information such as date of birth/death, subject identifier and gender; the ‘ADMISSIONS’ table, containing a quantity of demographic information on the patient, patient and admission identifiers, and diagnosis information; and the ‘DIAGNOSES\_ICD’ table, containing a list of all diagnoses for a given admission, provided in the form of ICD-9 codes.

**3.2 Event log extraction [146]**

In order to use the data, it first needs to be cleaned to remove irrelevant columns and handle missing values. As well as this, the data needs to be adapted into a format appropriate for the clustering algorithms discussed later to handle.

The data needed includes patient identifier, admission identifier, primary diagnosis for the admission, a comma-separated list of subsequent secondary diagnoses and an event flag to indicate whether the patient was discharged or diseased at the end of the admission. Clustering analysis will be performed on diagnoses for a given admission, rather than for a given patient.

Further, the diagnoses need to be in one-to-one primary-secondary diagnosis pairs, in order to be able to perform clustering analysis. This can be achieved through use of Python’s MultiLabelBinarizer class.

Finally, the data needs to be normalised. This can be achieved through Principal Component Analysis at two dimensions.

Need to talk about why I chose to get rid of those non-numerical ICD9 codes.

**3.3 Clustering algorithms**

**3.3.1 k-Means Algorithm [28]**

Three clustering algorithms were used in order to perform a comparative clustering analysis of the data. The first algorithm selected was the k-Means algorithm, a commonly-used partitioning algorithm.

[pseudocode and explanation]

**3.3.2 M-Algorithm [138]**

Sieranoja and [Fränti](https://link.springer.com/article/10.1007/s10115-021-01623-y#auth-Pasi-Fr_nti) [] derived two algorithms from the k-Means algorithm in order to perform graph clustering; the K-algorithm with good local optimisation and better local optima than the k-Means algorithm, and the M-algorithm, which solves the K-algorithm’s tendency to get stuck on a local optimum. I have implemented an adapted form of these algorithms, as the data is not of graph format. Another change is I have adapted their cost function to instead use a relative risk. [Why?] The relative risk between two data points (diagnoses) A and B can be calculated by:

where N is the total number of diagnoses within the dataset, ∑A is the number of times diagnosis A appears in the dataset, and ∑B is the number of times diagnosis B appears in the dataset.

[pseudocode and explanation]

**3.3.3 Agglomerative algorithm**

**3.4 Evaluation metrics**

**3.4.1 Calinski-Harabasz Index**

**3.4.2 Davies-Bouldin Index**

**3.4.3 Sum of Squared Error**

**3.5 Summary**

**Chapter 4**

**Results**

**4.1 k-Means Algorithm**

**4.2 M-Algorithm**

**4.3 Agglomerative algorithm**

**4.4 Summary**

**Chapter 5**

**Discussion and Analysis**

**5.1 …**

**5.2 Significance of the findings**

**5.3 Limitations**

Cannot handle the possibility of overlapping clusters

Scalability challenge – extremely computationally expensive for large data set

**5.4 Summary**

**Chapter 6**

**Conclusions and Future Work**

**6.1 Conclusions**

**6.2 Future work**

A SVM could be trained in accordance with the cluster results, and on strings of sequential diagnoses (both primary and secondary) in order to make predictions on future health outcomes for a patient.

**Chapter 7**

**Reflection**

**References**

[https://medinform.jmir.org/2022/5/e35422 - clustering with icd10](https://medinform.jmir.org/2022/5/e35422%20-%20clustering%20with%20icd10) codes

<https://link.springer.com/article/10.1007/s10115-021-01623-y> - k algorithm and m algorithm

<https://github.com/uef-machine-learning/gclu/blob/main/graphclu.cpp> - the code from ^

<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000353> - how i could change the data points to graph nodes

[] Johnson, A., Pollard, T., Shen, L. *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* **3**, 160035 (2016). <https://doi.org/10.1038/sdata.2016.35>

[] Sieranoja, S., Fränti, P. Adapting *k*-means for graph clustering. *Knowl Inf Syst* **64**, 115–142 (2022). https://doi.org/10.1007/s10115-021-01623-y

[] K. J. Wong (Dec 9, 2022). ‘7 Evaluation Metrics for Clustering Algorithms’. *Towards Data Science*, [Online]. Available: <https://towardsdatascience.com/7-evaluation-metrics-for-clustering-algorithms-bdc537ff54d2>