

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**: …

**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**
  2. **Problem statement**
  3. **Aims and objectives**
  4. **Solution Approach**
  5. **Summary of contributions and achievements**
  6. **Organisation of the report**

**Chapter 2**

**Literature Review**

**2.1 …**

**2.2 Critique of the review**

**2.3 Summary**

**Chapter 3**

**Methodology**

**3.1 The MIMIC-III Clinical Database**

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**

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.

**3.3 Clustering algorithms**

**3.3.1 k-Means Algorithm**

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**

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 Silhouette Score**

**3.4.2 Calinski-Harabasz Index**

**3.4.3 Sum of Squared Error**

**3.5 Summary**

**Chapter 4**

**Results**

**4.1 k-Means Algorithm**

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of clusters** | **Silhouette score** | **Calinski-Harabasz Index** | **SSE** |
| 4 | 0.79056 | 126,935.55987 | 2285.80086 |
| 6 | 0.80718 | 182,846.68838 | 1021.75360 |
| 8 | 0.82795 | 239,103.41295 | 571.66082 |
| 10 | 0.83472 | 296,118.63822 | 362.93479 |
| 20 | 0.88428 | 502,508.73482 | 102.66046 |
| 40 | 0.89442 | 799,498.38638 | 31.55203 |
| 60 | 0.91122 | 958,614.13696 | 17.39238 |
| 80 | 0.92990 | 1,161,406.22648 | 10.72061 |
| 100 | 0.93601 | 1,313,088.07549 | 7.56483 |
| 150 | 0.95509 | 1,911,284.88200 | 3.45045 |
| 200 | 0.96455 | 2,623,582.10237 | 1.88037 |
| 250 | 0.97282 | 3,526,287.91801 | 1.11701 |
| 300 | 0.97832 | 4,812,015.48296 | 0.68101 |
| 400 | 0.98581 | 10,145,911.96694 | 0.24156 |
| 500 | 0.99101 | 23,774,438.60676 | 0.08227 |
| 600 | 0.99323 | 62,929,571.14199 | 0.02583 |

**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**

**5.4 Summary**

**Chapter 6**

**Conclusions and Future Work**

**6.1 Conclusions**

**6.2 Future work**

**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