

1 Introduction

Diabetes is a fast-growing problem, rising from 108 million cases in 1980 to 422 million cases in 2014, and is the 9th leading cause of death in the world. In Singapore, 1 in 3 people are at risk of developing diabetes in their lifetime, and PM Lee declared a war on Diabetes in the 2017 National Day Rally.

The purpose of our project is to analyse diabetes health records to create statistics and rules of thumb which clinicians and patients can use to predict outcomes.

We chose the scope of records we analyse to be the UCI Machine Learning Repository Data Set as it provided many different fields of data that accurately captured the entire process from entering hospital to diagnosis to procedures to discharge. From there we chose to focus on the diagnoses corresponding to Circulatory, Digestive, and Diabetes, as they are 3 of the most common diagnosis types in the dataset and account for over half of all diabetes cases.

Our objective would be to find out which diagnosis codes or combinations of diagnosis codes correlate strongly to each other, and which ones lead to a deterioration in patient condition.

2 Description of Dataset

We use the dataset from the UCI Machine Learning Repository [1]. Our analysis is confined to diagnosis codes and discharge codes. The dataset has 101793 rows, representing 101793 seperate hospital admissions across non-distinct patients.

2.1 Diagnosis codes

In this report, there are 916 diagnosis codes in total. The diagnosis codes are in ICD-9. For over 99% of diagnoses, all 3 diagnosis codes are listed.

In our database, there are three types of diagnosis, namely Primary diagnosis, Secondary diagnosis and Additional secondary diagnosis.

2.2 Discharge codes

Based on the dataset, 30 discharge codes were attached to each of the patients.

discharge_disposition_id	description
1	Discharged to home
2	Discharged/transferred to another short term hospital
3	Discharged/transferred to SNF
4	Discharged/transferred to ICF
5	Discharged/transferred to another type of inpatient care institution
6	Discharged/transferred to home with home health service
7	Left AMA

Figure 1: The first 7 of the 30 diagnosis codes assigned in the dataset

We categorised these into 8 categories based on the changes they represented to the patient's lifestyle and the intensity of care provided, with their relative frequencies shown below:

Category	Name	Diagnosis Codes	Relative Percentage
1	Sent Home	1	59.2
2	Inpatient	2,5,9,23,27,28,29	3.8
3	Home Care or Nursing	6,8,22,24	14.8
4	Death	11,19,20,21	1.6
5	Hospital Monitoring	12,16,17	<0.1
6	Palliative Care	13,14	0.8
7	Specialised and Intensive Care	3,4,15	14.6
8	Ignored	7,10,18,25,26	5.2

Figure 2: Categories of discharge codes, with percentage occurrence

In particular some codes like "7: Discharged against Medical Advice" are ignored due to lack of specificity, while "10: Neonate discharged for neonatal aftercare" is ignored as it is specific to only infant diabetes, and the dataset only has 6 such encounters out of 101793.

3 Graph Model

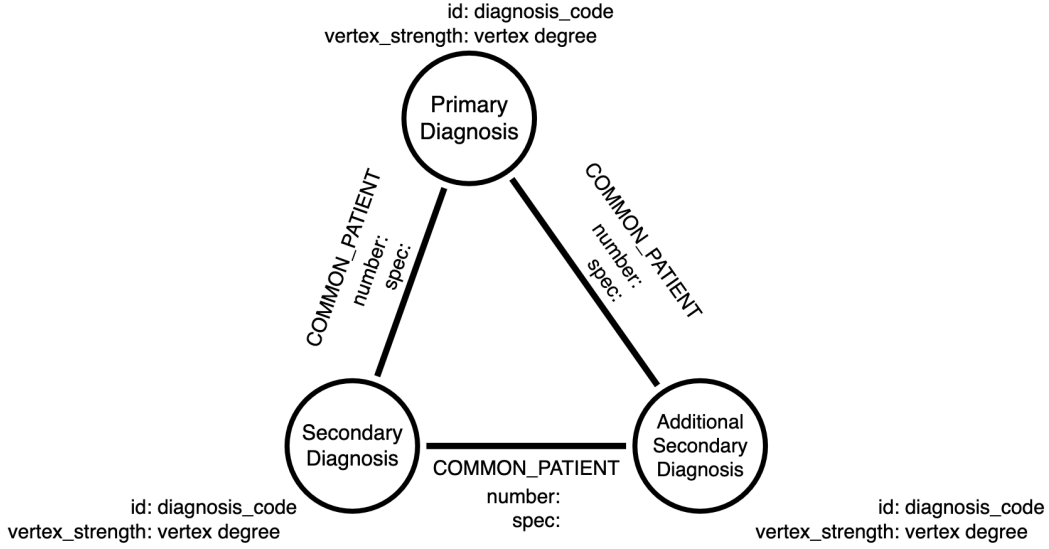


Figure 3: Categories of discharge codes, with percentage occurrence

3.1 Nodes

Diagnosis codes are represented by nodes. Note that the same node is used for the same diagnosis code, regardless if the diagnosis code happens to be a primary diagnosis, secondary diagnosis or additional secondary diagnosis for a particular encounter.

3.2 Relationships

When a patient is given 2 separate diagnosis codes, the 2 codes are linked by an undirected relationship between them (except when measuring degree centrality, which will be discussed in section 4.1). The weight of a relationship (property of "number") is equal to the number of encounters which the 2 diagnosis codes co-occur.

3.3 Discharge codes

Before the creation of the graph model, a set of diagnosis codes S_D (a combination of the 8 categories in Figure 2) is deemed "of interest". The relationship property "spec" counts the number of encounters where diagnosis codes A and B co-occur and where the patient was discharged with a discharge code in S_D .

3.4 Vertex Degree

Vertex degree is denoted as a graph property called "vertex_degree". The graph follows a power-law distribution of degree (in other words, $P(k) \propto k^{-\gamma}$, where $\gamma = 1.75$).

The most common diagnosis codes are: 250 [35722 encounters, diabetes mellitus], 428 [35562 encounters, congestive heart failure], 276 [26775 encounters, disorders of fluid electrolyte and acid-base balance].

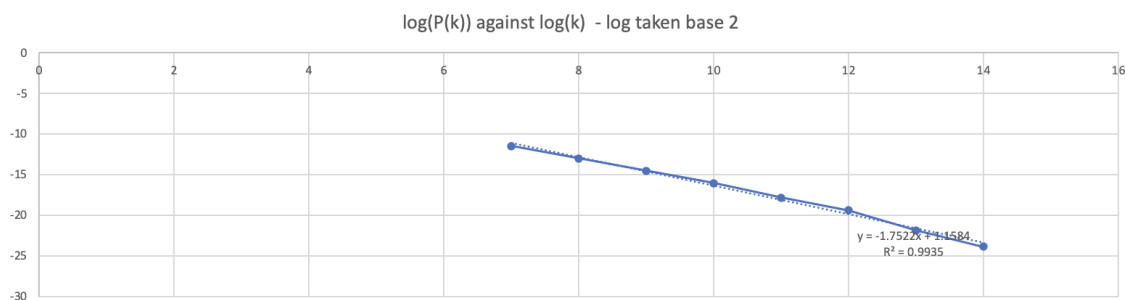


Figure 4: This is a log-log plot of $P(k)$ against k , with vertices with degree less than $2^7 = 128$ (representing a frequency of less than 0.1% of Encounters) ignored, for a lack of computational and clinical significance.

4 Centrality analysis

4.1 Degree centrality

The degree centrality algorithm can be used to find popular nodes within a graph. Degree centrality measures the number of incoming or outgoing (or both) relationships from a node. The higher the degree of a node, the more important it is in the graph.

There are two types of degree centrality: In-degree centrality and Out-degree centrality. In-degree refers to the number of edges/connections incident on a certain node. Out-degree refers to the number of edges/connections from a certain node to other nodes.

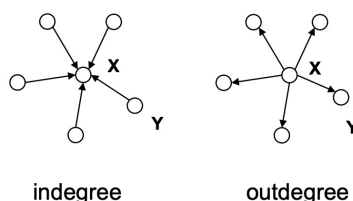


Figure 5: In-degree centrality and Out-degree centrality

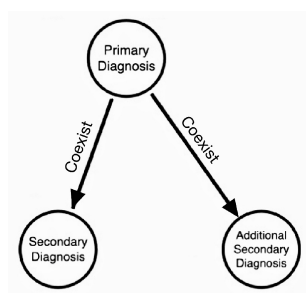


Figure 6: graph model of out-degree algorithm

Primary diagnosis describes the diagnosis that was the most serious during the hospitalisation, which is also the condition that occasioned the patient to the hospital. Secondary and Additional secondary diagnosis describe those conditions that coexist at the time of the patient, or develop subsequently. Our research uses directed graph to calculate out-degree of each diagnosis nodes, measuring the most important diagnosis which brings other coexisting diagnosis.

According to the result, neoplasm diagnosis(ICD code 682, 780, 276), circulatory diagnosis(ICD code 428, 427, 414), and respiratory diagnosis(ICD code 486, 786, 491) are the most important diagnoses that bring the patients to the hospital.

name	followers
682	331
780	325
428	314
486	306
786	304
276	301
427	286
414	284
491	261
996	254

Figure 7: Ten diagnosis with highest out-degree centrality score

4.2 Betweenness centrality

Betweenness centrality calculates the shortest paths between all pairs of nodes in the graph. Each node receives a score, based on the shortest paths that pass through the node. Nodes that more frequently lie on shortest paths between other nodes will have higher betweenness centrality scores.

$$B_{(i)} = \sum_{a,b} \frac{g_{aib}}{g_{ab}}$$

a,b is any pair of nodes in the graph

g_{aib} is the number of shortest paths from node 'a' to 'b' passing through 'i'

g_{ab} is the number of shortest paths from node 'a' to 'b'

name	score
250	71774.483
276	48506.1813
401	39610.9298
428	34054.1707
427	30196.8309
780	29964.8982
682	25079.5526
599	24947.0633
250.01	22987.0975
250.02	22593.5168

Figure 8: Ten diagnosis with highest betweenness centrality score

According to the result, diabetes(ICD code 250, 250.01, 250.02), neoplasm diagnosis(ICD code 276, 780, 682), circulatory diagnosis(ICD code 401, 428, 427), and genitourinary diagnosis(ICD code 599) influence the whole network most, suggesting their high connectedness to other diagnoses.

4.3 Relationship between betweenness centrality and out-degree centrality

From this figure, we can conclude that betweenness centrality and out-degree centrality are highly correlated. Diagnoses with higher betweenness centrality scores are more likely to have higher out-degree centrality, suggesting that the more important diagnoses are always more connected to the other diagnoses.

However, there are still exceptions. Diagnoses with high betweenness centrality but relatively low out-degree centrality have few crucial connections for the whole network flow. Diagnoses with low betweenness centrality but relatively high out-degree centrality are important but have relatively redundant connections.

5 Clustering analysis

Clustering is clinically relevant for several reasons: It informs patients about conditions they are at risk at, so patients can perform targeted self-management; It informs clinicians about commonly coexisting conditions, improving

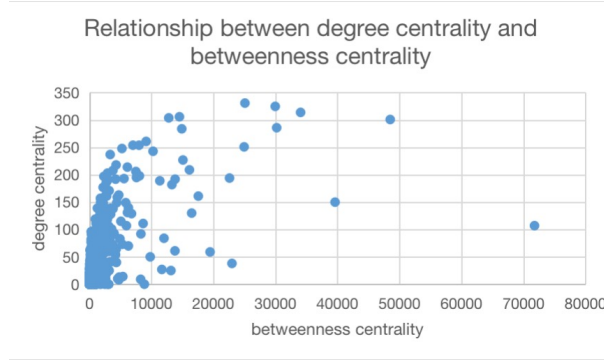


Figure 9: Ten diagnosis with highest betweenness centrality score

healthcare resource allocation; It allows researchers to connect disparate conditions to a common cause. Clustering coefficient is used as a metric to measure the degree which nodes tend to cluster together

5.1 Definition of clustering coefficient

On unweighted graphs, the local clustering coefficient of a node i has a single definition, which is the probability that the nodes in the neighbourhood of i are linked together. c_i has a minimum of 0 when its neighbours have no interconnections, and has a maximum of 1 if its neighbourhood subgraph is a clique. It is defined as $c_i = \frac{n_i}{\pi_i}$ where n_i is the number of triangles in the graph containing vertex i , or $\frac{1}{2} \sum a_{ij} a_{jk} a_{ki}$ where a_{ij} is 1 when the edge is present and 0 when it is absent, and π_i is the theoretical maximum number of triangles that could contain vertex i , equal to the maximum possible number of connections between neighbours of i , which is $\frac{1}{2}(w_i)(w_i - 1)$ if i has w_i neighbours. However our graph model is a weighted graph, and the definition of clustering coefficient for weighted graphs lacks an academic consensus.

We use the definition of clustering coefficient for weighted graphs in [2] due to its ability to handle weighted graphs present in biology and exhibit stable behaviour when low-weighted vertices are added (of which there are many in this graph)[3].

Under this definition, Clustering Coefficient $c_i = \frac{n_i}{\pi_i}$ where $n_i = \sum a_{ij} a_{jk} a_{ki}$ across all vertices $j, k \neq i$, and $\pi_i = K((\sum a_{ij})^2 - \sum a_{ij}^2)$. K is a proportionality constant equal to the greatest a_{ij} across all i, j . If we were to substitute $b_{ij} = a_{ij}/K$, we could formulate c_i in terms of b_{ij} without K in the equation, and The value of K also ensures $c_i \leq 1$ and that c_i does not change with scaling.

5.2 Null hypothesis

We define our null hypothesis under the assumption that diagnosis codes behave as independent outcomes, that is, two diagnoses D_1 and D_2 co-occur with a frequency proportional to $(F_1 F_2)$, where F_i is the degree of diagnosis code D_i . Note this assumption allows us to convert all mentions of a_{ij} into an expression in $F(i)$ and $F(j)$.

Ignoring K we can set the weight of D_i (hereafter called P_i) to be $F_i/101793$, which gives the relative frequency of D_i across all diagnoses. This normalisation allow us to set $a_{ij} = P_i P_j$.

$$n_i = \frac{1}{2} \sum a_{ij} a_{jk} a_{ki} = \frac{1}{2} ((P_i)^2) (\sum_{j \neq i} (P_j^2 (\sum_{k \neq j, i} P_k^2))) = \frac{1}{2} ((P_i)^2) [(S_2 - P_i^2)^2 - (S_4 - P_i^4)]$$

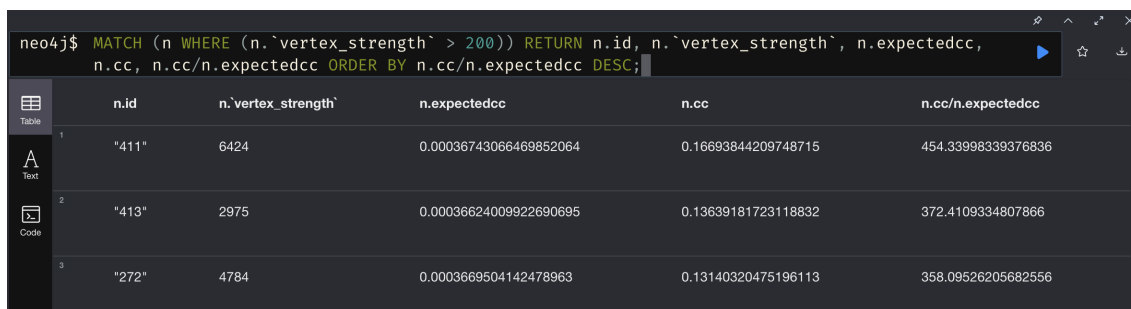
$$\pi_i = K((\sum a_{ij})^2 - \sum a_{ij}^2) = K(\frac{1}{2} ((P_i)^2) [(S_1 - P_i)^2] - S_2 + P_i^2)$$

where $K = P_{largest} P_{second-largest} = 0.175464 * 0.174678$ and $S_i = \sum P_j^i$ across all j .

The formula for Clustering Coefficient can be run in linear time, we find that $c_{(i_{expected})}$ a decreasing function, but remains roughly constant at 3.65×10^{-4} .

5.3 Test against data

We detect highly clustered vertices by computing the ratio of expected clustering coefficients against the real ones. For all vertices with degree > 200 , $\frac{c_i}{c_{i_{expected}}}$ is between 10 and 460, strongly suggesting that edges do not follow an independent distribution, and that the graph contains modular structure. The 5 diagnosis codes with the highest $\frac{c_i}{c_{i_{expected}}}$ ratio are 411, 413, 272, 410 and 412. Of particular interest is the fact that 410-413 all correspond to ischemic heart disease. 272 corresponds to disorders of lipid metabolism.



```
neo4j$ MATCH (n WHERE (n.`vertex_strength` > 200)) RETURN n.id, n.`vertex_strength`, n.expectedcc, n.cc, n.cc/n.expectedcc ORDER BY n.cc/n.expectedcc DESC;
```

	n.id	n.`vertex_strength`	n.expectedcc	n.cc	n.cc/n.expectedcc
1	"411"	6424	0.00036743066469852064	0.16693844209748715	454.33998339376836
2	"413"	2975	0.00036624009922690695	0.13639181723118832	372.4109334807866
3	"272"	4784	0.0003669504142478963	0.13140320475196113	358.09526205682556

Figure 10: Real Clustering Coefficient vs Expected Clustering Coefficient

We shall now do a comparison with the discharge code for one of them - compare death rate across discharge codes, and then comment on the cluster.

5.4 Analysis of neighbourhood of ischemic heart disease diagnosis codes

Diagnosis code 411, which has the highest clustering coefficient, has 6424 (6.4%) different co-occurrences. Isolating the nodes in the set of nodes S_{411} , where S_{411} is the union of neighbourhood of Diagnosis Code 411 and Diagnosis code 411 itself, we can query those nodes separately to find out the prevalence discharge types within the cluster.

The total weight within S_{411} is 240801, almost 80% of the total weight. This shows that ischemic heart disease commonly coexists with a majority of the common diabetes complications. The percentage of escalations to specialised healthcare (Category 7 diagnosis) or expired/sent to palliative care in S_{411} is 14.3% and 2.4% respectively, roughly equal to the corresponding overall statistics. Similar results are seen for diagnosis codes 410-413.

This is in line with knowledge that the pathophysiology of ischemic heart disease in diabetes patients is significantly researched but not fully understood [4]. However, the high clustering coefficient may suggest that the risk of ischemic heart disease is the result of a lifestyle that causes diabetes (common cause), or that it is a consequence from conditions in other organs (common effect).

6 Analysis based on Discharge Codes



```
neo4j$ MATCH (n:Diagnosis) WHERE n.`vertex_strength` > 100 RETURN toFloat(n.spectotal)/toFloat(n.`vertex_strength`), toFloat(n.cc)/n.expectedcc, n.id, n.`vertex_strength`, n.spectotal ORDER BY toFloat(n.spectotal)/n.`vertex_strength` DESC;
```

	toFloat(n.spectotal)/toFloat(n.`vertex_strength`)	toFloat(n.cc)/n.expectedcc	n.id	n.`vertex_strength`	n.spectotal
1	0.6017699115044248	91.62830302679174	"821"	339	204
2	0.5891472868217055	97.18871324447335	"808"	387	228
3	0.5483449477351916	118.61395058227683	"820"	2296	1259

Figure 11: Diagnosis codes with the largest fraction of encounters referred to specialised care

As shown, diagnosis codes with the highest rates of referral to specialised care are 821, 808, 820, E885 and 290. In particular, the first 3 correspond to bone fractures, and E885 corresponds to injuries from falls. These 5 diagnosis

codes have average clustering coefficients compared to nodes with similar vertex strength, which is in line with the knowledge that physical injuries are not particularly associated with any cluster of medical diagnoses, but rather are generally more serious for diabetes patients as a whole.

The diagnosis codes with the highest rates of death/transferral to palliative care are brain conditions (431, 199, 348) and lung conditions (507, 197).

7 Discussion

Diabetes is a serious disease that leads to the deterioration of many of the body's systems over time. By modelling diagnosis codes as nodes and their co-occurrences as relationships, we represented the cumulative diagnosis statistics of over 100,000 patients, analysing them by applying algorithms from network science.

By investigating centrality, we realise that diagnosis of neoplasms, circulatory disease, and respiratory diseases carry the most weightage in causing patients to go to hospital, where they are found to be diabetic. The former two, along with diabetes, are the diagnoses with highest betweenness centrality. The high clustering coefficient of diagnosis nodes suggests that our graph model is highly modular, and that symptoms of diabetes come in clusters. Ischemic heart disease, in particular is well-associated with other secondary ailments. Analysing diagnosis codes, we found that falls are associated with high rates of being referred to specialised healthcare.

We have left investigating the effects of age, race, lab procedures and medication on risk factors of diabetes to future work. Another limitation of this project is our dataset only contained patients who were confirmed as diabetic. We are unable to examine situations where the patient's conditions related to diabetes affect their quality of life, but have not been admitted to hospital.

Graph data models, while well-established in gene expression networks, are not often applied to analyse disease complications. We believe that graph databases, in particular because of how they lend themselves well to analysis of clustering and centrality, can shed more light on causes and risk factors of diabetes-related complications.

References

- [1] Clore, J., Cios, K. J., & DeShazo, J. (2014, March 5). Diabetes 130-US hospitals for years 1999-2008 Data Set. Retrieved July 30, 2022, from <https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008>.
- [2] Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4, Article17. <https://doi.org/10.2202/1544-6115.1128>
- [3] Fardet, T. ; Levina, A. (2021, November 17). Weighted directed clustering: Interpretations and requirements for heterogeneous, inferred, and measured networks. *Physical Review Research*. Retrieved December 21, 2022, from <https://link.aps.org/doi/10.1103/PhysRevResearch.3.043124>
- [4] Severino, P., D'Amato, A., Netti, L., Pucci, M., De Marchis, M., Palmirotta, R., Volterrani, M., Mancone, M., & Fedele, F. (2018). Diabetes Mellitus and Ischemic Heart Disease: The Role of Ion Channels. *International journal of molecular sciences*, 19(3), 802. <https://doi.org/10.3390/ijms19030802>
- [5] Valente, T., Coronges, K., Lakon, C., Costenbader, E. (2008). How Correlated Are Network Centrality Measures?. *Connections (Toronto, Ont.)*, 28(1), 16. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/article>
- [6] Disney, A. Disney, A. (2020). Social network analysis: Understanding centrality measures. Retrieved 4 January 2023, from <https://cambridge-intelligence.com/keylines-faqs-social-network-analysis/>
- [7] Graph Analytics—Introduction and Concepts of Centrality. (2019). Retrieved 4 January 2023, from <https://towardsdatascience.com/graph-analytics-introduction-and-concepts-of-centrality-8f5543b55de3>
- [8] List of ICD-9 codes 390–459: diseases of the circulatory system - Wikipedia. (2022). from https://en.wikipedia.org/wiki/List_of_ICD-9_codes_390