A replication of “Multimorbidity states associated with higher mortality rates in organ dysfunction and sepsis: a data-driven analysis in critical care” by Zador et al.

Diego Zavalza dzavalza@ucsd.edu
Asif Mahdin amahdin@ucsd.edu
Colin Tran ctt005@ucsd.edu
Zhuji Zhang zhz044@ucsd.edu

Mentor: Professor Kyle M. Shannon
kshannon@ucsd.edu

Abstract

In this paper, we are trying to replicate the results from the paper “Multimorbidity states associated with higher mortality rates in organ dysfunction and sepsis: a data-driven analysis in critical care” by Zador et al., where they focused on the issue that no consideration were given to the medical background of sepsis patients. They hypothesized the existence of patient subgroups in critical care with distinct multimorbidity states and certain multimorbidity states are associated with higher rates of organ failure, sepsis, and mortality co-occurring with these clinical problems. The primary objective of this replication study is to rigorously verify and validate the findings of the original research conducted by Zador et al. on multimorbidity states and their association with higher mortality rates in organ dysfunction and sepsis in a critical care setting. Through this replication, we aim to confirm the robustness and reproducibility of the original study’s conclusions, thereby strengthening the evidence base surrounding the impact of multimorbidity on critical care outcomes.

Code: https://github.com/kshannon-ucsd/ucsd-dsc180ab-team1

1 Introduction

Sepsis is one of the most complicated and emergency medical conditions in the world, and the diagnosis for the disease needs to be timely in order to reduce the mortality rate. However, due to its heterogeneity and unspecific nature, along with the countless possible combinations of morbidity, most of the warning systems have low predictive values and are often subject to inappropriate treatments (Ifedayo Kuye 2018).

In the paper by Zador et al, they tried to tackle the problem by through identifying patient groups from the seemingly cluttered patient history data. By identifying unique patient groups, treatment plans can be made more efficiently and accurately, and the findings can also be beneficial to future patients.

We are using a dataset called MIMIC-III (Johnson et al. 2016) from PhysioNet (PhysioBank 2000). It is a relational database consisting of 26 tables that can be linked by identifiers, comprising deidentified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. It is an extremely useful dataset when it comes to analysis since it includes information such as demographics, vital sign measurements made at the bedside, laboratory test results, procedures, medications, caregiver notes, imaging reports, diagnoses (represented by ICD-9 codes) and mortality (including post-hospital discharge). We will be using some of these variable in our analysis (Johnson, Pollard and Mark III 2016).

2 Methods

2.1 Cohort selection

Our initial approach with cohort selection used the filter provided in the paper by Zador et al, which included two key criteria: the patients where age is 16 and over and also the first time admissions for each of them. However, our first endeavor proved to be unsuccessful. Even though we did come close to the cohort in the paper in terms of number, the distribution of age groups is significantly off when compared with the supplementary table provided in the paper. Our second approach is more successful after we located the Github repository (Johnson et al. 2018) from the contributors of the MIMIC-III data, who provided a detailed guidance on the cohort selection method, including formulas for calculating age and retrieving the first admission.

2.2 Clustering and Latent Class Analysis

The paper started the analysis with k-means clustering, and that naturally became our first step too. We calculated the disease similarities, which is based on the prevalence of diseases in the population, within different age groups and employed Euclidean distance as a measure of similarity to establish cluster relationships just like the paper.

For the subsequent exploration of patient subgroups, we utilized latent class analysis (LCA), incorporating age, admission type (elective vs. non-elective), and morbidity composition using 30 Elixhauser categories. This method assumes the existence of unobserved (”latent”) subgroups within the study cohort and identifies them by fitting a series of mixture models to the data. The optimal number of subgroups was chosen based on a combination of achieving the lowest Bayesian information criteria (BIC) and Akaike information criteria (AIC), with the additional criterion that subgroup size should not be smaller than 5% of the entire study cohort. This approach corresponds with the method used in the paper we are trying to replicate.

To compare the characteristics of the latent subgroups, we employed the chi-square test for categorical variables and one-way ANOVA for continuous variables. Residual diagnostics were applied to ensure that ANOVA assumptions were not violated, and expected values were calculated to confirm that the chi-square test assumptions were upheld.

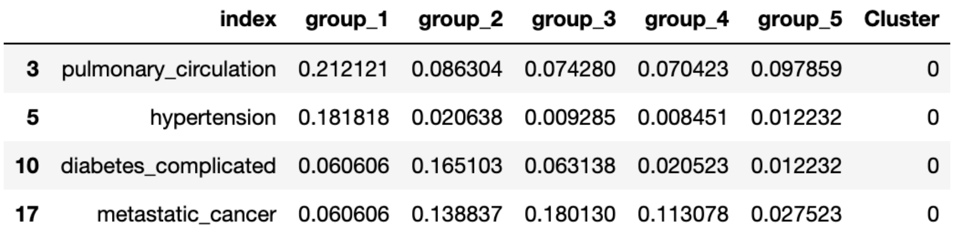
The methods described by this section all followed the procedures used in the paper (Zador et al. 2019).

3 Results

3.1 K-Means

Using our results from LCA, we can use these subgroups and visualize clusters that will inform us of any patterns we may see between the age groups and the Elixhauser categories. After combining all of our subgroups data into a single DataFrame and adding a column to determine the age group, we were able to pass it into our K-Means class. The result was three different clusters, illnesses that were more prevalent within younger age groups over older age groups, illnesses that were more prevalent in older age groups over younger age groups, and illnesses that had a generally low prevalence throughout all age groups.

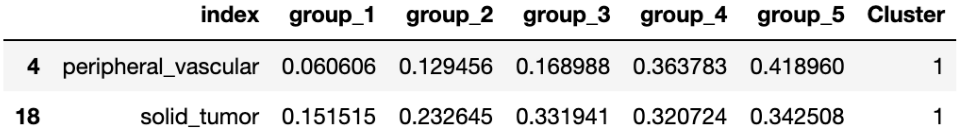
In our first cluster, we got the data of all illnesses that were generally more prevalent in younger age groups. This cluster consisted of Pulmonary Circulation, Hypertension, Diabetes Complicated, and Metastatic Cancer. The results are shown in the table below:



	index	group_1	group_2	group_3	group_4	group_5	Cluster
3	pulmonary_circulation	0.212121	0.086304	0.074280	0.070423	0.097859	0
5	hypertension	0.181818	0.020638	0.009285	0.008451	0.012232	0
10	diabetes_complicated	0.060606	0.165103	0.063138	0.020523	0.012232	0
17	metastatic_cancer	0.060606	0.138837	0.180130	0.113078	0.027523	0

Figure 1: Illnesses More Prevalent in Younger Age Groups

Similarly, in our second cluster, we got the data of illnesses that were more prevalent in older patients than younger patients. We gathered that Peripheral Vascular and Solid Tumors were much more prevalent in older age groups. The results are shown in the table below:



	index	group_1	group_2	group_3	group_4	group_5	Cluster
4	peripheral_vascular	0.060606	0.129456	0.168988	0.363783	0.418960	1
18	solid_tumor	0.151515	0.232645	0.331941	0.320724	0.342508	1

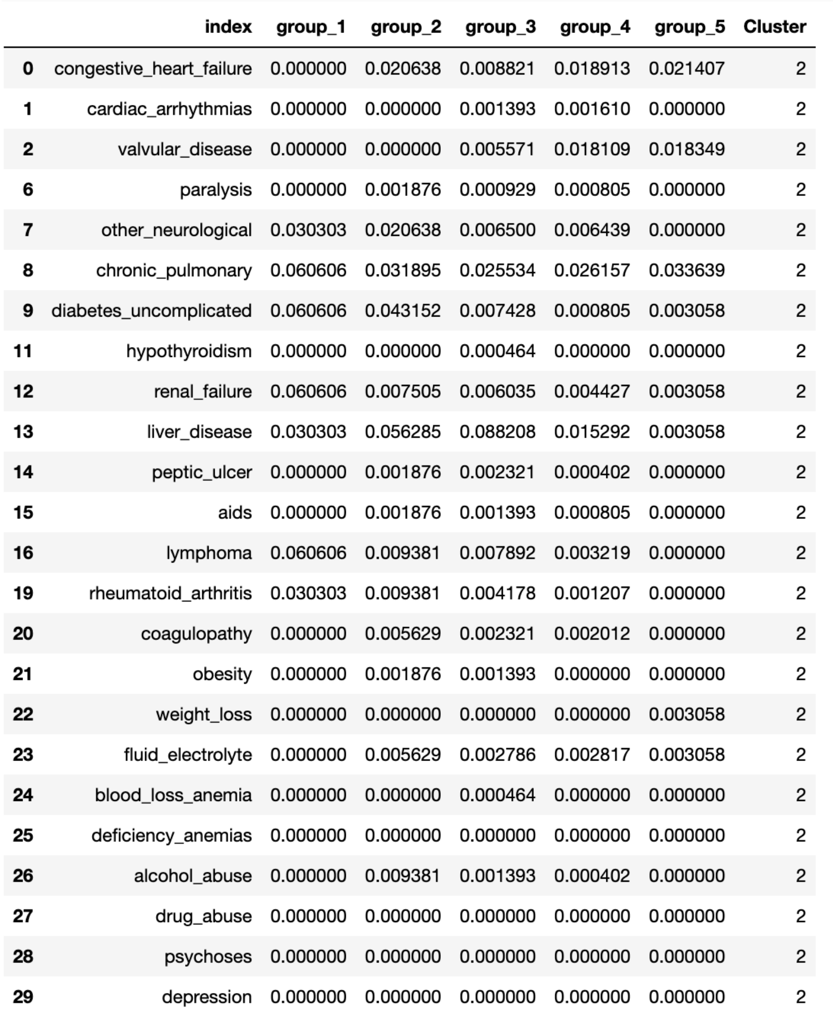
Figure 2: Illnesses More Prevalent in Older Age Groups

In the third cluster, we got the data of illnesses that had a low prevalence throughout all age groups. This cluster consisted of Congestive Heart Failure, Cardiac Arrhythmias, Valvular Disease, Paralysis, Other Neurological, Chronic Pulmonary, Diabetes Uncomplicated, Hypothyroidism, Renal Failure, Liver Disease, Peptic Ulcer, Aids, Lymphoma, Rheumatoid Arthritis, Coagulopy, Obesity, Weight Loss, Fluid Electrolyte, Blood Loss Anemia, Deficiency Anemias, Alcohol Abuse, Drug Abuse, Psychoses, and Depression. The results are shown on page 5 (Figure 3).

3.2 LCA

Our efforts to replicate the original findings revealed a notable deviation despite the robustness of our applied methodologies. We began with a preliminary analysis using k-means clustering to evaluate disease similarities, considering prevalence across various age brackets and utilizing Euclidean distance for cluster similarity assessments. This was followed by a deeper examination through Latent Class Analysis (LCA), where we factored in variables like age, admission type, and morbidity composition. Our analysis identified six distinct subgroups within the study cohort, generally aligning with the original study's conclusions. However, our replication highlighted a divergence from the original study: we found that two of these subgroups fell under the 5% threshold, contrary to the original findings. This discrepancy, while within a reasonable margin, underscores the intricate and sometimes unpredictable nature of data-driven research in medical studies. It also emphasizes the necessity for ongoing scrutiny and adaptation in the application of machine learning in healthcare, as even well-established methodologies can yield varied results under different circumstances.

Subgroup	Count	Percentage
1	7422	7.19
2	2188	2.12
3	9262	8.98
4	12861	12.47
5	4736	4.59
6	66703	64.65



	index	group_1	group_2	group_3	group_4	group_5	Cluster
0	congestive_heart_failure	0.000000	0.020638	0.008821	0.018913	0.021407	2
1	cardiac_arrhythmias	0.000000	0.000000	0.001393	0.001610	0.000000	2
2	valvular_disease	0.000000	0.000000	0.005571	0.018109	0.018349	2
6	paralysis	0.000000	0.001876	0.000929	0.000805	0.000000	2
7	other_neurological	0.030303	0.020638	0.006500	0.006439	0.000000	2
8	chronic_pulmonary	0.060606	0.031895	0.025534	0.026157	0.033639	2
9	diabetes_uncomplicated	0.060606	0.043152	0.007428	0.000805	0.003058	2
11	hypothyroidism	0.000000	0.000000	0.000464	0.000000	0.000000	2
12	renal_failure	0.060606	0.007505	0.006035	0.004427	0.003058	2
13	liver_disease	0.030303	0.056285	0.088208	0.015292	0.003058	2
14	peptic_ulcer	0.000000	0.001876	0.002321	0.000402	0.000000	2
15	aids	0.000000	0.001876	0.001393	0.000805	0.000000	2
16	lymphoma	0.060606	0.009381	0.007892	0.003219	0.000000	2
19	rheumatoid_arthritis	0.030303	0.009381	0.004178	0.001207	0.000000	2
20	coagulopathy	0.000000	0.005629	0.002321	0.002012	0.000000	2
21	obesity	0.000000	0.001876	0.001393	0.000000	0.000000	2
22	weight_loss	0.000000	0.000000	0.000000	0.000000	0.003058	2
23	fluid_electrolyte	0.000000	0.005629	0.002786	0.002817	0.003058	2
24	blood_loss_anemia	0.000000	0.000000	0.000464	0.000000	0.000000	2
25	deficiency_anemias	0.000000	0.000000	0.000000	0.000000	0.000000	2
26	alcohol_abuse	0.000000	0.009381	0.001393	0.000402	0.000000	2
27	drug_abuse	0.000000	0.000000	0.000000	0.000000	0.000000	2
28	psychoses	0.000000	0.000000	0.000000	0.000000	0.000000	2
29	depression	0.000000	0.000000	0.000000	0.000000	0.000000	2

Figure 3: Illnesses With Low Prevalence Throughout All Age Groups

4 Discussion

4.1 K-Means

While our results from K-Means do show us correct separation into the same three categorical groups, they are not an exact match to the paper we were trying to replicate. For instance, in the paper, they obtained that the illnesses that were more prevalent for younger age groups than older age groups were other neurological disorders, coagulation, depression, liver disease, alcohol abuse, and drug abuse. While none of them were in our cluster, this is because our results from LCA may have not given us the same subgroups as the papers results from LCA. Therefore, this throws off the results found in the cluster. This does not mean that our results are incorrect, it just means that for our subgroups found from LCA as mentioned above. These are the clusters we received. Furthermore, by looking at the prevalence of each illness you'll actually see that they are categorized correctly.

4.2 LCA

Our analysis of Latent Class Analysis (LCA) outcomes differs from those presented in the referenced paper due to the lack of access to the original study's code and methodology for handling missing data. This limitation highlights the significant impact that data cleaning and preprocessing have on research outcomes, especially in LCA. Our findings emphasize the importance of transparent and detailed data preprocessing methods in research to ensure reproducibility and accurate interpretation of results

References

Ifedayo Kuye, Chanu Rhee. 2018. “Spotlight: Overdiagnosis and Delay: Challenges in Sepsis Diagnosis.” Oct. [Link]

Johnson, A, T Pollard, and R Mark III. 2016. “MIMIC-III Clinical Database (version 1.4). PhysioNet. 2016.”

Johnson, Alistair EW, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. “MIMIC-III, a freely accessible critical care database.” *Scientific data* 3 (1): 1–9

Johnson, Alistair EW, David J Stone, Leo A Celi, and Tom J Pollard. 2018. “The MIMIC Code Repository: enabling reproducibility in critical care research.” *Journal of the American Medical Informatics Association* 25 (1): 32–39

PhysioBank, PhysioToolkit. 2000. “Physionet: components of a new research resource for complex physiologic signals.” *Circulation* 101 (23): e215–e220

Zador, Zsolt, Alexander Landry, Michael D Cusimano, and Nophar Geifman. 2019. “Multimorbidity states associated with higher mortality rates in organ dysfunction and sepsis: a data-driven analysis in critical care.” *Critical Care* 23: 1–11