**Group Project Phase 3**

**Asthma Data Analysis Project**

Chandana Suggu (cs1178)

Devi Pranathi Vinturi (dv0346)

Siri Vaishnavi Somanchi Venkata (svs0083)

Sri Lakshmi Vellanki (sv1792)

Sunith Kumar Kanakuntla (sk3124)

Department of Information Science, University of North Texas

INFO 5770: Introduction to Health Data Analytics

Dr. Heejun Kim

December 11, 2024

**Project Phase 3**

**Asthma Data Analysis Project**

**Introduction**

Asthma is a severe lung disease that causes difficulty breathing and affects millions of people of all ages worldwide. It always results from a combination of genetic development with environmental factors, including pollen, respiratory diseases, dust mites, air pollution, and tobacco smoking. Poor and middle-income countries have over 400,000 asthma-related deaths recorded annually due to increasing lifestyle changes, air pollution, and urbanization. Global health initiatives aim at overcoming these obstacles through improved disease control, education, and access.

Asthma among the younger age group, which lowers the quality of life, commonly leads to hospitalization. The types of asthma include allergic asthma, exercise-induced asthma, non-allergic asthma, and occupational asthma. The risk increases with an increase in age, but symptoms can decrease or completely disappear as an adult. Gender also affects asthma with asthma being more prevalent in women as adults and more common in males as children. Obesity has also been connected to increased symptoms of asthma due to augmented inflammation. Other factors will include adherence to medication, mental health status, and stress.

**Causes and Treatment**

Asthma is a multifarious illness influenced by genetic predisposition, genetic factors, and environmental ones. It is a syndrome affecting respiratory organs such as the lungs, which can be managed by pharmacological treatment and improvement in lifestyle accordingly. The usual medications adapted to treat asthma are quick-relief drugs, a short-acting beta-agonist like albuterol, medication for long-term control, and inhaled corticosteroids. All these drugs have their own adverse side effects, such as throat irritation or oral thrush.

There is no permanent medication for asthma. However positive changes can be obtained by improvement in lifestyle in the form of abstaining from nicotine, regular workouts, taking foods containing omega -3 fatty acids, fruits full of antioxidants, magnesium and vegetables. Side effects of such medications are somnolence, nausea, viral infections in the throat, sever cough and irregular heartbeats. Certain steroidal drugs and lifestyle changes can help manage asthma and detailed health data analytics can help healthcare professionals, particularly pulmonologists, make informed decisions.

**Research questions**

* How is Aspirin intake related to Total health expenditures for the year?
* How is non-Aspirin intake related to Total health expenditure for the year?
* What is the correlation between AGE19X (age) and ADAPPT42 (Number of times went to doctor’s office or clinic to get care)?

**Related work**

**Aspirin and Asthma disease**

Aspirin is an anti-inflammatory drug utilized for anti-pyretic action, swelling and aches. Prostaglandin in the human body is responsible for pain and inflammatory responses. During fever, there is an increase in production of prostaglandins in the body. The aspirin mode of action includes inhibiting prostaglandins production by blocking COX (cyclooxygenase) enzyme mediating. This results in a reduction of ache and swelling after ingestion of Aspirin drug.

A diagram of a plant

Description automatically generated

**Fig 1. Above chart is drawn using graphviz software**

As a result of Aspirin intake, there are other reactions as well such as narrowing/constriction of airways (Jenkins et al., 2004). This leads to breathing problems, persistent wheezing and aggravated asthma attacks. This condition is called aspirin exacerbated asthma disease (AERD), only noticed in 5-20. % of population.

**Environmental factors affect Asthma**

Air pollen from surrounding trees and nests worsens the asthma condition by triggering the symptoms. Dust, chemicals and smoke affect the airways of lungs and increase the severity of the disease. Occupational exposure to varied chemicals such as PPE induces the on-set of Asthma and its associated effects (Strachan, 2000). Pet mites from the fur of dogs and cats causes irritation, itching and other asthma related symptoms. Mold from unhygienic and moist areas triggers respiratory problems. Nicotine in cigarette smoke triggers respiratory problems to a greater extent. Certain seasons also influence the onset of disease, such as cold and chilly weather. Infestations of viruses, including the flu and the common cold, would make asthma worse.

**Asthma and Total health expenditure**

Health care is expensive in any region of the world, especially in developed countries. So, health insurance and policies play a crucial role in health care sectors. Asthma constitutes the major health expenditure in a year of total health care costs of the patient (Kamble & Bharmal, 2009). Moderate to severely asthma affected patients face this to a greater extent. Factors influencing / triggering asthma would increase the health expenditure for the patients at any phase of the year. This results in an increase in overall total health expenditure of the year (TOTEXP19).

**Methods**

To start with our group selected asthma as the topic. Data was retrieved from the Medical Expenditure Panel Survey (MEPS) and filtered the records using ICD-10 code J45, which is related to asthma and during filtering we ensured that only patients diagnosed with asthma were included in the analysis. The preprocessing of the data is taken in two phases which are explained below:

**Phase 1**

I the phase 1 of the project, we conducted background research, collected reference articles, and reviewed literature on asthma etiology and its major influencing factors. This involved studying academic literature to understand the topic comprehensively. We analyzed the data and examined treatment statuses. Throughout the study process we collected the data, edited it and provided key insights. We also understood disease management during this process. To conclude, this phase entailed substantial literature review, data analysis, and collaboration to evaluate and interpret asthma etiology and treatment data.

**Phase 2**

In phase 2 of the project, we focused on removing the unnecessary variables and handling missing data. We filtered the dataset using the ICD-10 code J45 to identify asthma patients and removed duplicate entries based on unique patient identifier which is DUPERSID. We also identified and addressed noisy data, calculated numeric data correlation coefficient, and determined degrees of freedom and critical values for categorical data using chi-square method. These steps ensured the dataset was accurate and reliable, preparing it for statistical testing and analysis.

**Statistical Tests**

In phase 3 of the project, descriptive analysis of the pre-processed data was conducted to address the research questions. We used the variable CHPMED42 (aspirin intake status) in the descriptive analysis to compare with TOTEXP19 (total healthcare expenditure). Additionally, we examined the correlation between AGE19X (age) and ADAPPT42 (number of healthcare visits). To find the solutions, we used the following tests:

1. Pearson Correlation test
2. Chi-square test
3. Mann-Whitney test

These tests were applied to both numerical and categorical variables.

* 1. **Pearson Coefficient**

The Pearson correlation test was used to find the correlation between AGE19X (age) and ADAPPT42 (number of healthcare visits). Pearson correlation has been used to find the direction of a linear relationship between two continuous variables, which helps to know how one variable shows an effect on another variable. A value near 1 implies a strong positive association, while -1 suggests a strong negative correlation.

* 1. **Chi-Square Test**

The Chi-square test is used to find the correlation between categorical variables. The chi-square value and p-value help us know the normal distribution of data and find the null hypotheses based on our results, which show no rejection of the null hypothesis, and the variables are in a normal distribution.

* 1. **Mann-Whitney U test**

The Mann-Whitney test is used because we compare two independent variables, and it is also chosen based on the abnormal results of the skewness and kurtosis. Two groups (aspirin users and non-users) are the reasons for choosing this statistical Mann-Whitney test. This test is used to provide answers to show association between aspirin intake, healthcare expenditure, and healthcare visits.

**Results**

By doing the statical tests, the taken variables show a correlation between each other, which is explained below by interpreting the below test results:

**Mean and Maximum**

We are finding the mean and maximum for both aspirin users and TOTEXP19 and non-users and TOTEXP.

A screenshot of a computer code

Description automatically generatedA screenshot of a computer code

Description automatically generated

**Fig 2. Aspirin intake and TOTEXP19 Fig 3. Non-users and TOTEXP19**

The mean and maximum values of aspirin users and TOTEXP19 are 12224.790 and 87972.0, respectively. Meanwhile, the mean and maximum values of non-users and TOTEXP19 are 10381.190 and 187613.0, respectively. The above results show that patients who take aspirin tend to have higher average healthcare expenditures (TOTEXP19) compared to non-aspirin users, indicating more consistent management of their condition.

**Graphs**

The histogram below shows the comparison between the patients of aspirin users and TOTEXP19 (Healthcare expenditure) and non-users and TOTEXP19 (Healthcare expenditure). The graph shows that non-aspirin users have higher variation and outliers in healthcare expenditures, where aspirin users have more consistent and lower expenditures.

**A graph with many colored bars

Description automatically generated with medium confidence**

**Fig 4. Histogram between Aspirin intake, non-users and TOTEXP19**

The boxplots for aspirin users and non-users reveal distinct differences in healthcare expenditures (TOTEXP19). Non-aspirin users exhibit greater variability, with a wider interquartile range (IQR) and more extreme outliers, including expenditures exceeding $150,000. In contrast, aspirin users show a narrower IQR, indicating less variability and more consistent healthcare costs, with outliers present but within a more moderate range. The median healthcare expenditure is slightly higher for aspirin users, suggesting a marginally increased typical cost. Additionally, the upper whisker for non-aspirin users extends significantly higher than that of aspirin users, reflecting a higher likelihood of very large expenditures among non-users. Overall, the results suggest that aspirin users experience more stable and predictable healthcare costs, while non-users face greater expenditure variability, potentially due to more severe or acute health conditions.

A graph of a graph

Description automatically generated with medium confidence

**Fig 5. Box Plot between Aspirin intake, non-users and TOTEXP19**

**Statistical Tests**

**Pearson Correlation test**

**1.Correlation between ADAPPT42 (Number of healthcare visits) and AGE19X (Age)**

The Pearson Correlation test between ADAPPT42 (number of healthcare visits) and AGE19X (age) shows a value of 0.201914, showing a weak positive relationship between the number of healthcare visits and the age of the patient. This indicates that as age increases, the number of healthcare visits tends to increase slightly. However, this relationship is not strong, and other factors beyond age are likely to influence the frequency of healthcare visits for asthma patients.

A number with numbers on it

Description automatically generated with medium confidence

**Fig 6. Correlation between ADAPPT42 and AGE19X**

**2.Correlation between TOTEXP19 (Total healthcare expenditure) and AGE19X (Age)**

The Pearson Correlation test between AGE19X and TOTEXP19 shows a value of 0.140840, indicating a weak positive relationship between age and healthcare expenditure. This means that if age has some effect on healthcare expenditure, other factors also play a important role in determining the total expenditure for healthcare services.

**A screenshot of a computer

Description automatically generated**

**Fig 7. Correlation between TOTEXP19 and AGE19X**

**Chi-Square Test**

The Chi-square correlation coefficient and test results present the Chi-square statistic (97.4643299366905), p-value (5.483526947190896e-23), degree of freedom (1), and a contingency table for expected frequencies [[737.7203458, 737.27965342], [936.27965342, 935.7203458]]. The chi-square value is 97.46, and the data shows a statistically significant relationship. The p-value is extremely small 5.483526947190896e-23, indicating strong evidence to reject the null hypothesis of independence. This suggests that the variables are correlated. The degree of freedom is 1, showing the chi-square test involved two categories of comparison. To conclude, the test confirms a significant relationship between the variables, rejecting the null hypothesis.

A computer screen shot of a computer code

Description automatically generated

**Fig 8. Chi-Square test**

**Skewness**

Based on the results, the distribution is asymmetric due to skewness exceeding the normality threshold of >2, with skewness values of 3.6339 for x and 5.1689 for y.

**A screenshot of a computer

Description automatically generated**

**Fig 9. Skewness**

**Kurtosis**

Based on the results, the low probability distribution is peaked due to kurtosis exceeding the normality threshold of 7, with values of 17.9158 for x and 39.4945 for y.

**A screenshot of a phone

Description automatically generated**

**Fig 9. Kurtosis**

**Mann-Whitney test**

Based on the abnormal results of the skewness and kurtosis, two groups (aspirin users and non-aspirin users) were compared using the statistical Mann-Whitney test. The test produced a result of Mann Whitney uResult (statistic=1710175.0, pvalue=1.5375668254733864e-32). The extraordinarily low p-value (1.537e-32) fails to support the null hypothesis because the p-value is significantly less than 0.05. This indicates a significant variation in total healthcare expenditures (TOTEXP19) between aspirin users and non-aspirin users.

**A screenshot of a computer code

Description automatically generated**

**Fig 10. Mann-Whitney test**

**Discussion**

**The above results clearly answer the research question, including:**

Based on the Mann-Whitney test, the results show a Mann-Whitney U statistic of 1710175.0 and a p-value of 1.5375668254733864e-32 which is very low and shows a meaningful relationship between aspirin usage and healthcare expenditures (TOTEXP19). The low p-value shows that there is a statistically significant difference in total healthcare expenditures between aspirin users and non-aspirin users. Aspirin users have more consistent healthcare expenditures, but non-users show greater variability and higher maximum expenditures, indicating potentially more severe or mismanaged health conditions among non-users.

The Pearson correlation between ADAPPT42 (Number of Healthcare Visits) and AGE19X (Age) shows a weak positive correlation with a value of 0.201914, suggesting a slight relationship between age and the frequency of healthcare visits. However, other factors, such as disease severity and socioeconomic factors, likely contribute to the variation in healthcare visits. At the same time, the Pearson correlation between TOTEXP19 (Total Healthcare Expenditure) and AGE19X (Age) shows a weak positive correlation with a value of 0.140840, indicating that age may have some effect on healthcare expenditures, other factors such as treatment options and associated conditions also play an important role. These findings highlight the multifactorial nature of healthcare utilization and expenditures in asthma management.

**Reflection**

A detailed study on healthcare data analytics was conducted this semester focusing on analyzing healthcare expenditures among aspirin users and non-users. The process involved extensive literature review, data preprocessing, and statistical analysis to understand relationships between variables and evaluate healthcare patterns. We ensured accuracy and reliability by addressing missing values, filtering irrelevant data, and identifying outliers. Applying correlation analysis, Chi-square tests, and the Mann-Whitney U test helped us interpret relationships between variables such as age, healthcare visits, and expenditures. Visualizations like histograms and boxplots were generated to better understand data distribution and variability. The results show significant differences in healthcare expenditures between aspirin users and non-users, focusing on the influence of preventive measures and management strategies on costs. While data quality and complexity posed challenges, this project showcased the potential of health data analytics to enhance decision-making in disease management and healthcare expenditure analysis. The study underscored the importance of robust techniques and interdisciplinary approaches in this evolving field.

**Conclusion**

To conclude, based on the results, patients who are taking aspirin tend to have slightly higher average healthcare expenditures compared to non-aspirin users. However, non-aspirin users exhibit greater variability and significantly higher maximum expenditures, indicating potential disparities in healthcare needs or management strategies. This is important in asthma patients as aspirin can sometimes exacerbate symptoms in some individuals such as those with Aspirin-Exacerbated Respiratory Disease (AERD). The analysis also shows a weak positive correlation between age (AGE19X) and healthcare visits (ADAPPT42) suggesting that age may also slightly influence the frequency of healthcare visits. Other factors like disease severity, environmental triggers, and healthcare access likely contribute as well. Similarly, skewness and kurtosis results indicate that the distributions of expenditures are highly asymmetric and peaked, reflecting the variability and extremity of healthcare costs among asthma patients. These findings emphasize the importance of tailored healthcare management strategies for both aspirin users and non-users to optimize outcomes and expenditures.

**Student Contribution in the Group**

**Chandana Suggu**Contributed to the research literature review, analysis of nominal data, determining the degree of freedom for categorical data, and performing the Pearson correlation test.

**Devi Pranathi Vinturi**Focused on data preprocessing, formulating research questions, handling noisy data, conducting statistical analyses, performing the Chi-square test, and participating in discussions.

**Siri Vaishnavi Somanchi Venkata**  
Played a key role in literature review, data preprocessing, generating and analyzing random samples, creating graphs, and interpreting results to answer the research questions.

**Sri Lakshmi Vellanki**Was responsible for identifying and filtering indiscriminative variables, addressing missing data, performing the Mann-Whitney test, interpreting statistical results, and contributing to the conclusion.

**Sunith Kumar Kanakuntla**Assisted in refining the methodology, cross-checking statistical analyses, compiling results, and writing the final report, ensuring clarity and coherence across the document.

The team members equally distributed the work to successfully complete the project.

**Reference**

Jenkins, C., Costello, J., & Hodge, L. (2004). Systematic review of prevalence of aspirin induced asthma and its implications for clinical practice. *BMJ*, *328*(7437), 434. <https://doi.org/10.1136/bmj.328.7437.434>

Strachan, D. P. (2000). The role of environmental factors in asthma. *British Medical Bulletin*, *56*(4), 865–882. <https://doi.org/10.1258/0007142001903562>

Kamble, S., & Bharmal, M. (2009). Incremental direct expenditure of treating asthma in the United States. *Journal of Asthma*, *46*(1), 73–80. <https://doi.org/10.1080/02770900802503107>