Which demographic and socioeconomic factors affect the prevalence of childhood allergies the most?

By Neeza Singh, Sanjeev Pal

Allergies are unique to various other diseases in that they can only be developed if the individual comes into contact with the substance or food item. Food allergies occur when the immune system creates IgE antibodies in response to a specific allergen from consuming or touching it. IgE attaches to mast cells and basophils, triggering an immune response upon subsequent exposure. However, the surprising aspect of allergies is that they can emerge even after coming into contact with the food. In fact, millions of adults in the US have developed a sudden allergy to a food they've eaten their entire lives, a phenomenon we wish to explore. The random occurrence of allergies creates scope for more statistical analyses about their origins. From a purely observational standpoint, we noticed that allergies are also sporadic in their geographical distribution. In many countries in the east, peanut or wheat allergies are nearly unheard of, probably because of their prevalence in daily cuisine. For example, In Israel, there is a joke that the first three words out of every toddler's mouth are: abba(dad), ima(mom) and Bamba, a popular peanut snack.

When analyzing the origin and spread of diseases, two general categories are given the most prominence: individual genetics and environment. In our project, we wanted to delve deeper into both genetic and environmental factors that may lead to the susceptibility of childhood allergies. Although allergies are a medical condition that many believe are developed at birth, many dismiss their emergence in adulthood.

Due to the arbitrary nature of childhood allergies, we felt that it was interesting to analyze multiple factors, such as the role of gender, race, and ethnicity allergy development. Additionally, the accessibility of health care resources, age, and location are socioeconomic categories that have historically influenced the epidemiology of other diseases. We will be analyzing these different variables in relation to different childhood food-related allergies through the findings of the 2016 BMC Pediatric research study. This data set analyzes a plethora of childhood allergies, from common nuts to fish, and other miscellaneous food groups in a cohort study.

About the Dataset

Our data set is part of a research article that focuses on the epidemiologic characteristics of healthcare provider-diagnosed eczema, asthma, allergic rhinitis, and food allergy in children. The author, David Hill followed a retrospective cohort study, a study in which the medical records of groups of individuals are alike in most ways but differ in a few key characteristics. The researchers employed a care network of 1,050,061 urban and sub-urban children to establish two retrospective cohorts. The first cohort comprised 29,662 children who were part of a closed birth group, while the second cohort included 333,200 children in a cross-sectional study. These cohorts enabled the researchers to examine the epidemiologic characteristics of the medical conditions under study. To determine the association between food allergy and respiratory allergy, they utilized logistic regression analysis.

The dataset contains 333,000 rows or unique individuals and 55 columns or variables. The variables include demographic information such as age, gender, race, and ethnicity, as well as clinical information such as the presence or absence of the four conditions of interest (eczema, asthma, allergic rhinitis, and various food allergy), as well as other comorbidities. Our research mainly focuses on food allergies and the most prevalent childhood conditions. The dataset also includes information about the healthcare provider who diagnosed the conditions, the date of the diagnosis, and the source of the diagnosis. Most of the variables are categorical, such as the four conditions of interest, race, and ethnicity, while others are numerical, such as age. The units for age are years.

The data was collected retrospectively from electronic health records, which may introduce some biases. For example, the data may not be representative of the general population, as it only includes children receiving medical care in a specific healthcare system. Additionally, there may be some selection bias, as children with certain conditions or demographics may be more likely to seek medical care than others. These biases could limit the generalizability of the study's conclusions.

The dataset includes specific alphabet-number codes for different variables such as Race and Gender. For the sake of simplicity, we plan on replacing the codes with the name itself. In terms of missing values, the dataset contains many rows with “NA” entries, indicating that the child does not have the listed illness. Although there is no real “data” in these entries, we believe that it is important to include them to analyze the prominence of the different illnesses. There appears to be no noticeable outlier or outlier since the patients are very similar in terms of age and the variables tend to be binary: for example, sex as male or female, healthcare coverage as Medicaid or No Medicaid, and whether an illness is present or not. Overall, the dataset provides valuable information about the epidemiological characteristics of healthcare provider-diagnosed eczema, asthma, allergic rhinitis, and food allergy in children, but care should be taken when interpreting the results due to potential biases in the data collection process.

Hypotheses

The first hypothesis within our research explores where or not government subsidized healthcare affects the prevalence of food allergy diagnosis. In the dataset, researchers from the epidemiological study simultaneously focus on government aid within their evaluation of childhood food allergies. The cost of maintaining a nutritious and healthy lifestyle can be a toll on some, especially since this study takes into account children from both suburban and urban settings. An urban setting is most likely to be more dense in population as well as house a greater proportion of low-income families, due to the expenses associated with a suburban lifestyle. Keeping this socioeconomic context in mind, we have decided to create the following hypotheses:

$H0$- The prevalence of food allergy diagnosis does not differ between populations of children with and without Medicaid, a government-subsidized health care service.

Ha: The prevalence of food allergy diagnosis does differ significantly between population of children with and without Medicaid, a government-subsidized health care service.

The population of interest in this scenario would be individuals with food allergy diagnosis. To test this hypothesis, we will be focusing on whether or not a child comes from a family with Medicaid (Y or N) and the average number of food allergy diagnoses between the two populations. This can be done by iterating through each subject ID in our dataset (which corresponds to one of the 330,000 rows of unique individuals) and partitioning between Medicaid holders and non-Medicaid holders, the populations that will be used for comparison as the independent variables. Each subject ID within these populations will have an associated number of food allergies which can then allow us to calculate a population average for the purpose of testing our hypothesis.

The second hypothesis follows a similar socio-economic approach as the first. We aim to explore whether or not being part of a minority group (defined as non-White ethnicity) plays a role in the diagnoses of various childhood illnesses. In the United States, many minority populations are uninsured or underinsured and cannot afford healthcare services that allow for the diagnosis of diseases. Additionally many minorities face language and cultural barriers that prevent them from seeking routine medical checkups. With these factors in mind, among many others, we felt that it was worth analyzing the difference in average rate of diagnosis for people in certain demographic groups through the following hypotheses:

$H0$: There is no significant difference in the prevalence of food allergy diagnosis between minority and non-minority children populations.

Ha: The prevalence of food allergy diagnosis does, on average, differ between minority and non-minority children populations.

The population of interest in this hypothesis would again be children with a food allergy diagnosis reported in the healthcare database. To test this hypothesis we will be clustering the children by minority status and then taking an average number of childhood allergy diagnosis for each subject ID in order to produce an overall average for both variables (minority and non-minority).

The third hypothesis we have chosen for this study analyzes the time in which different childhood illnesses develop. Oftentimes, some allergies and common chronic diseases like asthma develop later on in life, sometimes due to environmental factors that can manipulate genetic expression. Using the socioeconomic variables discussed previously in our list of chosen hypotheses, we can easily evaluate links between external factors and the time range in which illnesses develop. For the purpose of simplicity, though, our third hypotheses are addressed as followed:

$H0$: The time range in which childhood allergies develop does not have an effect on the type of allergy and disease a child develops within this study.

Ha: The time range in which childhood allergies develop do have an effect on the type of allergy and disease a child develops within this study.

The population of interest in this study is again any child with a childhood allergy or illness diagnosis. The goal of this hypothetical approach is to gauge whether certain childhood illnesses develop in certain time periods in life, either early childhood and infancy or during the middle of childhood and before puberty. We will be measuring this with the average number of diagnoses per age group (as described in the previous sentence) and then labeling main childhood illnesses with the most probable time of development within childhood, a categorical measurement that does not require units.

The last hypothesis aims to compare illnesses rather than factors that determine the prevalence of the diseases studied in our dataset. Our provided dataset includes the start and end of every disease that is studied. Through the following hypotheses, we aim to compare food allergies and Asthma to see which one is most common among the 333,000 children present in this study:

$H0$: There is no significant difference between the number of children who have a food allergy vs the number of children who have asthma.

Ha: There is a significant difference between the number of children who have a food allergy vs the number of children who have asthma.

The population of interest in this study is the children who have asthma as well as the children who have a food allergy. We will be comparing these populations to see which one of the illnesses has a higher chance of containing a null “end year” value in the dataset. In terms of the formal parameters and units, we will be calculating the average number of null values within the “end date” column, indicating the end of the disease in each subject ID (children).

Data Visualizations

In order to test the hypotheses we discussed previously, we decided to create simple data visualizations in order to gauge which of the hypotheses are supported by the dataset. The first hypothesis focuses on the correlation between possessing Medicaid and developing a childhood allergy. The goal is to either reject or accept the following statement: The prevalence of food allergy diagnosis does not differ between populations of children with and without Medicaid, a government-subsidized health care service. First, we created a bar graph to determine whether the population of children possessing Medicaid is similar to the overall US population. In Figure 1, we can see that there is about a 15% difference between the proportion of children who have Medicaid in our sample and the proportion of children who have Medicaid in the United States. This is a significant difference which may lead us to believe that the sample may not be representative of the population. Having a large sample size (333,000+) can provide more confidence in the representativeness of the sample when compared to the US population. Since many external factors are at play, such as the lack of clarity in the population parameters, we cannot say for sure if the sample is representative. To increase our confidence, we compared the relative frequencies of children who have an illness in this sample vs in the population, as visualized in Figure 2.

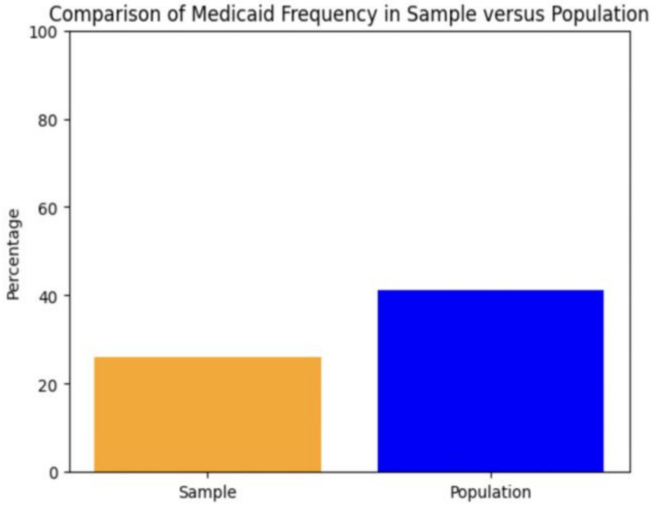

Figure 1

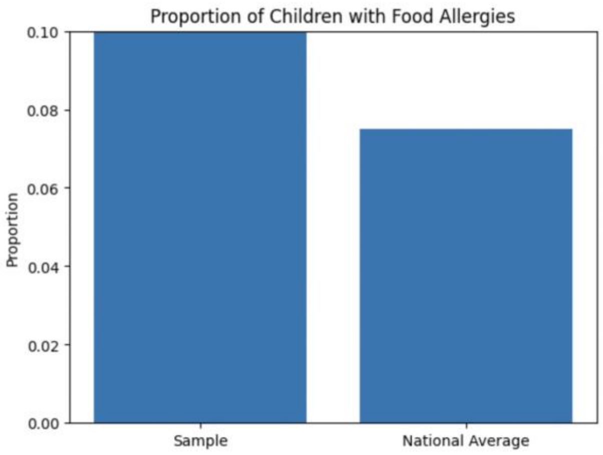

Figure 2

In Figure 2, we can see that there is roughly a 1.5% difference between the national average number of children who have food allergies and the sample average number of children who have food allergies. Since we are utilizing a large dataset and since the researchers in this study confirmed the randomness of this study, we are confident enough to create further data visualizations. To determine whether or not Medicaid affects the likelihood of being diagnosed with a food allergy or asthma, we created a pie chart of our data, as shown in Figure 3. Out of the population of children who have an allergy, about 70% of them do not have Medicaid. From our general research, we learned that there is a minimum income needed to qualify for Medicaid, indicating that a lack of accessibility to proper nutrition or medical resources and knowledge may result in a higher likelihood of intercepting any sort of illness. The lack of diversity in diets may also lead to developing an allergy towards a specific food.

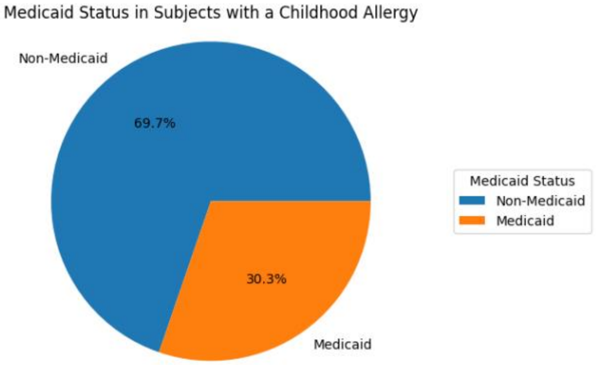

Figure 3

With this information in mind, we cannot reject or accept any hypothesis; however, there is an obvious majority in the Non-Medicaid section that can allow us to assume that there is a correlation between Medicaid Status and food allergy diagnosis.

Our second hypothesis focused on race and food allergy susceptibility. We wanted to see if a certain racial group is more likely to get an allergy in the US. First, we tried to determine whether the racial breakdown present in this sample resembles that of the US population, as seen in Figure 4.

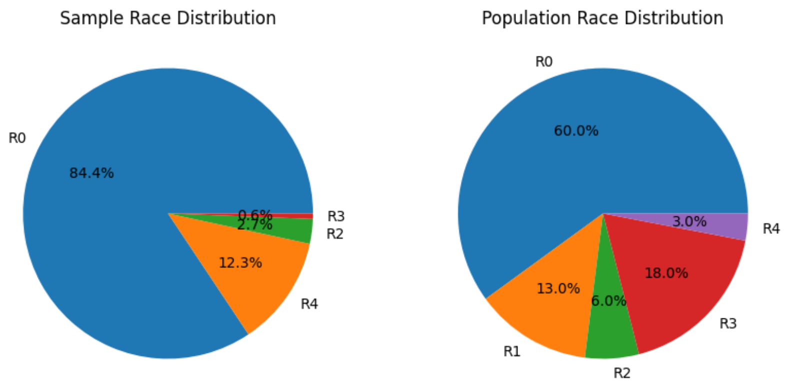

Figure 4

From these race distribution pie graphs we can see that there seems to be a sizable difference between the white population (R0) and other (R3) in both distributions, leading us to believe that the data assigned to white children may not be representative. Since our sample dataset does not include as many race categories as the population dataset, the other category may be less significant to focus on. Other race categories seem to resemble each other in terms of proportion between the sample and population. We proceeded to create another pie chart, as seen in Figure 5, which showed the breakdown of each race and the percentage of children within each race who have a food allergy.

Race Distribution of Children with Food Allergies

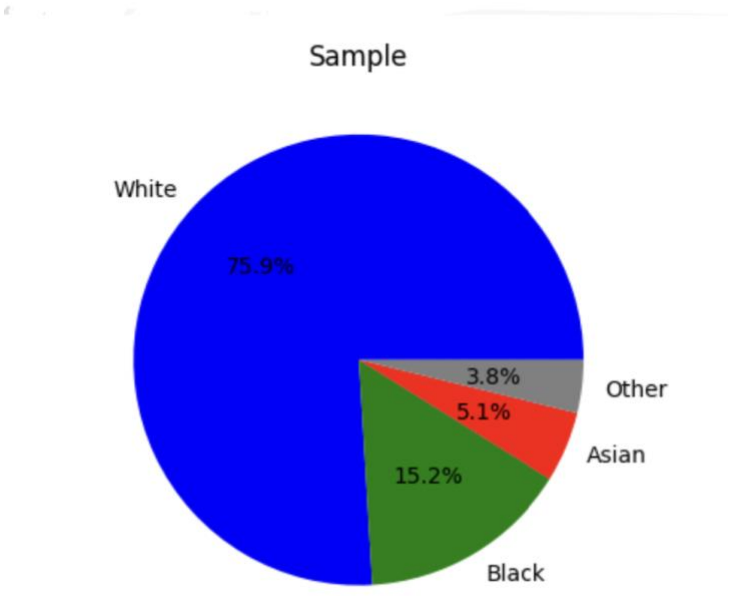

Figure 5

In Figure 5, we can see that the large majority of children with a food allergy are white while the remaining 25% include the other 3 race categories. The R4(Unknown) category is not present in this pie graph due to a lack of data. Nonetheless, this visualization helped us make contextual predictions on how food allergies work and their behavior in different circumstances. We decided to perform a Chi-square test of independence to determine if there is a significant association between peanut allergy and Hispanic status in the sample. Our function returned the chi-square statistic, p-value(0.05), degrees of freedom, and expected frequencies where we found that there is indeed a significant association between peanut allergy and hispanic status. More detailed hypothesis testing for each hypothesis will be discussed later on in this research. Food allergies specifically tend to develop at a young age and disappear by adulthood; if a child is not exposed to that type of food in their childhood, then they are more likely to develop an allergy. One racial cohort in particular, Asian/Pacific Islanders, are known to have the most peanuts in their diet. In Figure 6, we compared Peanut Allergy numbers between White and Asian/Pacific Islander population. The bar graph shows that there are about 60,000 more white children that experience peanut allergy as compared to their Asian counterparts. However, this bar graph may be misleading due to the misrepresentation of minority populations within the dataset. There was a higher percentage of white children being studied as compared to minority students which causes a statistical error in our research, as shown in figure 6.2. Without an equal representation of each demographic group, it is difficult to conclude whether a certain allergy affects a certain group of people more. Any claims made about race as a demographic factor that may contribute to allergy development should be taken with a grain of salt in the context of this specific dataset.

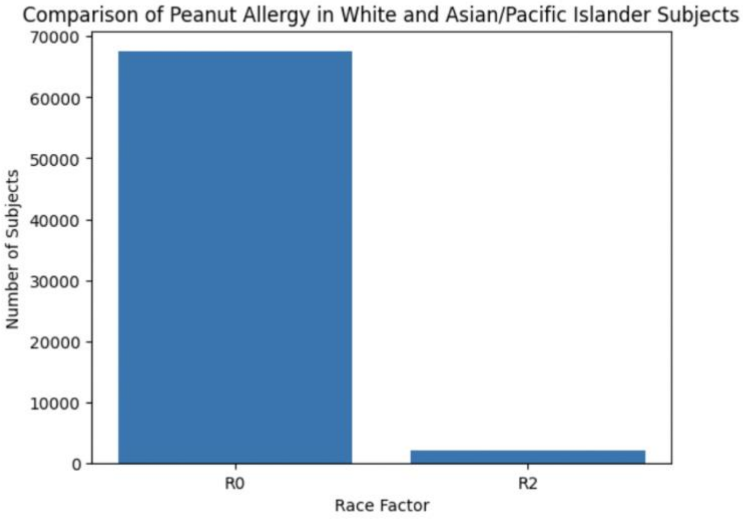

Figure 6

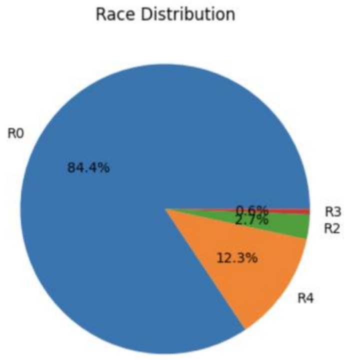

Figure 6.2: Race Distribution of Children in Food Allergy Dataset

Our third hypothesis focuses on the behavior of the twelve allergies discussed, more specifically when they tend to develop and terminate. We wanted to analyze the time range in which childhood allergies developed and if there are any visible patterns among them.

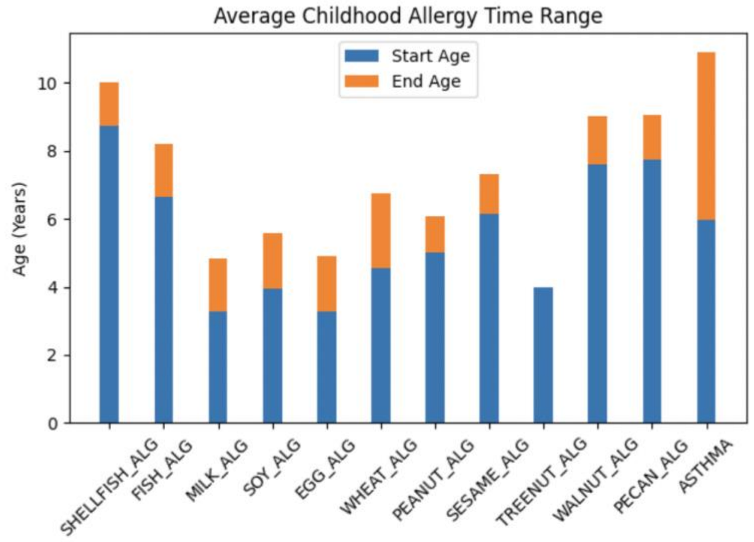

Figure 7

We created a stacked bar graph which shows the average start age and end age of each illness, as seen in Figure 7. From what we can see, many of the illnesses behave the same in terms of the time range they are present in a child's life. We felt it was important to note that asthma usually starts later on in early childhood, after age six. On the other hand, common food allergies tend to develop in the first few years of life and terminate between age six and eight. Tree Nut allergies were very rarely presented in this dataset and usually last longer than the ages presented in this dataset, which explains the lack of an end year section in the stacked bar graph. From the sporadic patterns we see in the bar graph, we do not believe that the type of allergy has an effect on the time range it is present; other individual confounding variables may affect when an allergy starts and ends.

For our final hypothesis, we were interested in comparing the prevalence of Asthma vs Food allergies in young children. To evaluate this, we sorted children into "Asthma," "Food Allergy," or "Both" categories into a simple bar graph. From Figure 8, it is clear that there is an overwhelming difference between Asthma and Food Allergy numbers with Asthma affecting over 60,000 children in this study. It is reasonable to assume that Asthma is far more prevalent than Food allergy, especially with a hypothesis test.

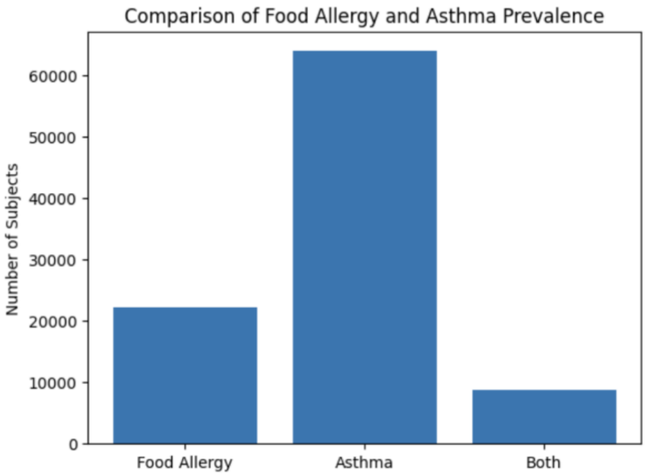
Comparison of Food Allergy and Asthma Prevalence
Number of Subjects
60000
50000
40000
30000
20000
10000
0
Food Allergy
Asthma
Both

Figure 8

Hypothesis Testing

Visualizations aided our team in bringing the dataset to life while simultaneously providing a basis for the hypothesis testing we conducted afterwards. In many cases, our visualizations were disproved completely through our computation of p values and testing as a whole; however, some hypotheses appeared to match the results in our graphs that were previously discussed. In this study, we conducted hypothesis testing on four different hypotheses to explore the associations between demographic factors, age of allergy onset, and the prevalence of food allergies and asthma in a cross-sectional dataset of 333,000 children.

In hypothesis testing, the alpha value is determined by the researchers and represents the level of significance, which is the probability of rejecting the null hypothesis when it is true. In other words, it is the probability of making a Type I error. An alpha value of 0.05, or 5%, is a common threshold used in many scientific studies. An alpha value of 0.05 provides a balance between Type I errors (rejecting a true null hypothesis) and Type II errors (failing to reject a false null hypothesis), which is why we chose 0.05 as our alpha value for each hypothesis test we ran. We mostly dealt with two-proportion z tests and concluded that the sample size is large enough with 333,000 rows of data and that the standard deviations and means were given (after basic calculations).

For the first hypothesis, we created two separate data frames for children with and without Medicaid using the pandas package in Python. We then calculated the proportions of food allergy diagnosis in each dataset and used the two-proportion z-test from the statsmodels package in Python to compare the proportions between the two groups. The binary organization of the new data frame produced a more efficient method of hypothesis testing. With a test statistic, z, of 0.48 and a p-value of 0.63062, we failed to reject the null hypothesis that the prevalence of food allergy diagnosis does not differ between populations of children with and without Medicaid. This finding suggests that there is no significant evidence to indicate that the prevalence of food allergy diagnosis is different between children with and without Medicaid, implying that access to government-subsidized health care may not play a major role in the diagnosis of food allergies among these children.

For the second hypothesis, we created two separate data frames for minority and non-minority children using pandas. Minority status was determined by whether or not a child falls into the White category; in other words, a subject ID is labeled as a Minority when they have anything but an “R0” in the Race Factor column. We then calculated the proportions of food allergy diagnosis in each group (Minority vs. Non-Minority) and used the two-proportion z-test from statsmodels to compare the proportions between the two groups. With a test statistic, z, of 0.56 and a p-value of 0.57736, we failed to reject the null hypothesis that there is no significant difference in the prevalence of food allergy diagnosis between minority and non-minority children populations. This indicates that there is no significant evidence to support the claim that the prevalence of food allergy diagnosis differs between minority and non-minority children populations. One possible explanation for this could be due to the fact that there is an uneven proportion of white children being tested vs minority children in this dataset. Although our visualization in Figure 6 shows an overwhelming majority of white children with a peanut allergy diagnosis when compared to their minority counterparts, we also know that there are more Subject ID’s who identify as White. Due to this misrepresentation, we cannot make any further conclusions on whether race is a significant factor in any sort of childhood illness diagnosis.

To conduct our hypothesis testing for the third hypothesis, we used a slightly different approach than the two proportion z tests that were previously conducted. We used the Chi-Square test of independence since it's a common statistical test used to determine whether there's a significant association between two categorical variables. In this case, we wanted to test the association between the time range in which childhood allergies develop (categorized into age ranges) and the type of allergy and disease a child develops. Both of these variables are categorical, making the Chi-Square test of independence an appropriate choice for this analysis. In order to computationally conduct a Chi-Square test, we imported the chi2_contigency module in Python and created a new dataframe to store the allergy occurrences and age range for each illness. Our test statistic is the chi-square value, which we computed to be 24,802. This value is a measure of how much the observed frequencies in a contingency table differ from the expected frequencies under the assumption that the null hypothesis is true. The assumptions of the chi-squared test include having a large enough sample size and having categorical data. In this case, we have a large dataset of 333,000 children, so the sample size assumption is satisfied. The p-value associated with the test statistic is 0.005. This p-value indicates the probability of observing a chi-square value as extreme as or more extreme than the one calculated, assuming the null hypothesis is true. In this case, a p-value of 0.005 suggests that the observed association between the time range and type of allergy is very unlikely to have occurred by chance alone.

Based on the obtained p-value, we reject the null hypothesis that the time range in which childhood allergies develop does not have an effect on the type of allergy and disease a child develops within this study. Instead, we accepted the alternative hypothesis and concluded that there is a significant relationship between the time range of allergy development and the type of allergy a child develops. Since the hypothesis test we conducted allowed us to reject the null hypothesis, we decided to further analyze which disease in particular tends to have the highest time range amongst the subjects in this study. The disease with the longest time range is Asthma, with the maximum time range being over 20 years.

For our final hypothesis, we used a standard two proportion z-test from the statsmodels module in Python to compare the proportions between the two groups: children with and without an Asthma diagnosis. From our previous hypothesis, it was clear that Asthma tends to affect an individual for a longer period of time than common childhood food allergies, but we wanted to see if Asthma affects children at a higher rate in general. The computational aspect of this hypothesis testing was fairly simple; we created a table of subject IDs who had Asthma (determined by whether or not there were entries in the Asthma Start column and later carried out the z-test. After running our test a few times for accuracy and precision, we obtained a large test statistic, z, of 31.48 (rounded to two decimal places). A Z-statistic of 31 is quite large and would typically indicate that the difference between the two proportions is highly significant. However, statistical significance does not always imply practical significance. In some cases, a large z-statistic may be driven by a very large sample size that can make small differences between proportions appear significant, which is true in respect to this study. We decided to calculate the effect size using Cohen's h statistic, a measure that is used to describe the difference between two groups or conditions. We calculated this statistic by dividing the difference in means between the two groups by the pooled standard deviation of those groups. Our code gave us an output of 0.3 which indicated a small to medium effect size; in other words, there is a noticeable difference between the two groups (Asthma vs No Asthma), but it may not have practical significance if the sample size is not large or if the nature of the variables prevent us from reaching statistical significance. To further test our hypothesis, we obtained a p value of 0.003 from the z-test, a value that is much smaller than our alpha value of 0.05, allowing us to reject the null hypothesis that there is no significant difference between difference between the proportion of children who have a food allergy vs the number of children who have asthma. Our hypothesis testing allowed us to conclude that our data is unlikely to have occurred by chance alone if the null hypothesis were true. In this case, we decided to accept the alternative hypothesis that there is a significant difference between the proportion of children who have a food allergy vs the number of children who have asthma, especially after our calculation of Cohen's h statistic.

Conclusions

In this research study, we aimed to investigate the relationship between childhood allergies and the development of various types of allergies and diseases later in life. We conducted fourth hypothesis tests, each focusing on a different aspect of childhood allergies and how external factors affect their prevalence and outcomes. Our results provide insights into the characteristics and behavior of childhood allergies and highlight the importance of early intervention and prevention.

Through our first hypothesis we aimed to determine if there was a significant difference in the proportion of children who possessed Medicaid and the proportion of children who did not possess Medicaid, specifically in regards to allergy development. For each hypothesis, we utilized different packages and libraries present within Python. The general method for the computational aspect of our hypothesis testing involved creating new columns in the dataset, through pandas dataframes, and then calculating the proportions present in the column. For example, we used the Medicaid Factor column and created 2 new columns, "Medicaid" and "No Medicaid." From here, we were able to calculate the proportion of children with an allergy diagnosis in each of the two new columns, allowing us to then use the statsmodel.api package in Python to carry out the two-proportion z test. Our results found no significant difference between children who had Medicaid and did not have Medicaid in terms of allergy diagnosis. For the second hypothesis, we also conducted a two proportion z test using a similar computational method used in the first hypothesis, leading to a failure to reject the null hypothesis that there is no significant difference in food allergy prevalence between non minority and minority populations. We were unable to keep analyzing race as a factor for food allergy in this specific dataset due to the disproportionate presence of white children as compared to minority children. Nearly 75% of subjects were white, making it difficult to make any conclusions about race as a variable in our study. For the future, a dataset with equal population sizes for each race would be more beneficial in terms of creating and testing a hypothesis.

Through our third hypothesis, we analyzed the association between food allergies and the specific time range in which they are active using a chi square test in Python. Our results showed that there was an association, particularly with asthma patients, in terms of the time range of childhood allergies. Asthma is considered a chronic respiratory disease which explains why the time range that a child has asthma is usually far greater than if they have a food allergy. The final hypothesis also allowed us to make more inferences about the nature of allergies vs Asthma in general. Although we found a significant difference between the proportion of children with an allergy as opposed to Asthma, it is important to note that Asthma acts differently than food allergies in general since it involves the capacity of the volume of air that children inhale. Allergies, on the other hand, allergies largely concern the immune system. For a future study, it would be more efficient to separate asthma and food allergies all together in terms of studying their emergence and cause.

While our study provides important insights into the long-term consequences of childhood allergies, there are several limitations that should be addressed in future research. Firstly, our study's data was collected through randomly selected electronic health records, which may not always lead to a representative sample. Future studies could consider using medical records from a wide geographical pool and also account for the representativeness of each group of people. Additionally, our study did not account for potential confounding variables such as family history of allergies, environmental factors, and lifestyle factors. Future studies could explore the role of these variables in the development of allergies and diseases in adulthood, especially since allergies tend to act differently depending on the region, context, and culture in which a child grows up in.

References

Hill DA, Grundmeier RW, Ram G, Spergel JM. The epidemiologic characteristics of healthcare provider-diagnosed eczema, asthma, allergic rhinitis, and food allergy in children: a retrospective cohort study. BMC Pediatr. 2016 Aug 20;16:133. doi: 10.1186/s12887-016-0673-z. PMID: 27542726; PMCID: PMC
4992234.

Appendix

The next page contains Python code that was used in creating our visualizations as well as hypothesis tests. Each block of code contains comments that describe what each function does. The code is a direct export from Google Colaboratory.