Justin Harding 450221916

INFO2150 Report

Table of Contents

**1. Introduction3**

1.1 Purpose3

1.2 Hypothesis3

1.3 Key Documents3

1.4 Raw Data3

**2. Methods4**

2.1 Data Usage4

2.2 Data Cleaning4

2.3 Data Quality4

2.4 Data Analysis4

2.5 Data Profiling9

**3. Results10**

3.1 Reflections10

1. **Introduction**

This report outlines the findings of possible associative relationships within the stroke registry data set. These relationships may signal that these medications are able to better prevent a stroke from happening again.

* 1. **Purpose**

The purpose of the analysis undertaken is to find possible relationships between Aspirin and re-strokes, and determine if taking Aspirin will lower the possibility of having a re-stroke. Re-strokes are what all stroke patients fear the most after being taken out of care, so prevention of a re-stroke using medications such as Aspirin are their largest concern.

* 1. **Hypothesis**

If there is association between Aspirin and re-strokes, then taking Aspirin will result in lower chances of re-stroke.

* 1. **Key Documents**

Verro, Piero, et al. “Aspirin Plus Dipyridamole Versus Aspirin for Prevention of Vascular Events After Stroke or TIA” Stroke, American Heart Association, Inc., 24 Mar. 2008, stroke.ahajournals.org/content/39/4/1358.

* 1. **Raw Data**

The raw data given had several fields not used for our research. In the updated Excel file, I have removed these fields, since they do not apply for my research. The fields used are: Aspirin before admission, Aspirin during admission, Aspirin upon discharge, re-stroke at 3-month follow-up and re-stroke at 6-month follow-up. In general, this data was relatively easy to use for my research. Most data were complete and did not require any cleaning. Some re-stroke fields were incomplete, therefore the rows which had no follow-ups or unclear entries were completely removed from the data. Apart from that, the raw data required minimal cleaning. The only main issue with the data was the lack of positive re-stroke entries. This lead to the sample size of positive re-strokes to be very small.

Statistical analysis of the raw data can be found in the analysis sheet of the excel file attached. These include frequencies and percentiles of all fields used for my research, including all the Aspirin fields. Patients using Aspirin before admission was 33.04%, which raised to 69.13% during admission, and fell slightly back down to 62.69% upon discharge. At 3-month follow-ups, only 0.2% of all patients recorded had re-stroke. This raised to 0.71% at the 6-month follow-up. In total, out of the 5063 follow-up appointments, only 23 had reported re-strokes. This is a mere 0.45% of all follow-ups.

1. **Methods**
   1. **Data Usage**

Due to the nature of our hypothesis, only a few fields from the dataset are considered. These fields are: Aspirin before admission, Aspirin during admission, Aspirin upon discharge, re-stroke at 3-month follow-up and re-stroke at 6-month follow-up. Other fields are not related to the hypothesis. These fields can be used to answer the research question since they show all stages of using Aspirin and when re-strokes have occurred. The other fields did not relate to our research, and therefore were not included.

* 1. **Data Cleaning**

The fields regarding Aspirin in the dataset did not require cleaning since all fields were complete with either true or false records. However, the re-stroke fields required some cleaning due to incomplete or unknown values. Rows were completely removed from the data if they did not relate to the hypothesis, or entries were not clear. Rows where values were incomplete in both fields were also removed. However, rows where one follow-up record was blank while the other was not were included, since they may have not had the other follow-up. These cleaning steps should help the research be more accurate in terms of representing what happened. After cleaning the data, it is ready to be used for analysis and only relevant data is used for proving the research question.

* 1. **Data Quality**

There existed only two main issues with the data collected. Firstly, the data had inconsistent values for the re-stroke fields. The entries which were unclear had to be removed. There were also a lot of missing entries in the re-stroke fields, which had to be accounted for. Missing values can mean many things, perhaps the patient did not go to their follow-up appointments and it is then unknown if they had a stroke or not. If there was no entry in the 3-month follow-up appointment, but there is a “no stroke” entry in the 6-month follow-up then it can be assumed that there was no-stroke for the first three months too. Therefore, this data could be used. However, if it is reversed, it cannot be assumed that an empty 6-month follow-up entry means there is no re-stroke. Hence, entries with no re-stroke at the 3-month follow-up and empty 6-month follow-ups had to be removed. Secondly, there only existed 23 cases of re-stroke out of the 5063 usable cases within the data. This was only 0.45% of the total responses in follow-ups. While this is fortunate for the patients involved, it leads to a very small sample size when it comes to performing statistical research on the effects of Aspirin on re-stroke rates. A smaller sample size increases the chances that our research will not properly represent reality.

* 1. **Data Analysis**

All data being analyzed for this hypothesis is categorical, therefore frequency counts and percentile were calculated. Group comparisons of these variables will be made by using the chi-square test, to determine if there is association between Aspirin and re-stroke.

Null Hypothesis: There is no association between Aspirin and re-stroke

Alternative Hypothesis: There is association between Aspirin and re-stroke

Significance value: 0.05

To test if there is any possible association with Aspirin and re-stroke, we will have to perform multiple chi-square tests to see if Aspirin taken before admission, during, or upon discharge has an association with re-stroke rates at 3-month follow-ups or at 6-month follow ups. For each of these cases, a crosstabulation will be generated. It shows the actual count and the expected count for each pair of fields. There will also be Chi-Square Tests, which show the degree of freedom, and the p-value. If the p-value is greater than 0.05, our null hypothesis is accepted. If it is less than 0.05, the null hypothesis is rejected, and our alternative hypothesis will be accepted instead. In order to perform these tests in SPSS, the entries in 6-month follow-up for strokes where they could either be ICH, infarct or unknown mechanism of stroke were all change to simply “restroke”.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Aspirin\_Before\_Admission \* Restroke at 3 mth FU Crosstabulation** | | | | | | | | | | |
|  | | | | | | Restroke at 3 mth FU | | | | Total |
|  | | Infarct | No restroke |
| Aspirin\_Before\_Admission | | 0 | | Count | | 712 | | 2 | 982 | 1696 |
| Expected Count | | 730.5 | | 3.3 | 962.2 | 1696.0 |
| 1 | | Count | | 379 | | 3 | 455 | 837 |
| Expected Count | | 360.5 | | 1.7 | 474.8 | 837.0 |
| Total | | | | Count | | 1091 | | 5 | 1437 | 2533 |
| Expected Count | | 1091.0 | | 5.0 | 1437.0 | 2533.0 |
| **Chi-Square Tests** | | | | | | |
|  | Value | | df | | Asymptotic Significance (2-sided) | |
| Pearson Chi-Square | 4.297a | | 2 | | .117 | |
| Likelihood Ratio | 4.170 | | 2 | | .124 | |
| N of Valid Cases | 2533 | |  | |  | |
| a. 2 cells (33.3%) have expected count less than 5. The minimum expected count is 1.65. | | | | | | |

Aspirin Before Admission and Re-stroke at 3-month follow-up: Degrees of freedom: 2

As seen under the chi-square tests table, 2 cells have expected count less than 5. This means that we cannot use the Pearson Chi-Square results, since the expected count less than 5 is above 20%. Instead, we must use the likelihood ratio. The resultant p-value is .124, which is greater than our significance value of 0.05, meaning that the null hypothesis is accepted, and there is no association between Aspirin before admission and re-stroke at 3-month follow-up.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Aspirin\_During\_Admission \* Restroke at 3 mth FU Crosstabulation** | | | | | | |
|  | | | Restroke at 3 mth FU | | | Total |
|  | Infarct | No restroke |
| Aspirin\_During\_Admission | 0 | Count | 342 | 4 | 436 | 782 |
| Expected Count | 336.8 | 1.5 | 443.6 | 782.0 |
| 1 | Count | 749 | 1 | 1001 | 1751 |
| Expected Count | 754.2 | 3.5 | 993.4 | 1751.0 |
| Total | | Count | 1091 | 5 | 1437 | 2533 |
| Expected Count | 1091.0 | 5.0 | 1437.0 | 2533.0 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Chi-Square Tests** | | | |
|  | Value | df | Asymptotic Significance (2-sided) |
| Pearson Chi-Square | 5.960a | 2 | .051 |
| Likelihood Ratio | 5.443 | 2 | .066 |
| N of Valid Cases | 2533 |  |  |
| a. 2 cells (33.3%) have expected count less than 5. The minimum expected count is 1.54. | | | |

Aspirin During Admission and Re-stroke at 3-month follow-up: Degrees of freedom: 2

Once again, two cells have expected count less than 5, and having 33.3% is greater than 20% expected count, therefore we cannot use Pearson Chi-Square, and must once again use the Likelihood Ratio. The p-value of 0.66 is greater than our significance value of 0.05, so the null hypothesis is accepted, and there is no association between Aspirin during admission and re-stroke at 3-month follow-up.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Aspirin\_Upon\_Discharge \* Restroke at 3 mth FU Crosstabulation** | | | | | | |
|  | | | Restroke at 3 mth FU | | | Total |
|  | Infarct | No restroke |
| Aspirin\_Upon\_Discharge | 0 | Count | 380 | 5 | 560 | 945 |
| Expected Count | 407.0 | 1.9 | 536.1 | 945.0 |
| 1 | Count | 711 | 0 | 877 | 1588 |
| Expected Count | 684.0 | 3.1 | 900.9 | 1588.0 |
| Total | | Count | 1091 | 5 | 1437 | 2533 |
| Expected Count | 1091.0 | 5.0 | 1437.0 | 2533.0 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Chi-Square Tests** | | | |
|  | Value | df | Asymptotic Significance (2-sided) |
| Pearson Chi-Square | 12.963a | 2 | .002 |
| Likelihood Ratio | 14.438 | 2 | .001 |
| N of Valid Cases | 2533 |  |  |
| a. 2 cells (33.3%) have expected count less than 5. The minimum expected count is 1.87. | | | |

Aspirin Upon Discharge and Re-stroke at 3-month follow-up: Degrees of freedom: 2

Once again, two cells have expected count less than 5, and having 33.3% is greater than 20% expected count, therefore we cannot use Pearson Chi-Square, and must once again use the Likelihood Ratio. The p-value of 0.001 is much less than the significance value of 0.05, so the null hypothesis is rejected and there is strong association between Aspirin Upon Discharge and Re-stroke at 3-month follow-up.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Aspirin\_Before\_Admission \* Restroke\_at\_6\_mth\_FU Crosstabulation** | | | | | | |
|  | | | Restroke\_at\_6\_mth\_FU | | | Total |
|  | No restroke | restroke |
| Aspirin\_Before\_Admission | 0 | Count | 540 | 1148 | 8 | 1696 |
| Expected Count | 531.6 | 1152.3 | 12.1 | 1696.0 |
| 1 | Count | 254 | 573 | 10 | 837 |
| Expected Count | 262.4 | 568.7 | 5.9 | 837.0 |
| Total | | Count | 794 | 1721 | 18 | 2533 |
| Expected Count | 794.0 | 1721.0 | 18.0 | 2533.0 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Chi-Square Tests** | | | |
|  | Value | df | Asymptotic Significance (2-sided) |
| Pearson Chi-Square | 4.570a | 2 | .102 |
| Likelihood Ratio | 4.284 | 2 | .117 |
| N of Valid Cases | 2533 |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 5.95. | | | |

Aspirin Before Admission and Re-stroke at 6-month follow-up: Degrees of freedom: 2

0 cells have expected count less than 5, so Pearson Chi-Square can be used in this case. The p-value of 0.102 is greater than the significance value of 0.05, so the null hypothesis is accepted, and it has been found that there is no association between Aspirin before admission and re-stroke at 6-month follow-up.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Aspirin\_During\_Admission \* Restroke\_at\_6\_mth\_FU Crosstabulation** | | | | | | |
|  | | | Restroke\_at\_6\_mth\_FU | | | Total |
|  | No restroke | restroke |
| Aspirin\_During\_Admission | 0 | Count | 281 | 498 | 3 | 782 |
| Expected Count | 245.1 | 531.3 | 5.6 | 782.0 |
| 1 | Count | 513 | 1223 | 15 | 1751 |
| Expected Count | 548.9 | 1189.7 | 12.4 | 1751.0 |
| Total | | Count | 794 | 1721 | 18 | 2533 |
| Expected Count | 794.0 | 1721.0 | 18.0 | 2533.0 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Chi-Square Tests** | | | |
|  | Value | df | Asymptotic Significance (2-sided) |
| Pearson Chi-Square | 12.318a | 2 | .002 |
| Likelihood Ratio | 12.375 | 2 | .002 |
| N of Valid Cases | 2533 |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 5.56. | | | |

Aspirin During Admission and Re-stroke at 6-month follow-up: Degrees of freedom: 2

0 cells have expected count less than 5, so Pearson Chi-Square can be used in this case. The p-value of 0.002 is much less than the significance value of 0.05, so the null hypothesis is rejected, and the alternative hypothesis is accepted. There is association between Aspirin during admission, and re-stroke at 6-month follow-up.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Aspirin\_Upon\_Discharge \* Restroke\_at\_6\_mth\_FU Crosstabulation** | | | | | | |
|  | | | Restroke\_at\_6\_mth\_FU | | | Total |
|  | No restroke | restroke |
| Aspirin\_Upon\_Discharge | 0 | Count | 352 | 589 | 4 | 945 |
| Expected Count | 296.2 | 642.1 | 6.7 | 945.0 |
| 1 | Count | 442 | 1132 | 14 | 1588 |
| Expected Count | 497.8 | 1078.9 | 11.3 | 1588.0 |
| Total | | Count | 794 | 1721 | 18 | 2533 |
| Expected Count | 794.0 | 1721.0 | 18.0 | 2533.0 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Chi-Square Tests** | | | |
|  | Value | df | Asymptotic Significance (2-sided) |
| Pearson Chi-Square | 25.499a | 2 | .000 |
| Likelihood Ratio | 25.369 | 2 | .000 |
| N of Valid Cases | 2533 |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 6.72. | | | |

Aspirin Upon Discharge and Re-stroke at 6-month follow-up: Degrees of freedom: 2

0 cells have expected count less than 5, so Pearson Chi-Square can be used in this case. The p-value of 0.000 is much less than the significance value of 0.05, so the null hypothesis is rejected, and the alternative hypothesis is accepted. There is association between Aspirin upon discharge, and re-stroke at 6-month follow-up.

Summary:

In summary, there is no association between Aspirin before admission and re-stroke at follow-ups. There is association between Aspirin during admission and re-stroke at 6-month follow-up, but not 3-month follow-up. There is association between Aspirin upon discharge and re-stroke at follow-ups.

* 1. **Data Profiling**

For this analysis, column profiling was used to show the distribution of variables over the fields used. In the analysis sheet of the data excel file, there are several tables and charts which display the frequency and percentile of the variables in the fields. This includes the frequency and percentile of all Aspirin fields, and all re-strokes. An additional summary is included for the two follow-up fields combined, to show the number of total re-strokes, and the percentile.

1. **Results**

Our analysis finds that there is some association between Aspirin and re-stroke rates. Our findings suggest that there is no association between Aspirin before admission and re-stroke follow-ups, and there is no association between Aspirin during admission and 3-month re-stroke follow-ups. This means that if a patient takes Aspirin before admission, it will not affect their chances of having re-stroke at 3 or 6-month follow-ups. If a patient takes Aspirin during admission, it will not affect their chances of having re-stroke at their 3-month follow-up. However, it was found that there is association between Aspirin during admission and re-stroke at 6-month follow-up, as well as Aspirin upon discharge and re-stroke at follow-up. This suggests that taking Aspirin during admission will lessen the chances of re-stroke at 6-month follow-ups, and taking Aspirin upon discharge will lessen the chances of re-stroke at 3 and 6-month follow-ups.

These findings are important for stroke patients globally, as the biggest fear they can have is a possible re-stroke. The findings of this project also align with that of the cited paper, which found that there is a significant reduction in overall risk ratio in favor of aspirin plus dipyridamole for stroke alone. Hopefully this project can support medical decisions to give Aspirin to patients during admission and upon discharge, to lessen the chances of a re-stroke.

* 1. **Reflections**

When reflecting upon this project I do not have many suggestions to make to improve the project. There were minimal issues handling data, analysis, cleaning and profiling in the project. The only issue was the sample size of positive re-strokes. If there were more positive re-stroke entries, then the project could draw conclusive results much more confidently.