A logo with text on it

Description automatically generated

**MH6321**

**STATISTICAL MODELLING & DATA ANALYSIS**

**Project Report**

Our Professor: Prof. Xiang Liming

Team members:

|  |  |
| --- | --- |
| **NAME** | **MATRICULATION NUMBER** |
| Ang Shu Wei | G2302794C |
| Mai Xiyang | G2404171F |
| Singh Shivang | G2404095K |

1. **Background**

There have been many studies on whether drugs can help increase the survival rates in HIV patients. One of the earliest drugs used in combating HIV and AIDS is AZT (zidovudine) which is an antiretroviral drug specifically used to treat HIV infection. While AZT was groundbreaking in the early days of HIV treatment, there are limitations including side effects and development of drug resistance. Thus, the introduction of other drugs such as zalcitabine (ddC) and didanosine (ddI) as second-line therapies for HIV patients who had failed or interolerant of AZT therapy started as promising alternatives to improve treatment outcomes.

A few reasons why we are proposing research questions to investigate using this dataset.

1. Dynamic drug effects: Drug effects can be complex and may vary over time. Some drugs can be complex and may vary over time. Some drugs may have immediate effects.
2. Clinical implications:

* Early intervention: Identifying when a drug is losing effectiveness can lead to early intervention with alternative treatments, potentially improving patient outcomes.
* Safety monitoring: Monitoring changes in drug effects over time can help identify potential safety concerns or adverse events.

1. **Dataset**

The dataset we are using for our project is the **aids** dataset. It is a randomized clinical trial in which both longitudinal and survival data were collected to compare the efficacy and safety of two antiretroviral drugs (ddC/ddl) in treating patients who had failed or were intolerant of zidovudine (AZT) therapy.

A data frame with **1405** **observations** on the following **12 variables**. The table description of the variables can be found in Table 1 of the Appendix. There are no missing values observed in the dataset.

**Data Exploration**

**2.1 Descriptive Statistics**

The descriptive statistics for continuous and discrete variables are presented in the Table 2 and Table 3 of Appendix respectively. There are 3 continuous variables and 9 discrete variables. Out of the 9 discrete variables, there are 2 binary variables and 4 categorical variables with 2 levels as seen in Table 3 and Table 4 respectively. Obstime is a categorical variable with 5 levels.

**2.2 Visualization**

To understand more about the dataset, we explored different visualizations from histogram plot to understand the distribution of the variable of our interest such as CD4 count to the breakdown of facet plots showing different groups like by gender and by AZT and so on.

From the histogram diagram shown below, we can see that CD4 is a right-skewed distribution. Thus, we may consider applying transformations like log transformations to make the distribution more symmetric or we may want to use non-parametric tests rather than parametric tests that assume normality.

A graph of a number of bars

Description automatically generated

1. **Study problem questions**
2. Does the effect of a drug on the occurrence of an event (i.e. death during the study period) change over time?
3. Does the effect of a drug on the occurrence of an event change over time, and how are these effects influenced by gender, AZT and previous opportunistic infections (prevOIAIDS)?
4. How does CD4 cell count change over time for patients on ddC and ddI?
5. Do factors like gender, prevOI, and AZT factors influence CD4 cell count changes?
6. Does the rate of survival increase with the use of one medicine over another?
7. Is there a difference in survival between male and female patients?
8. Do patients with a previous AIDS diagnosis have a higher risk of death compared to those without?
9. **Methods**

The model we use for study research questions A and B is Generalized Estimating Equations (GEE). As the dataset involves clustered data whereby the patients within each cluster are correlated since same patient is measured repeatedly over time on CD4 count. And binomial distribution is used to model the outcome variable when the response (i.e. event) is binary (e.g., 0 or 1). Since the obstime values are 0, 2, 6, 12, 18, we have run the GEE model with the original dataset and run the same GEE model after removing the obstime values of 2 so that the repeated measures are equally spaced at 6 months.

The correlation structure may be appropriate to use ar1 because repeated measures taken at obstime at equally spaced times work best using ar1. To pick the best model, we run the GEE models with each of the correlation structure (ar1, unstructured, and exchangeable) to see which one gives the lowest QIC score which signifies the best fit of the model.

Similarly for study research questions C and D we also picked GEE. As we are looking at the CD4 counts of which the distribution is slightly skewed but due to central limit theorem since the sample size is large enough, the distribution tends to a normal distribution, thus we can still apply the gaussian distribution in the model.

The model we use for study research questions E to G is the Cox Proportional Hazards (Cox PH) model because it is a powerful statistical tool widely used in survival analysis. And it is well-suited for analyzing time-to-event data, where the primary interest is in the time it takes for a specific event to occur. Thus, we used the Cox proportional hazards model on **aids.id** dataset to evaluate the effects of several factors on patient survival time, where the response variable is time to death or censoring, and covariates include drug type, gender, previous AIDS diagnosis, and AZT intolerance status. This aids.id dataset only shows the 1st visit of each patient.

**Analysis/Results**

1. **Does the effect of a drug on the occurrence of an event (i.e. death during the study period) change over time?**

* **QIC Score for** (ar1/ unstructured/ exchangeable): 762.567/ 762.31/ 762.42

From Table 12, it shows that the QIC score of correlation structure ar1 has improved from **1067.64** from the original dataset to **762.567** after removing obstime values of 2. As such, for research question A and B which includes obstime factor, we will be using the dataset without obstime values of 2 to proceed with our analysis. And out of the 3 correlation structures, although the QIC for the model with correlation structure of unstructured is the lowest at **762.31**, the difference is very minor as compared to the correlation structure of ar1 at **762.567** thus we will continue to use ar1 as our correlation structure in the model.

For the first study research question A, we only look at the factors such as drug, obstime values,and the interaction of drug and obstime values, and from Table 13 it shows that when obstime is at 6 months and the interaction of drugs and obstime at 18 months is statistically significant on the effects on the event.

The negative coefficient of **-12.163386** for the interaction of drugs and obstime at 18 months suggests that individuals who received the drug drugddI and when obstime is at 18 months (i.e. had their CD4 cell counts taken at 18 months) are significantly less likely to experience the outcome (i.e. death during the study period). Overall, from the anova Table, obstime and the interaction of obstime and drug is statistically significant since the p values are less than α = 0.05.

**Interpretation**: The GEE model suggests that the type of drug does not seem to have much impact on the event. However, the interaction of the drug and observation time at the 18-month time point shows that this time point is important in the outcome of the event which could mean that those who taken the drugs and survived till then are less likely to experience deaths. The 6-month time point is an important time point and intervention methods can be included to reduce the risk of the event happening.

1. **Does the effect of a drug on the occurrence of an event change over time, and how are these effects influenced by gender, AZT and previous opportunistic infections (prevOIAIDS)?**

The study research question B is an expansion of A because we included more variables such as gender, AZT and previous opportunistic infections (prevOIAIDS) to study how these variables affect the occurrence of the event.

From Table 14, the coefficient for gendermale is **-1.4782** this suggests that males have a significantly lower risk (e-1.4782 ≈ 0.229) **77.1%** less likely to experience the event as compared to females and males are statistically significant since the p value is 0.00035 (less than α = 0.01). However, overall males are only marginally statistically significant since p value is 0.087.

Individuals with a previous AIDS diagnosis are approximately **64.9%** more likely to experience the event compared to those without a previous diagnosis. This finding highlights the importance of considering prior medical history when assessing the risk of adverse outcomes in patients with HIV/AIDS.

Although in Table 14 the AZTfailure is not statistically significant, overall, in the anova Table AZT factor is statistically significant to the outcome of the event.

**Interpretation**: The GEE model suggests that all these factors (i.e. AZT, prevOIAIDS, and the interaction of the obstime with drug, gender and AZT) show that they are statistically significant to the outcome of the event. Similarly to the interpretation in Question A, the interaction of the drug, gender, and AZT with observation time at the 18-month time point shows that this time point is critical in determining the outcome of the event. As such, alternatives such as customized therapy targeting high risk group (such as those females with AZT failure and has previous opportunistic infection) might be more effective in reducing the risk of having the event.

1. **How does CD4 cell count change over time for patients on ddC and ddI?**

* **Intercept:** 3.1534
* **Coefficient for time:** 0.2455
* **Standard Error:** 1.339
* **QIC Score for (ar1/Unstructured/ exchangeable):** 32922.986/ 33272.84/ 32878.47

The estimate of the **intercept** (3.15341) represents the baseline CD4 count at time zero. The estimate for the effect of time (0.24553) is positive and highly significant (p<0.001). **drugddI** (0.70742) shows the effect of taking ddI compared to a reference category (ddC). This positive coefficient implies that ddI is associated with a higher CD4 count compared to the reference, and the significance of p=0.49 indicates this might not be statistically significant. The **interaction term between time and the ddI drug** (-0.01108) indicates a slight reduction in the effect of time on CD4 count when using ddI compared to the reference category. However, this effect is not statistically significant (p=0.89). Comparing the QIC scores, it indicates that correlation structure of **exchangeable** is the best GEE model.

**Interpretation:** The GEE model suggests that time is a significant factor in increasing CD4 counts, implying a general improvement for patients over the study period. The specific drug treatment (ddC vs ddI) and its interaction with time do not show significant differences, indicating that both drugs have similar effects on CD4 counts over the analyzed time frame.

1. **Do factors like gender, prevOI, and AZT factors influence CD4 cell count changes?**

* **Coefficient** for **time (gender/prevOI/AZT)**: 0.220/0.051/0.234
* **Coefficient** for **prevOIAIDS**: -5.437
* **Standard Error (gender/prevOI/AZT):** 1.34/1.2/1.26
* **QIC Score** for **(gender/prevOI/AZT):** 32883.70/27522.9/30944.72

The estimate for the effect of **time** (0.220/0.051/0.234) is positive, though its significance varies depending on the context (Only in model **AZT** is significant (p<0.01)). **prevOIAIDS** (-5.437) shows a significant negative impact on CD4 count levels and highly significant (p<0.01). Looking at the QIC scores, it indicates that GEE with **prevOI** factor is a better model as compared to the others.

**Interpretation:** Previous opportunistic infections have the most statistically significant impact on CD4 counts among the factors we analyzed. Time consistently shows a positive trend in influencing CD4 counts, while other interactions and direct effects like gender and AZT failure appear less impactful based on this analysis.

1. **Is there a significant difference in survival rates between the two drug groups (ddC vs. ddI)?**

* **Coefficient** for **drugddI**: 0.2652
* **Hazard Ratio (HR)**: 1.3037
* **95% Confidence Interval (CI)**: [0.7345, 2.3139]
* **p-value**: 0.3649

The hazard ratio of 1.3037 suggests that patients receiving **ddI** have a 30.37% higher risk of death compared to those receiving **ddC**. However, this result is not statistically significant (p = 0.3649), as the confidence interval (0.7345 to 2.3139) includes 1. Therefore, we do not have sufficient evidence to conclude that there is a significant difference in survival between the two drug groups.

**Interpretation**: The choice of drug (ddC vs. ddI) does not appear to significantly impact survival time in this study. Both drugs may be considered comparable in terms of their effect on survival, and the selection between them could be based on other considerations such as side effects or patient tolerance.

1. **Does gender significantly affect survival outcomes?**

* **Coefficient** for **gendermale**: -1.0617
* **Hazard Ratio (HR)**: 0.3459
* **95% Confidence Interval (CI)**: [0.1712, 0.6987]
* **p-value**: 0.00309

The hazard ratio for **male** patients compared to **female** patients is 0.3459, indicating that male patients have a **65.41% lower risk of death** than female patients. This effect is statistically significant (p = 0.00309), with a confidence interval that does not include 1, supporting the association between gender and survival.

**Interpretation**: Gender plays a significant role in survival outcomes, with male patients having a lower risk of death. This finding suggests that gender-related biological or behavioral factors might influence patient survival, highlighting the importance of considering gender in treatment and support strategies.

1. **How does a previous AIDS diagnosis affect survival?**

* **Coefficient** for **prevOInoAIDS**: -1.3936
* **Hazard Ratio (HR)**: 0.2482
* **95% Confidence Interval (CI)**: [0.09409, 0.6546]
* **p-value**: 0.00486

Patients with no previous AIDS diagnosis (**noAIDS**) have a hazard ratio of 0.2482 compared to those with a previous AIDS diagnosis, indicating a **75.18% lower risk of death**. This result is statistically significant (p = 0.00486), as the confidence interval (0.09409 to 0.6546) does not include 1.

**Interpretation**: Patients without a prior AIDS diagnosis at the start of the study have a significantly better survival prognosis than those with prior AIDS. This finding implies that early-stage intervention, before the development of AIDS-related opportunistic infections, may be crucial for improving survival outcomes. It underscores the potential benefit of prompt antiretroviral therapy in HIV-positive patients before the onset of AIDS.

1. **Conclusion**

A person diagnosed with AIDs has a lower chance of survival compared to someone without AIDs. Gender has marginal significance in determining the survival rate with males having higher chances of survival as compared to females. In contrast to the belief of drug influence on person survival, type of drugs is statistically insignificant in determining the mortality rate of a person. Time shows a positive trend in influencing CD4 count, suggesting gradual improvements, while a history of previous opportunistic infections significantly reduces CD4 count levels, highlighting the long-term impact of these infections on immune recovery.

**Appendices**

**Table 1: Description of the variables in the aids dataset**

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| patient | patients' identifier; in total there are 467 patients |
| Time | the time to death or censoring |
| death | a numeric vector with 0 denoting censoring and 1 death |
| CD4 | the CD4 cells count. The higher the CD4 cells count, the better the survival rate |
| obstime | the time points in terms of months (i.e. 0, 2, 6, 12, and 18) at which the CD4 cells count was recorded |
| drug | a factor with levels ddC denoting zalcitabine and ddI denoting didanosine |
| gender | a factor with levels female and male |
| prevOI | a factor with levels AIDS denoting previous opportunistic infection (AIDS diagnosis) at study entry, and noAIDS denoting no previous infection |
| AZT | a factor with levels intolerance and failure denoting AZT intolerance and AZT failure respectively |
| start | The start time of an interval during which an event may occur. In the context of survival analysis, it marks the time at which an individual begins to be at risk. The values for ’start’ is the same as the values for 'Obstime’ |
| stop | The end time of the interval. It is the time at which the individual is no longer at risk for the event (e.g., due to experiencing the event, censoring, or the end of the study) |
| event | Indicates whether the event of interest occurred. When death = 1 and stop time is the same as the Time => event is ‘1’ |

**Table 2: Descriptive Statistics of Continuous Variables**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **Std. Dev** | **Min.** | **1st Qu.** | **Median** | **3rd Qu.** | **Max.** |
| **Time** | 13.89 | 4.138 | 0.47 | 12.23 | 14.07 | 17.00 | 21.40 |
| **CD4** | 7.023 | 4.958 | 0.000 | 3.162 | 5.477 | 10.440 | 24.125 |
| **stop** | 8.412 | 5.511 | 0.470 | 2.000 | 6.000 | 12.070 | 21.40 |

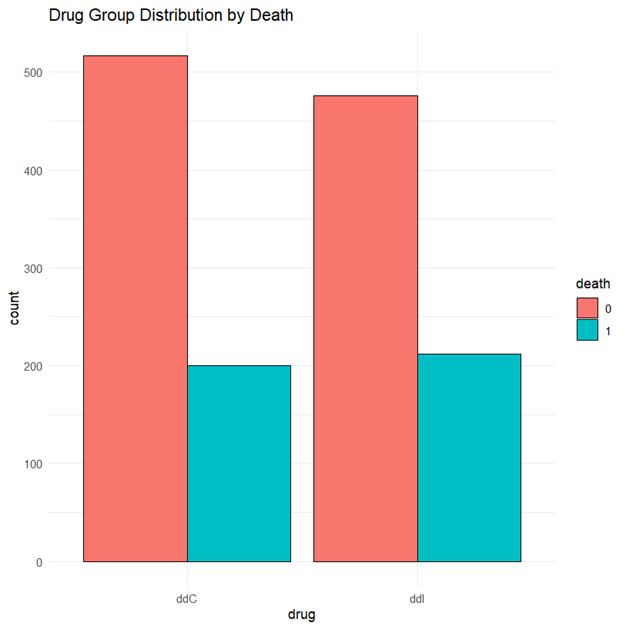
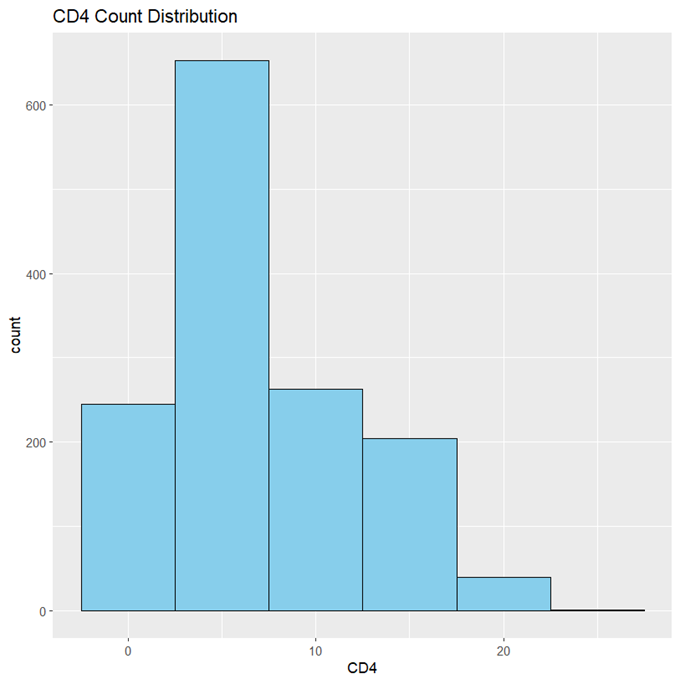
**Table 3: Descriptive Statistics of Discrete Variables that are binary**

|  |  |  |
| --- | --- | --- |
| **Variable** | **0** | **1** |
| **death** | 993 | 412 |
| **event** | 1217 | 188 |

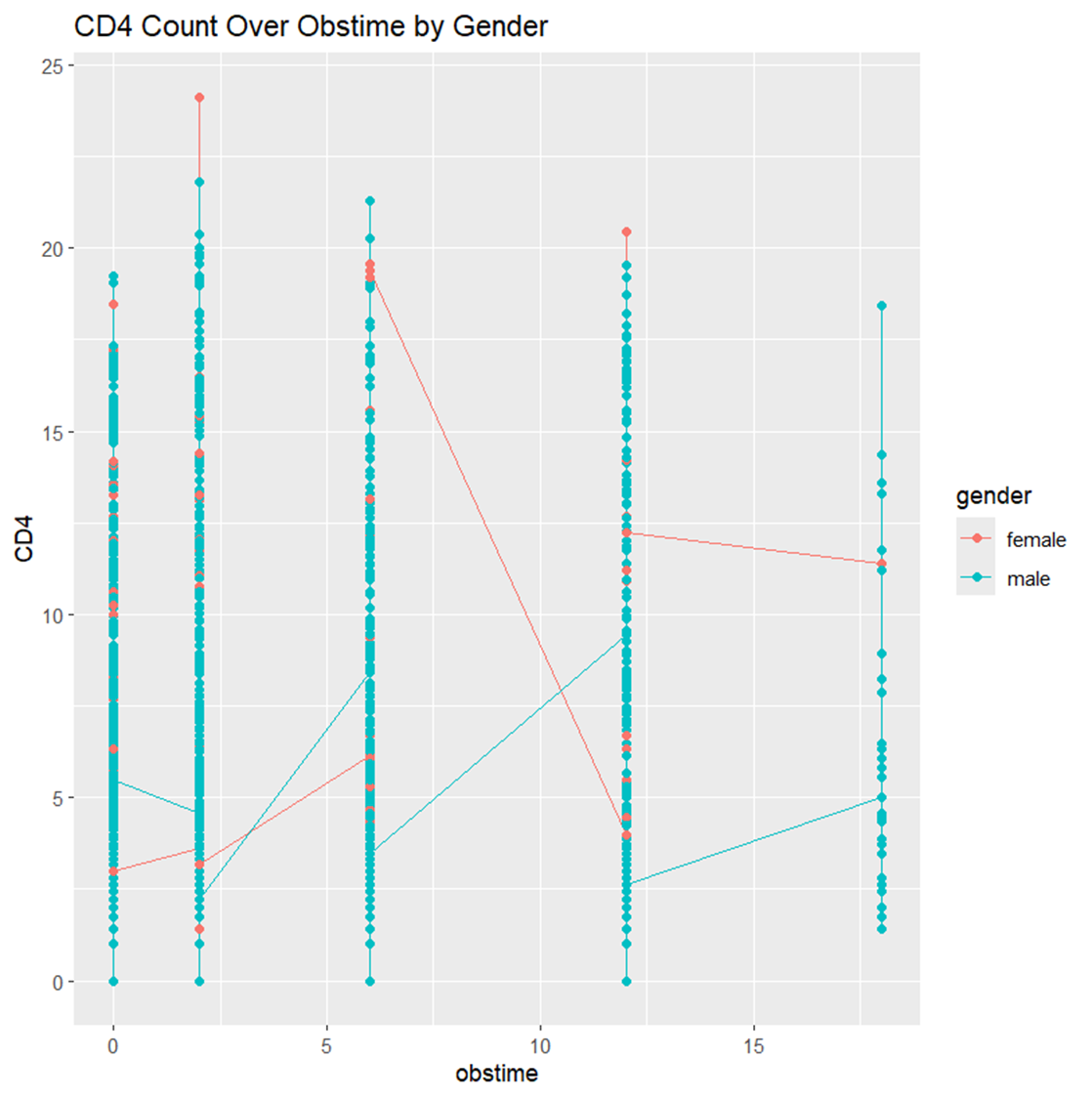
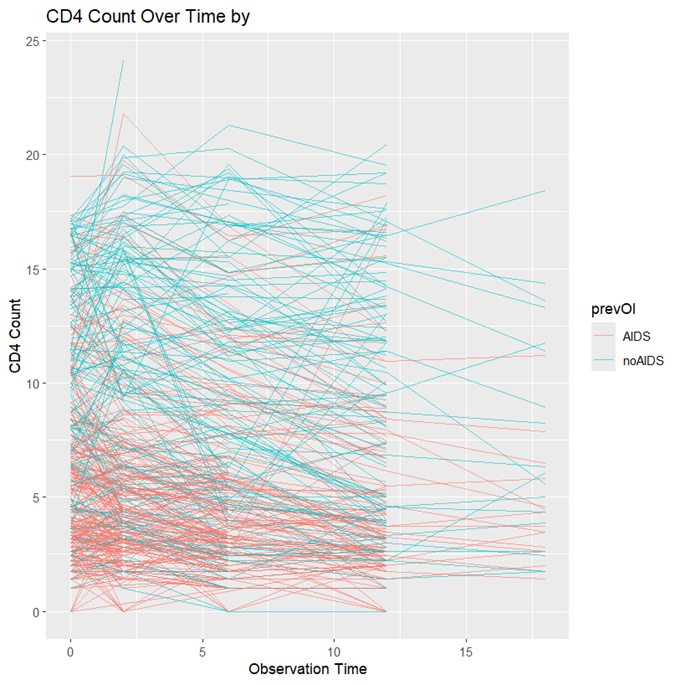
**Table 4: Descriptive Statistics of Categorical Variables that have 2 levels**

|  |  |  |
| --- | --- | --- |
| **Variable** | **0 (Reference Level)** | **1** |
| **drug** | ddC: 717 | ddI: 688 |
| **gender** | Female:117 | Male: 1288 |
| **prevOI** | AIDS: 863 | noAIDS: 542 |
| **AZT** | Failure: 491 | Intolerance: 914 |

**Table 5: CD4 Distribution Table 6: Drug Group Distribution by Death**



**Table 7: CD4 count Over Time by prevOIAIDS Table 8: CD4 Count Over Obstime by Gender**



**Table 9: Average CD4 count Over Obstime based on the drug type by Gender**

A graph of a number of people

Description automatically generated with medium confidence

**Table 10: Average CD4 count Over Obstime based on the drug type by AZT**

A graph of a number of red and blue bars

Description automatically generated

**Table 11: Average CD4 count Over Obstime based on the drug type by prevOIAIDS**

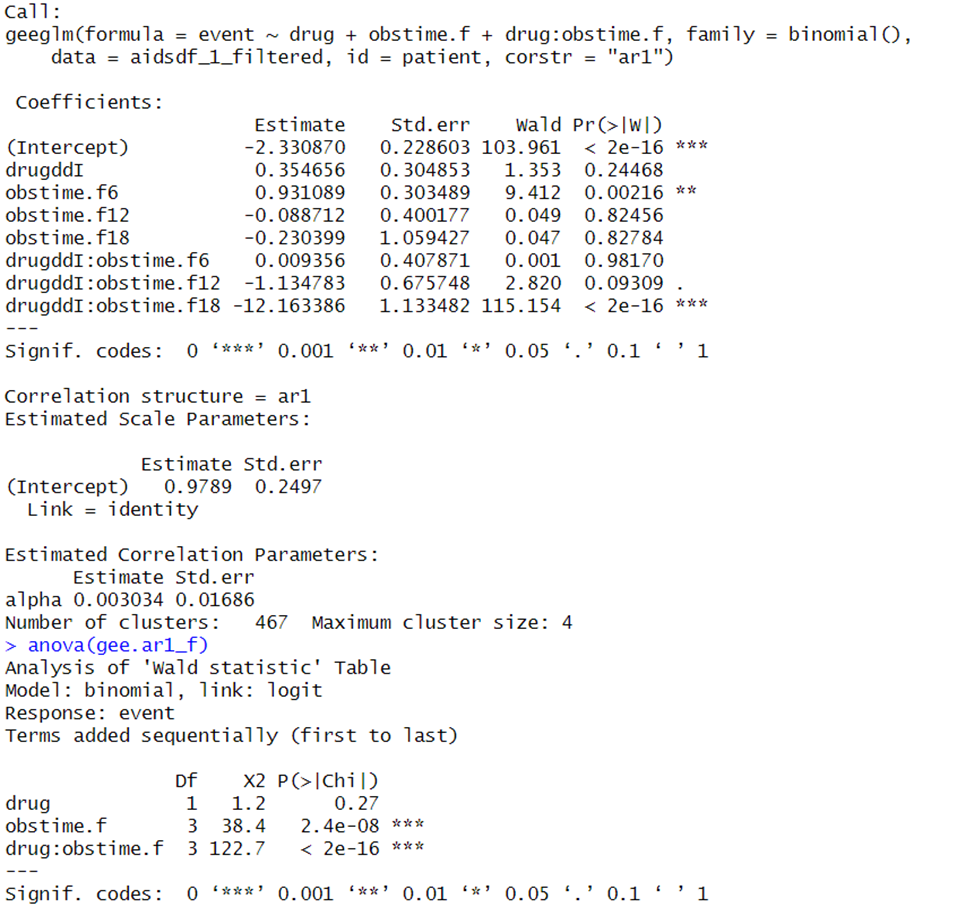
A graph of a number of red and blue bars

Description automatically generated

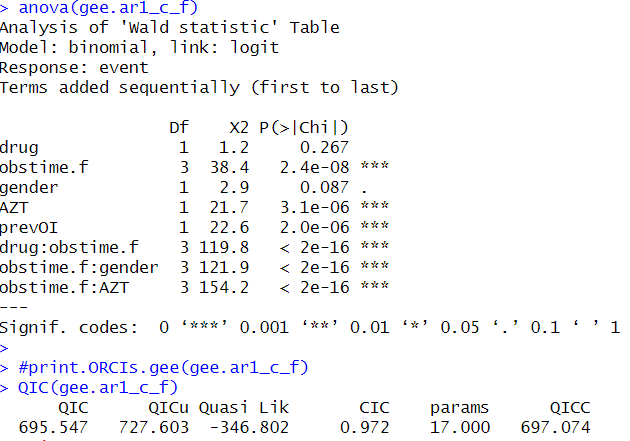
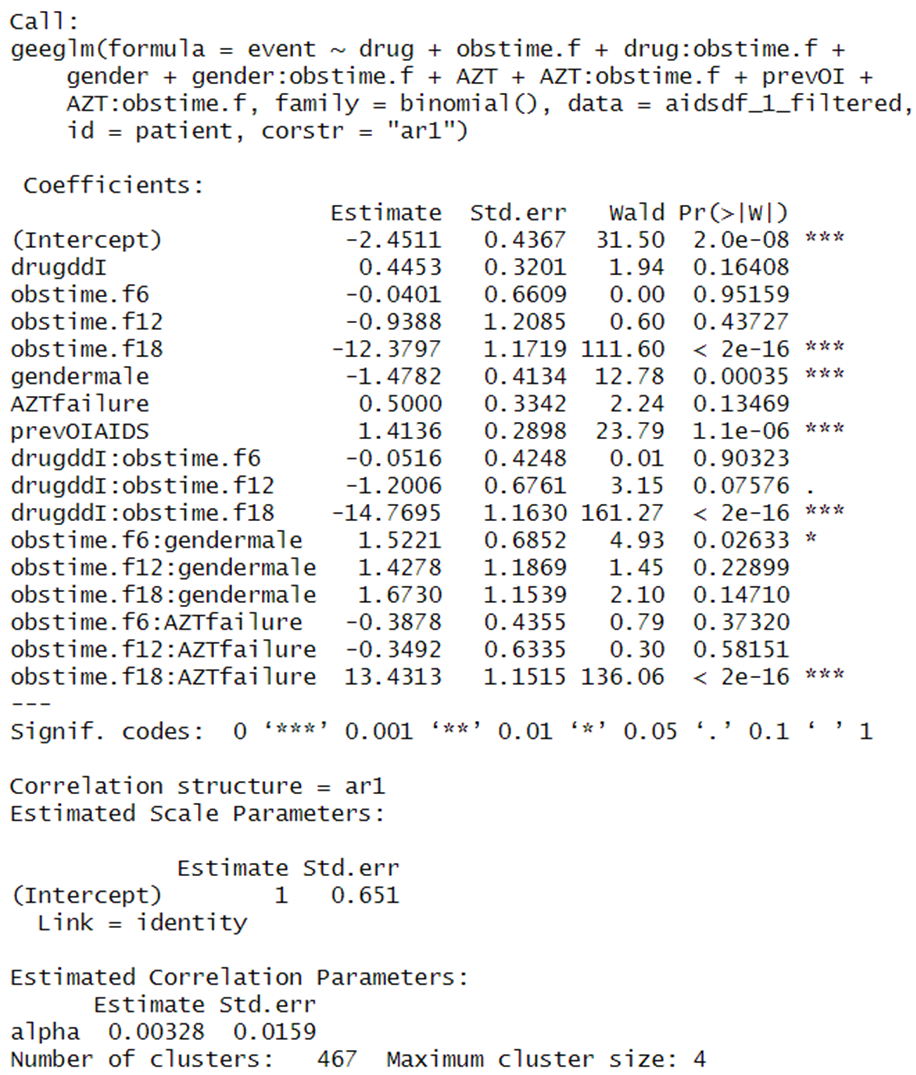
**Table 12: QIC scores for the 3 types of correlation structures used for the original dataset and the dataset without obstime =2 values respectively and the standard error in the GEE models.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **ar1** | **unstructured** | **exchangeable** |
| **Original dataset** | | | |
| QIC scores | 1067.64 | 1067.49 | **1067.27** |
| Std.err | 0.193 | 0.134 | 0.178 |
| **After removing rows where obstime=2** | | | |
| QIC scores | 762.567 | **762.31** | 762.42 |
| Std.err | 0.2497 | 0.212 | 0.243 |

**Table 13: GEE model output and ANOVA Table for Research question A**



**Table 14: GEE model output and ANOVA Table for Research question B**



**Table 15: GEE model output and ANOVA Table for Research question C**

文本

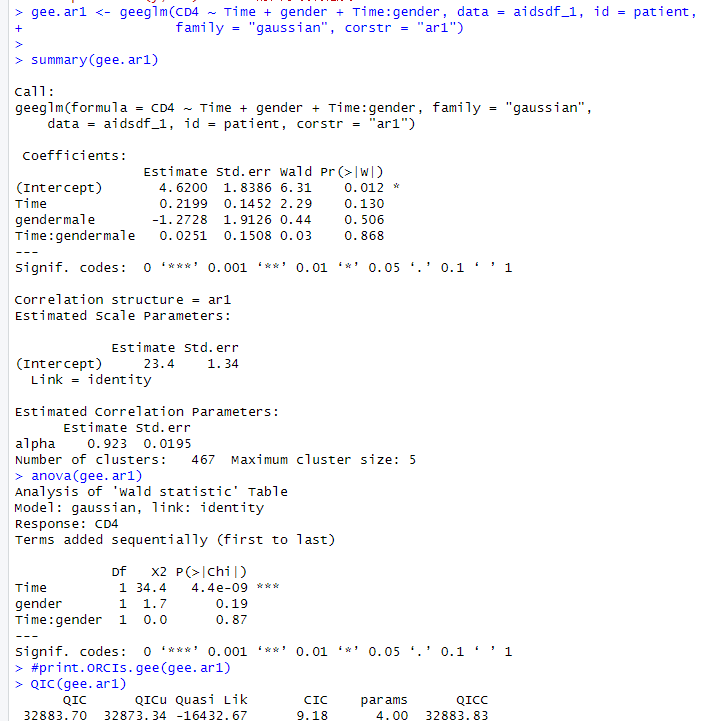
已自动生成说明文本

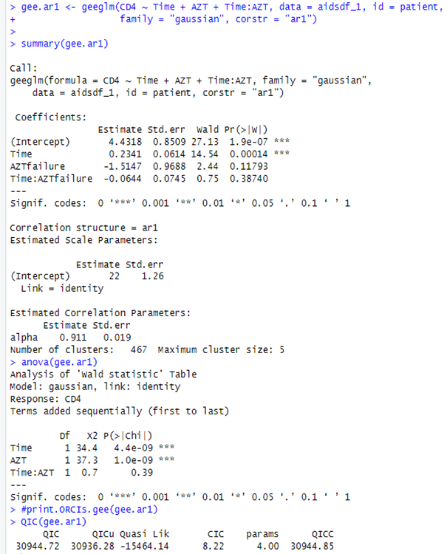
已自动生成说明

**Table 16: QIC scores for each factor (Gender, prevOI, AZT) in the GEE model for Research Question D**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Gender** | **prevOI** | **AZT** |
| QIC scores | 32883.70 | 27522.9 | 30944.72 |
| Std.err | 1.34 | 1.2 | 1.26 |

**Table 17: GEE models output and ANOVA Tables for Research question D**

**** **文本

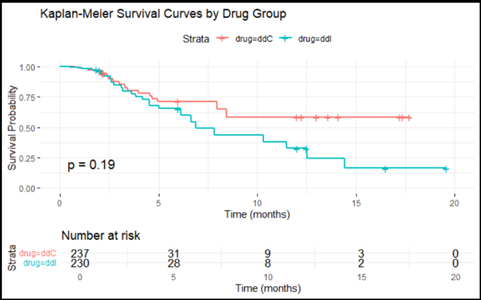
已自动生成说明**

**Table 18: KM Curves for survival analysis for Research question E to G**

A graph of a survival curve

Description automatically generatedA graph of a number of men and women

Description automatically generated



**Table 19: Cox Hazard Model Output for Research question E to G**

A screenshot of a computer

Description automatically generated