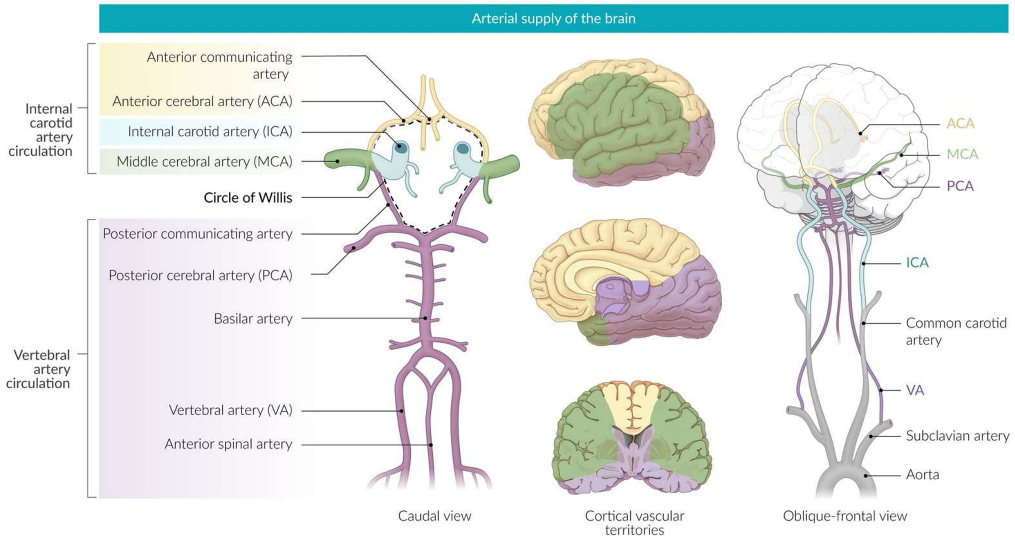
DATA 623 | Professor: Dr. Bakal
Marc McCoy (ID: 30136987) | MDSA 2022 & MD 2024

Analyzing the 1991-1996 International Stroke Trial Database:

Assigned Open-ended Research Question: Is there a difference in the outcome (acute ischemic stroke) between the aspirin and high heparin groups?

Background:

From the original study: The International Stroke Trial (IST) was a large, randomised, open trial of up to 14 days of antithrombotic therapy started as soon as possible after stroke onset. The aim was to provide reliable evidence on the safety and efficacy of aspirin and of subcutaneous heparin. Half the patients were allocated unfractionated heparin (5000 or 12,500 IU bd [twice daily]), and half were allocated "avoid heparin"; and, in a factorial design, half were allocated aspirin 300 mg daily and half "avoid aspirin". The primary outcomes were death within 14 days and death or dependency at 6 months.[1] 19,435 patients (467 hospitals, 36 countries) were randomised within 48 hours of symptom onset. Ischemic stroke is an acute neurological condition caused by impaired cerebral blood flow (vascular occlusion or systemic hypoperfusion). The most important risk factors for ischemic stroke are chronic systemic hypertension and cardiovascular disease.[2] Attributes capturing these risk factors are included within the randomisation data and include: *SEX, AGE, RATRIAL, and RSBP*. The inclusion criteria into the IST database were: clinical diagnosis of acute ischaemic stroke, with onset within the previous 48 hours and no clear indication for, or clear contraindication to, treatment with aspirin or subcutaneous heparin.[1] To sufficiently answer the assigned question, I decided to focus on the 14-day follow up period, as opposed to the 6-month follow up period. The initial IST was done between 1991 and 1996 to assess the impact of aspirin, heparin, both or neither on the clinical course of acute ischaemic stroke. I interpreted the original study to be assessing the efficacy of the medication on preventing future ischemic events without triggering potential haemorrhagic events.


Arterial supply of the brain
Internal carotid artery circulation
Anterior communicating artery
Anterior cerebral artery (ACA)
Internal carotid artery (ICA)
Middle cerebral artery (MCA)
Circle of Willis
Vertebral artery circulation
Posterior communicating artery
Posterior cerebral artery (PCA)
Basilar artery
Vertebral artery (VA)
Anterior spinal artery
ACA
MCA
PCA
ICA
Common carotid artery
VA
Subclavian artery
Aorta
Caudal view
Cortical vascular territories
Oblique-frontal view

Schematic one. Schematic showing the rough anatomical location of strokes corresponding to the *STYPE* categories within the initial randomisation data.

Data Source:

The IST database (doi:10.1186/1745-6215-12-101) included data on 19,435 patients with acute stroke. Background stroke care was limited and none of the patients received thrombolytic therapy. It was a large, prospective, randomised controlled trial, with 100% complete baseline data and over 99% complete follow-up data. *The aim of the trial was to establish whether early administration of aspirin, heparin, both or neither influenced the clinical course of acute ischaemic stroke.*[1]

Variables:

The following variables made up the selected subset of attributes that were utilized to answer the open-ended research question via a polytomous logistic regression model.

Final diagnosis of initial event (4 attributes)

Coded Variable	Description	Units	Variable Type
DDIAGISC	Ischaemic stroke	(Y/N/U=unknown)	Categorical
DDIAGHA	Haemorrhagic stroke	(Y/N/U=unknown)	Categorical
DDIAGUN	Indeterminate stroke	(Y/N/U=unknown)	Categorical
DNOSTRK	Not a stroke	(Y/N/U=unknown)	Categorical

Recurrent stroke within 14 days (3 attributes)

Coded Variable	Description	Units	Variable Type
DRSISC	Ischaemic recurrent stroke	(Y/N/U=unknown)	Categorical
DRSH	Haemorrhagic stroke	(Y/N/U=unknown)	Categorical
DRSUNK	Unknown type	(Y/N/U=unknown)	Categorical

Other events within 14 days (2 attributes)

Coded Variable	Description	Units	Variable Type
DPE	Pulmonary embolism	(Y/N/U=unknown)	Categorical
DDEAD	Dead on discharge form	(Y/N/U=unknown)	Categorical

Randomisation data (13 attributes)

Coded Variable	Description	Units	Variable Type
HOSPNUM	Hospital number	unitless	Categorical
RDELAY	Delay between stroke and randomisation	hours	Continuous
RCONSC	Conscious state at randomisation	(F - fully alert, D - drowsy, U - unconscious)	Categorical
SEX	Biological sex	M=male; F=female	Binary
AGE	Patient Age	years	Continuous
RSLEEP	Symptoms noted on waking	(Y/N)	Binary
RATRIAL	Atrial fibrillation	(Y/N)	Binary
RHEP24	Heparin within 24 hours prior to randomisation	(Y/N)	Binary
RASP3	Aspirin within 3 days prior to randomisation	(Y/N)	Binary
RSBP	Systolic blood pressure at randomisation	(mmHg)	Continuous
STYPE	Stroke subtype	(TACS/PACS/POCS/LACS/OTH=other)	Categorical
RXASP	Trial aspirin allocated	(Y/N)	Binary
RXHEP	Trial heparin allocated. The terminology for the allocated dose of unfractioned heparin changed slightly from the pilot to the main study. Patients were allocated either 12500 units subcutaneously twice daily (coded as H in the pilot and M in the main trial), 5000 units twice daily (coded as L throughout) or to 'avoid heparin' (coded as N throughout).	(M/L/N)	Categorical

Data collected on 14 day/discharge form about treatments given in hospital (13 attributes)

Coded Variable	Description	Units	Variable Type
DASP14	Aspirin given for 14 days or till death or discharge	(Y/N/U=unknown)	Categorical
DLH14	Low dose heparin given for 14 days or till death/discharge	(Y/N/U=unknown)	Categorical
DMH14	Medium dose heparin given for 14 days or till death/discharge	(Y/N/U=unknown)	Categorical
DSCH	Non trial subcutaneous heparin	(Y/N/U=unknown)	Categorical
DIVH	Non trial intravenous heparin	(Y/N/U=unknown)	Categorical
DAP	Non trial antiplatelet drug	(Y/N/U=unknown)	Categorical
DOAC	Other anticoagulants	(Y/N/U=unknown)	Categorical
DGORM	Glycerol or manitol	(Y/N/U=unknown)	Categorical
DSTER	Steroids	(Y/N/U=unknown)	Categorical
DCAA	Calcium antagonists	(Y/N/U=unknown)	Categorical
DHAEMD	Haemodilution	(Y/N/U=unknown)	Categorical
DCAREND	Carotid surgery	(Y/N/U=unknown)	Categorical
DTHROMB	Thrombolysis	(Y/N/U=unknown)	Categorical

Table 1. Subset of the original 112 coded attributes that were utilized to derive the final twenty-five attributes used within the final analysis dataframe (‘stroke.csv’).

The following variables remained after performing wrangling and cleaning operations on the original subset of attributes. The outcome variable is denoted by the yellow highlighting.

Remaining Attributes Post Wrangling (25)			
Coded Variable	Description	Units	Variable Type
RDELAY	Delay between stroke and randomisation	hours	Continuous
RCONSC	Conscious state at randomisation	(F - fully alert, D - drowsy, U - unconscious)	Categorical
SEX	Biological sex	M=male; F=female	Binary
AGE	Patient Age	years	Continuous
RSLEEP	Symptoms noted on waking	N=0, Y=1	Binary
RATRIAL	Atrial fibrillation	N=0, Y=1	Binary
RSBP	Systolic blood pressure at randomisation	(mmHg)	Continuous
STYPE	Stroke subtype	(TACS/PACS/POCS/LACS/OTH=other)	Categorical
DASP14	Aspirin given for 14 days or till death or discharge	N=0, Y=1	Binary
DLH14	Low dose heparin given for 14 days or till death/discharge	N=0, Y=1	Binary
DMH14	Medium dose heparin given for 14 days or till death/discharge	N=0, Y=1	Binary
DSCH	Non trial subcutaneous heparin	N=0, Y=1	Binary
DIVH	Non trial intravenous heparin	N=0, Y=1	Binary
DAP	Non trial antiplatelet drug	N=0, Y=1	Binary
DOAC	Other anticoagulants	N=0, Y=1	Binary
DGORM	Glycerol or manitol	N=0, Y=1	Binary
DSTER	Steroids	N=0, Y=1	Binary
DCAA	Calcium antagonists	N=0, Y=1	Binary
DHAEMD	Haemodilution	N=0, Y=1	Binary
DCAREND	Carotid surgery	N=0, Y=1	Binary
DTHROMB	Thrombolysis	N=0, Y=1	Binary
group	Treatment group at randomisation	N, A, H, AH	Categorical
medHx	No Medication, Aspirin within 3 days, Heparin within 24 hours, Aspirin within 3 days + Heparin within 24 hours	N, A, H, AH	Categorical
outcome	Outcome category within 14 days (No event, Unknown event, Ischemic event, Hemorrhagic event)	N, UE, IE, HE	Categorical

Table 2. Coded attributes remaining post wrangling, concatenation and grouping.

Analysis Plan:

To answer the posed question, I utilized a polytomous logistic regression model in RStudio after wrangling my data with Python in a conda environment. The polytomous logistic regression model is also known as a multiclass logistic regression, or a mlogit model. This type of model assumes that data are case-specific, which is intuitive in this context because each row within the database represents a separate patient. The main reason this model makes sense in this context is I set up the question so that the dependent variable (*outcome*) was categorical and nominal, but not ordinal; (no event, unknown event, ischemic event, haemorrhagic event). A benefit of this model is that there is no underlying assumption that the explanatory variables are independent from each other. If that assumption had to be made, it would be unrealistic, as several of the medications that may have been given in the initial 14-day period would be dependent on the randomization data from the initial patient presentation. For example, a patient with atrial fibrillation (*RATRIAL*) would receive a calcium antagonist (*DCAA*) as a preventative measure within the 14-day observation period. The concern of multicollinearity is still embedded within the model assumptions, which is addressed within the results section. Multinomial logistic regression is an attractive analysis in this setting as well because it does not assume normality, linearity, or homoscedasticity.

The likelihood function for the polytomous model utilizes the data involving all categories of the outcome variable in a single structure. A log-linear model is fitted with coefficients of zero for the first class (or those variables that have been set as a reference prior to running the model). Although this model creates an output that is not easily interpretable, it is great for facilitating “odds-like” expressions for comparisons. For example, P $(Y=1)/P(Y=0)$ …rather than odds as the following: P $(Y=1)$ / $[1-P$ $(Y=1)]$.

Wrangling:

The dataset originally contained 19,435 patients and 112 attributes. I wanted to determine clearly if there was a difference in outcome at 14 days between the aspirin and high heparin groups. I only utilized a subset of the available attributes: *DDIAGISC, DDIAGHA, DDIAGUN, DNOSTRK, DRSISC, DRSH, DRSUNK, DPE, DDEAD, RDELAY, RCONSC, SEX, AGE, RSLEEP, RATRIAL, RHEP24, RASP3, RSBP, STYPE, RXASP, RXHEP, DASP14, DLH14, DMH14, DSCH, DIVH, DAP, DOAC, DGORM, DSTER, DCAA, DHAEMD, DCAREND, DTHROMB*.

I first removed the pilot data (first 984 patients). I then removed all patients that were deemed to have had an initial diagnosis prior to randomisation of a haemorrhagic stroke, an indeterminate stroke, or "not" a stroke. I did this as the posed question and purpose of the original study was to establish whether early administration of aspirin, heparin, both or neither influenced the clinical course of *ischaemic stroke*. After filtering on these parameters, I subsequently dropped the *DDIAGISC*, *DDIAGHA*, *DDIAGUN*, and *DNOSTRK* attributes. As the question clearly states a focus on the *high* heparin group, I decided only include patients who got a high dose of heparin or no dose of heparin and therefore filtered rows in the *RXHEP* column that were equal to 'M' or 'N.' Patients in the high heparin group were allocated 12500 units subcutaneously twice daily (coded as M in the main trial). I then removed patients with an unknown *DRSISC* (ischemic recurrent stroke diagnosis) at 14 days. Of the recurrent stroke within fourteen days variables *DRSISC*, *DRSH*, and *DRSUNK* – only *DRSISC* had any patients with an "unknown" diagnosis.

Following this preliminary cleaning, I realized there was considerable amount of missing data, so I dropped the patients who had any missing data (NaN or null) within the thirty-one attributes I was analyzing. A graphical representation of the missing data is shown below.

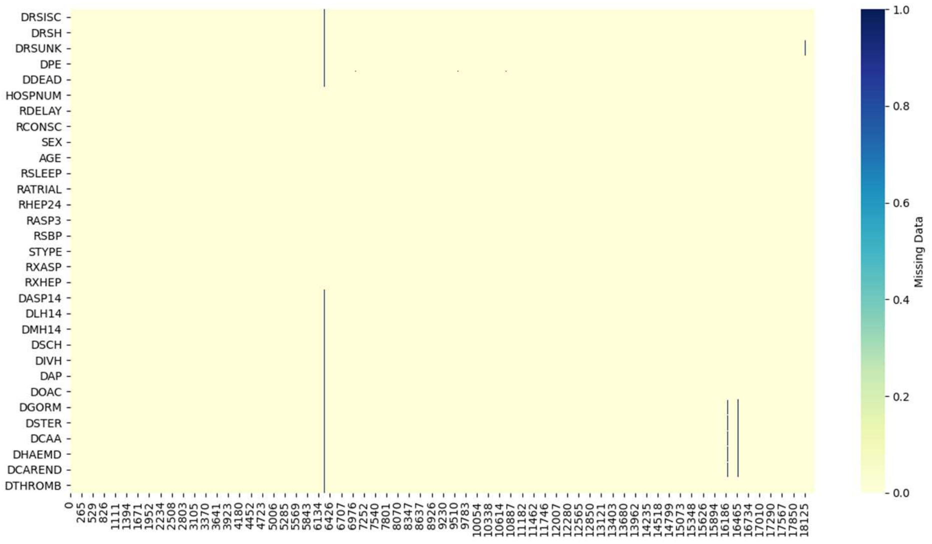
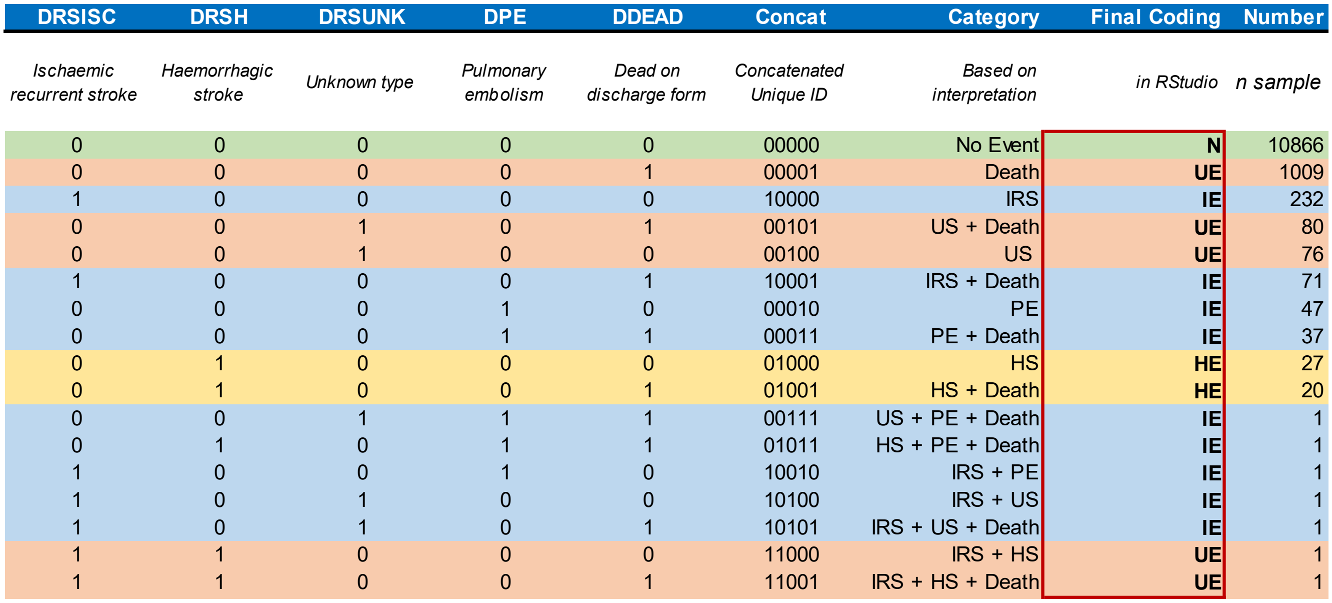


Figure 1. Missing data that was classified as NaN or null, after removing the 984 patients within the pilot study from 1991-1993, is shown in dark blue.

A haemorrhagic stroke and an ischemic stroke have fundamentally different pathophysiology. An ischemic stroke is an acute neurological condition caused by impaired cerebral blood flow. This most often is more of a "clotting event" as opposed to a "bleeding event" … whereby a thrombus may block an artery which is perfusing the brain.[2,3] A haemorrhagic stroke in contrast is an acute neurological condition characterized by hemorrhage within the brain parenchyma (intracerebral hemorrhage) or the subarachnoid space (subarachnoid hemorrhage). Resulting tissue compression and swelling are what drive clinical symptoms and long-term tissue damage.[4,5]

To use a polytomous logistic regression to model outcomes within the 14-day observation period, I had to create mutually exclusive and exhaustive categorical outcome levels that had no natural ordering. Keeping the pathophysiology in mind, I decided to categorize the outcome as being: *no event (N - green), an unknown event (UE - orange), an ischemic event (IE - blue), or a haemorrhagic event (HE - yellow)*. In the end, there were four categories which encompassed all outcomes within the 14-day observation period. This is shown in the table below. It is best to think of a bleeding event (haemorrhagic stroke) a HS, and a clotting event (pulmonary embolism, ischemic stroke) as an IE.



DRSISC	DRSH	DRSUNK	DPE	DDEAD	Concat	Category	Final Coding	Number
Ischaemic recurrent stroke	*Haemorrhagic stroke*	*Unknown type*	*Pulmonary embolism*	*Dead on discharge form*	*Concatenated Unique ID*	*Based on interpretation*	*in RStudio*	*n sample*
0	0	0	0	0	00000	No Event	**N**	10866
0	0	0	0	1	00001	Death	**UE**	1009
1	0	0	0	0	10000	IRS	**IE**	232
0	0	1	0	1	00101	US + Death	**UE**	80
0	0	1	0	0	00100	US	**UE**	76
1	0	0	0	1	10001	IRS + Death	**IE**	71
0	0	0	1	0	00010	PE	**IE**	47
0	0	0	1	1	00011	PE + Death	**IE**	37
0	1	0	0	0	01000	HS	**HE**	27
0	1	0	0	1	01001	HS + Death	**HE**	20
0	0	1	1	1	00111	US + PE + Death	**IE**	1
0	1	0	1	1	01011	HS + PE + Death	**IE**	1
1	0	0	1	0	10010	IRS + PE	**IE**	1
1	0	1	0	0	10100	IRS + US	**IE**	1
1	0	1	0	1	10101	IRS + US + Death	**IE**	1
1	1	0	0	0	11000	IRS + HS	**UE**	1
1	1	0	0	1	11001	IRS + HS + Death	**UE**	1

Figure 2. How the outcome variable was built via concatenation from the recorded outcomes within the 14-day observation period

Within the 14 days observation period, it was recorded if patients had an ischemic stroke (*DRSISC*), haemorrhagic stroke (*DRSH*), a stroke of an unknown type (*DRSUNK*), pulmonary embolism (*DPE*), and if a patient died within the 14 days post initial diagnosis (*DDEAD*). To analyze one response variable, I recoded these attributes from "Y" and "N" to 1 and 0. I then concatenated these five columns as objects (str), which returned seventeen unique combinations of outcomes shown above. Instead of using seventeen mutually exclusive and exhaustive outcomes within the polytomous logistic regression model, I classified the outcome as existing within one of four categories: *no event (N - green), unknown event (UE - orange), ischemic event (IE - blue), or a haemorrhagic event (HE - yellow)*. After concatenating the outcome variable and classifying it into one of the four groups, I dropped the attributes *DRSISC, DRSH, DRSUNK, DPE*, and *DDEAD*.

The next step was to assign the patients to one of the four groups that were created at randomisation. I recoded N as 0 and Y as 1 in the *RXASP* and *RXHEP* attributes and subsequently concatenated the values, which resulted in outcomes 00, 01, 10, 11 – the patients were then grouped into the '00' *no medication (N)*, '01' *high heparin (H)*, '10' *aspirin (A)*, and '11' *aspirin and high heparin (AH)* groups. The attributes *RXASP* and *RXHEP* were subsequently dropped.

The last grouping of interest revolved around the *RHEP24* and *RASP3* attributes which were binary variables that recorded if patients had received heparin within the 24 hours prior to their acute ischemic stroke or aspirin within the three days prior to their acute ischemic stroke. Via the same concatenation procedure, patients were placed into a *medHx* group of *no medication (N)*, *aspirin within 3 days prior (A)*, *heparin within 24 hours prior (H)*, and *both aspirin 3 days prior and heparin within 24 hours prior (AH)*. The *RHEP24* and *RASP3* attributes were subsequently dropped.

The decision was made to only include patients who were between the ages of 40 and 90 years of age. As patients younger than forty are unlikely to have a stroke that was not precipitated by unusual comorbidities. There were also very few patients under the age of forty and over the age of ninety.

All patients who had an unknown status ("U") for alternative medication or treatment within the hospital during the 14-day observation period were removed from the dataset. Alternative medication or treatments were coded by the following variables: *DASP14, DLH14, DMH14, DSCH, DIVH, DAP, DOAC, DGORM, DSTER, DCAA, DHAEMD, DCAREND, DTHROMB.* All remaining binary attributes were coded as N=0 and Y=1. As a last step, the datatypes of each remaining attribute were adjusted to either 64-bit int or string.

Results:

MODEL A:

I utilized RStudio to conduct my statistical analysis to see if there was a difference in the outcome (i.e., acute ischemic stroke) between the aspirin and high heparin groups. As can be seen below, I first load the wrangled stroke.csv dataset that I produced within a Jupyter Notebook. I then set the reference levels to no medication (N) for the group attribute, no event (N) for the outcome, no medication (N) for the medHx attribute, and total anterior circulation syndrome (TACS) for the STYPE attribute. Setting these references allowed me to interpret the results of the model more easily.

I then utilized the multinom function via the nnet library to fit a multinomial log-linear model using a neural network (polytomous logistic regression). Please see the documentation and explanation via the hyperlink.

```
#Load the data
stroke_df <- read.csv ("stroke.csv", header = TRUE)

#set reference levels
stroke_df$group <- as.factor(stroke_df$group)
stroke_df$group <- relevel(stroke_df$group, ref="N")

stroke_df$outcome <- as.factor(stroke_df$outcome)
stroke_df$outcome <- relevel(stroke_df$outcome, ref="N")

stroke_df$medHx <- as.factor(stroke_df$medHx)
stroke_df$medHx <- relevel(stroke_df$medHx, ref="N")

stroke_df$STYPE <- as.factor(stroke_df$STYPE)
stroke_df$STYPE <- relevel(stroke_df$STYPE, ref="TACS")

modelA <- multinom (outcome~group+medHx+STYPE, data = stroke_df) #this is my model

## # weights:  48 (33 variable)
## initial value 16812.978012
## iter  10 value 5657.628980
## iter  20 value 5497.335528
## iter  30 value 5463.866589
## iter  40 value 5416.412962
## final  value 5413.810601
## converged
```

```
summary(modelA)
```

```
## Call:
## multinom(formula = outcome ~ group + medHx + STYPE, data = stroke_df)
##
## Coefficients:
##    (Intercept)      groupA     groupAH      groupH      medHxA    medHxAH
## HE   -5.553437 -0.21035617  1.36739060  1.64251041  0.27592644 1.22633654
## IE   -2.712169 -0.46177605 -0.70000923 -0.35876289  0.03898283 0.57859832
## UE   -1.352600 -0.08299948 -0.01256821 -0.04695307 -0.10775751 0.06152859
##        medHxH  STYPELACS  STYPEOTH  STYPEPACS  STYPEPOCS
## HE 0.24323401 -2.1222332 -8.133826 -0.7652881 -0.7377289
## IE 0.31657422 -0.5928425 -0.524495 -0.4220204 -0.1184391
## UE 0.03075816 -2.1523724 -2.145884 -1.0299685 -0.8728789
##
## Std. Errors:
##    (Intercept)     groupA    groupAH     groupH     medHxA   medHxAH    medHxH
## HE  0.42818104 0.55715349 0.46422990 0.44364923 0.34032458 1.0315858 1.0231992
## IE  0.11902643 0.12449026 0.17332556 0.15231306 0.12767881 0.5213149 0.3479757
## UE  0.06886269 0.07883471 0.09598669 0.09599921 0.08079087 0.3857832 0.2345086
##    STYPELACS  STYPEOTH  STYPEPACS STYPEPOCS
## HE 0.6204778 95.304886 0.32715544 0.5026027
## IE 0.1507664  1.019888 0.12960950 0.1711820
## UE 0.1232062  1.015415 0.07211493 0.1069335
##
## Residual Deviance: 10827.62
## AIC: 10893.62
```

```
coeftest(modelA) #this shows the pvalues for each estimated coefficient
```

```
##
## z test of coefficients:
##
##                  Estimate Std. Error  z value  Pr(>|z|)
## HE:(Intercept) -5.553437   0.428181 -12.9698 < 2.2e-16 ***
## HE:groupA      -0.210356   0.557153  -0.3776 0.7057611
## HE:groupAH      1.367391   0.464230   2.9455 0.0032243 **
## HE:groupH       1.642510   0.443649   3.7023 0.0002137 ***
## HE:medHxA       0.275926   0.340325   0.8108 0.4174953
## HE:medHxAH      1.226337   1.031586   1.1888 0.2345232
## HE:medHxH       0.243234   1.023199   0.2377 0.8120990
## HE:STYPELACS   -2.122233   0.620478  -3.4203 0.0006255 ***
## HE:STYPEOTH    -8.133826  95.304886  -0.0853 0.9319869
## HE:STYPEPACS   -0.765288   0.327155  -2.3392 0.0193241 *
## HE:STYPEPOCS   -0.737729   0.502603  -1.4678 0.1421539
## IE:(Intercept) -2.712169   0.119026 -22.7863 < 2.2e-16 ***
## IE:groupA      -0.461776   0.124490  -3.7093 0.0002078 ***
## IE:groupAH     -0.700009   0.173326  -4.0387 5.375e-05 ***
## IE:groupH      -0.358763   0.152313  -2.3554 0.0185012 *
## IE:medHxA       0.038983   0.127679   0.3053 0.7601228
## IE:medHxAH      0.578598   0.521315   1.1099 0.2670497    
## IE:medHxH       0.316574   0.347976   0.9098 0.3629493    
## IE:STYPELACS   -0.592843   0.150766  -3.9322 8.417e-05 ***
## IE:STYPEOTH    -0.524495   1.019888  -0.5143 0.6070652    
## IE:STYPEPACS   -0.422020   0.129610  -3.2561 0.0011296 ** 
## IE:STYPEPOCS   -0.118439   0.171182  -0.6919 0.4890064    
## UE:(Intercept) -1.352600   0.068863 -19.6420 < 2.2e-16 ***
## UE:groupA      -0.082999   0.078835  -1.0528 0.2924193    
## UE:groupAH     -0.012568   0.095987  -0.1309 0.8958252    
## UE:groupH      -0.046953   0.095999  -0.4891 0.6247719    
## UE:medHxA      -0.107758   0.080791  -1.3338 0.1822749    
## UE:medHxAH      0.061529   0.385783   0.1595 0.8732828    
## UE:medHxH       0.030758   0.234509   0.1312 0.8956487    
## UE:STYPELACS   -2.152372   0.123206 -17.4697 < 2.2e-16 ***
## UE:STYPEOTH    -2.145884   1.015415  -2.1133 0.0345745 *  
## UE:STYPEPACS   -1.029968   0.072115 -14.2823 < 2.2e-16 ***
## UE:STYPEPOCS   -0.872879   0.106934  -8.1628 3.273e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
exp(confint(modelA)) #this shows the 95% confidence intervals for the OR
```

```
## , , HE
## 
##                     2.5 %       97.5 %
## (Intercept) 1.673804e-03 8.966880e-03
## groupA      2.718902e-01 2.414868e+00
## groupAH     1.580145e+00 9.749973e+00
## groupH      2.166199e+00 1.233014e+01
## medHxA      6.763142e-01 2.567545e+00
## medHxAH     4.513400e-01 2.574415e+01
## medHxH      1.716669e-01 9.475102e+00
## STYPELACS   3.549557e-02 4.040894e-01
## STYPEOTH    2.207285e-85 3.901132e+77
## STYPEPACS   2.449990e-01 8.833134e-01
## STYPEPOCS   1.785633e-01 1.280633e+00
## 
## , , IE
## 
##                   2.5 %     97.5 %
## (Intercept) 0.05257806 0.08383696
## groupA      0.49372704 0.80430267
## groupAH     0.35355328 0.69746889
## groupH      0.51825338 0.94154346
## medHxA      0.80956110 1.33539710
## medHxAH     0.64200438 4.95479993
## medHxH      0.69388735 2.71446290
## STYPELACS   0.41133835 0.74278707
## STYPEOTH    0.08018338 4.36862838
## STYPEPACS   0.50862192 0.84536194
## STYPEPOCS   0.63511463 1.24243293
## 
## , , UE
```

```
##
##                   2.5 %    97.5 %
## (Intercept) 0.22592133 0.2959302
## groupA      0.78858690 1.0741329
## groupAH     0.81815893 1.1919162
## groupH      0.79048541 1.1516572
## medHxA      0.76635887 1.0518912
## medHxAH     0.49927781 2.2651699
## medHxH      0.65124441 1.6329474
## STYPELACS   0.09127740 0.1479483
## STYPEOTH    0.01598569 0.8558099
## STYPEPACS   0.30996027 0.4112205
## STYPEPOCS   0.33875954 0.5151521
```

The resulting ratio can be interpreted as the following:

$$OR = \frac{\frac{P(Y = IE|group = H)}{P(Y = N|group = H)}}{\frac{P(Y = IE|group = A)}{P(Y = N|group = A)}} = \frac{\exp^{\alpha_H+\beta_H}}{\exp^{\alpha_A+\beta_A}} = \frac{\exp^{(-2.712169-0.35876289)}}{\exp^{(-2.712169-0.46177605)}}$$

```
numerator = exp(-2.712169 - 0.35876289)
denominator = exp(-2.712169 - 0.46177605)
OR = numerator/denominator
OR

## [1] 1.108506
```

For patients diagnosed with an acute ischemic stroke, those patients who received high heparin at the randomization period, relative to patients who received aspirin, were 10.85% more likely to have another ’ischemic event (IE)" within the 14-day observation period. (OR = 1.108506).

The Z distribution is used to assess significance of the beta estimates:

- H0: β=0
- HA: β ≠0

It is important to note in this case that the p value for the β_A is 0.0002078, which is much lower than an alpha of 0.05. The p value for the β_H is 0.0185012 which is also much smaller than an alpha of 0.05. Before we can blindly use the estimated beta coefficients from the model, we must assess if they are different from 0 via a z distribution as is shown above. The null hypothesis for the estimated coefficient is that $\beta = 0$, and therefore we must see a p value of less than 0.05 to properly accept the alternative hypothesis that $\beta \neq 0$.

In this example, the result directly answers the posed problem, “*is there a difference in the outcome (i.e., acute ischemic stroke) between the aspirin and high heparin groups?*” It is important to understand that the result does **not** hold for other types of events (*unknown events (UE) and haemorrhagic events (HE)*), the p value for the aspirin and high heparin beta coefficients are clearly larger than an arbitrary alpha cutoff of 0.05 for these events.

LIKELIHOOD RATIO TESTS:

The likelihood ratio test is a test of the significance of the difference between the likelihood ratio for the modelA with predictors (final model) minus the likelihood ratio for baseline model with only a constant in it.

Chi square is used to assess significance of this ratio:

- H0: There is no difference between null model and final model.
- HA: There is difference between null model and final model.

```
lrtest(modelA, "group") # Chi-Square=53.942, p=1.936e-08***

## # weights:  36 (24 variable)
## initial  value 16812.978012
## iter  10 value 5626.569754
## iter  20 value 5489.198381
## iter  30 value 5442.296146
## final  value 5440.781547
## converged

## Likelihood ratio test
##
## Model 1: outcome ~ group + medHx + STYPE
## Model 2: outcome ~ medHx + STYPE
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  33 -5413.8
## 2  24 -5440.8 -9 53.942  1.936e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

lrtest(modelA, "medHx") # Chi-Square=5.3109, p=0.8064

## # weights:  36 (24 variable)
## initial  value 16812.978012
## iter  10 value 5631.813067
## iter  20 value 5495.657807
## iter  30 value 5451.683302
## iter  40 value 5416.472228
## final  value 5416.466055
## converged

## Likelihood ratio test
##
## Model 1: outcome ~ group + medHx + STYPE
## Model 2: outcome ~ group + STYPE
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  33 -5413.8
## 2  24 -5416.5 -9 5.3109     0.8064

lrtest(modelA, "STYPE") # Chi-Square=504.92, p=< 2.2e-16 ***

## # weights:  32 (21 variable)
## initial  value 16812.978012
## iter  10 value 5746.255358
## iter  20 value 5705.922296
## iter  30 value 5666.401103
## final  value 5666.270931
## converged

## Likelihood ratio test
##
## Model 1: outcome ~ group + medHx + STYPE
## Model 2: outcome ~ group + medHx
##   #Df  LogLik  Df  Chisq Pr(>Chisq)
## 1  33 -5413.8
## 2  21 -5666.3 -12 504.92  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I performed likelihood ratio tests on the _group_, _medHx_, and _STYPE_ variables which I left in my model. The results of the likelihood ratio tests ascertain the significance of predictors to the model. Within the results section _group_ (_Chi-Square=53.942, p=1.936e-08***_) and _STYPE_ (_Chi-Square=504.92, p=< 2.2e-16_ ***) had significant main effects on outcome (recurrence of stroke or complications).

MODEL B:

To alternatively assess the question, I compared the *patients who received aspirin and heparin to the patients who received no medication* and looked at the ischemic event results that occurred within the 14-day observation period. The resulting ratio can be interpreted as the following:

$$OR = \frac{\frac{P(Y = IE|group = AH)}{P(Y = N|group = AH)}}{\frac{P(Y = IE|group = N)}{P(Y = N|group = N)}} = \exp^{\beta_{AH}} = \exp^{(-0.70000923)}$$

```
OR = exp(-0.70000923)
OR

## [1] 0.4965807

percent_less_likely = (1 - OR)
percent_less_likely

## [1] 0.5034193
```

For patients diagnosed with an acute ischemic stroke, those patients who received aspirin and high heparin at the randomization period, relative to patients who received no medication, were 50.34% less likely to have another ’ischemic event (IE)" within the 14-day observation period. (OR = 0.4965807; 95% CI [0.35355328, 0.69746889], p-value = 5.375e-05). Here I show the utility of quoting the CI which are shown below.

```
modelB <- multinom(outcome~group, data = stroke_df)

## # weights:  20 (12 variable)
## initial  value 16812.978012
## iter  10 value 5736.950461
## iter  20 value 5680.475075
## final  value 5668.919619
## converged

summary(modelB)

## Call:
## multinom(formula = outcome ~ group, data = stroke_df)
##
## Coefficients:
##    (Intercept)     groupA     groupAH      groupH
## HE   -6.193240 -0.1963032  1.35332669  1.65639049
## IE   -3.016155 -0.4539634 -0.70138642 -0.35234198
## UE   -2.222131 -0.0551458 -0.02029803 -0.02015252
##
## Std. Errors:
##    (Intercept)     groupA    groupAH     groupH
## HE  0.37595147 0.55542803 0.46189703 0.44121242
## IE  0.07855591 0.12432576 0.17318093 0.15211564
## UE  0.05428967 0.07719783 0.09397941 0.09397443
##
## Residual Deviance: 11337.84
## AIC: 11361.84
```

```
coeftest(modelB) #this shows the pvalues for each estimated coefficient

##
## z test of coefficients:
##
##                   Estimate Std. Error  z value  Pr(>|z|)
## HE:(Intercept) -6.193240   0.375951 -16.4735 < 2.2e-16 ***
## HE:groupA      -0.196303   0.555428  -0.3534 0.7237684
## HE:groupAH      1.353327   0.461897   2.9299 0.0033904 **
## HE:groupH       1.656390   0.441212   3.7542 0.0001739 ***
## IE:(Intercept) -3.016155   0.078556 -38.3950 < 2.2e-16 ***
## IE:groupA      -0.453963   0.124326  -3.6514 0.0002608 ***
## IE:groupAH     -0.701386   0.173181  -4.0500 5.121e-05 ***
## IE:groupH      -0.352342   0.152116  -2.3163 0.0205431 *
## UE:(Intercept) -2.222131   0.054290 -40.9310 < 2.2e-16 ***
## UE:groupA      -0.055146   0.077198  -0.7143 0.4750145
## UE:groupAH     -0.020298   0.093979  -0.2160 0.8290004
## UE:groupH      -0.020153   0.093974  -0.2144 0.8301986
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

exp(confint(modelB)) #this gives the 95% confidence intervals for the OR

## , , HE
##
##                    2.5 %       97.5 %
## (Intercept) 0.0009779114  0.004268944
## groupA      0.2766721174  2.440775263
## groupAH     1.5652180652  9.569952334
## groupH      2.2069909596 12.442909623
##
## , , IE
##
##                  2.5 %     97.5 %
## (Intercept) 0.04199847 0.05714356
## groupA      0.49775991 0.81034968
## groupAH     0.35316680 0.69631159
## groupH      0.52179361 0.94724188
##
## , , UE
##
##                  2.5 %    97.5 %
## (Intercept) 0.09743818 0.1205458
## groupA      0.81346631 1.1009343
## groupAH     0.81505940 1.1780944
## groupH      0.81518597 1.1782544
```

Limitations:

If allotted more time, I would have chosen to fit an offset into the multinom model because of the differing sample size of *N, UE, IE, and HE* within the outcome attribute. I would have also conducted a stratification analysis and conducted the same polytomous logistic regression on data subsetted by type of stroke (*STYPE*) and by which medications the patient was initially on (*medHx*). Creating these subsets of data would have allowed for a more consistent comparison of patients across treatment groups.

I would have also considered using a more powerful test that is related to the multinomial logistic regression, which is a discriminant function analysis. The discriminant function analysis, however, requires that the assumptions of normality, linearity, homoscedasticity, independence, and multicollinearity be clearly met.

I did conduct a rough check on the multicollinearity assumption via a ggpairs plot using the ggplot2 library, but because my wrangled dataset still included 31 attributes, it was challenging to assess if this assumption was clearly met, which is why I ended up using all categorical explanatory variables within my “Model A” and “Model B” analyses.

I could have expanded my model to include more covariates (both continuous and categorical) and conducted a Pearson’s chi-square test to compare the models pre and post covariate addition. I could have continued this process until I was able to fit a parsimonious model that accounted for the maximum amount of variation in the outcome variable. **For the purposes of this analysis, and given the limited amount of allocated time, I wanted to keep my model simple, as interpreting the odds ratios within the model can be challenging when multiple variables are included as explanatory variables.**

References:

1. The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19435 patients with acute ischaemic stroke. International Stroke Trial Collaborative Group. *Lancet*. 1997;349(9065):1569-1581.

2. Powers WJ, Rabinstein AA, Ackerson T, et al. Guidelines for the Early Management of Patients with Acute Ischemic Stroke: 2019 Update to the 2018 Guidelines for the Early Management of Acute Ischemic Stroke: A Guideline for Healthcare Professionals from the American Heart Association/American Stroke Association. *Stroke* .2019; 50(12). doi: 10.1161/str.0000000000000211.

3. Kleindorfer DO, Towfighi A, Chaturvedi S, et al. 2021 Guideline for the Prevention of Stroke in Patients with Stroke and Transient Ischemic Attack: A Guideline from the American Heart Association/American Stroke Association. *Stroke* .2021. doi: 10.1161/str.0000000000000375.

4. Kasper DL, Fauci AS, Hauser SL, Longo DL, Lameson JL, Loscalzo J. *Harrison's Principles of Internal Medicine*. New York, NY: McGraw-Hill Education; 2015.

5. Caplan LR, Kasner SE, Dashe JF. Etiology, Classification, and Epidemiology of Stroke. In: Post TW, ed. *UpToDate* .Waltham, MA: UpToDate.https://www.uptodate.com/contents/etiology-classification-and-epidemiology-of-stroke.Last updated March 13, 2017. Accessed March 19, 2022.