HMS 520: Introduction to Programming, Version Control and Data Wrangling

Final Project

Cassidy Chang, Yaz Ozten &
Sinclair Carr

December 14, 2021
11:00am PST

Our data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	nid	sex	super	meas_value	meas_stdev	CD4_lower	CD4_upper	time_lower	time_upper	iso3	time_point	site	sample_size	age
2	1993	1	other	0.074296	0.099057	0	5	2007	2013	IND	24	AP_RJ	14	15_25
3	1993	1	other	0.148794	0.118096	0	5	2007	2013	IND	6	AP_RJ	14	15_25
4	1993	1	other	0.0749	0.099243	0	5	2007	2013	IND	12	AP_RJ	14	15_25
5	1993	1	other	0.121332	0.09795	5	10	2007	2013	IND	6	AP_RJ	17	15_25
6	1993	1	other	0.05463	0.079756	5	10	2007	2013	IND	12	AP_RJ	17	15_25
7	1993	1	other	0.051405	0.078686	5	10	2007	2013	IND	24	AP_RJ	17	15_25
8	1993	1	other	0.040957	0.039777	10	20	2007	2013	IND	12	AP_RJ	40	15_25
9	1993	1	other	0.047234	0.041537	10	20	2007	2013	IND	24	AP_RJ	40	15_25

- HIV mortality data from a range of studies worldwide
- Contains some variables directly related to HIV such as: CD4 cell count, probability of mortality in the study time period, time since last initiation of anti-retroviral therapy (ART), study period length
- Other variables such as: sex, region, country, age

Project aims

- Data wrangling - change variable names to be more readable, scale variables as necessary, create new variables

- Data analysis - descriptive analyses; population-level associations between HIV-related variables; predict time_since_art based on numeric variables

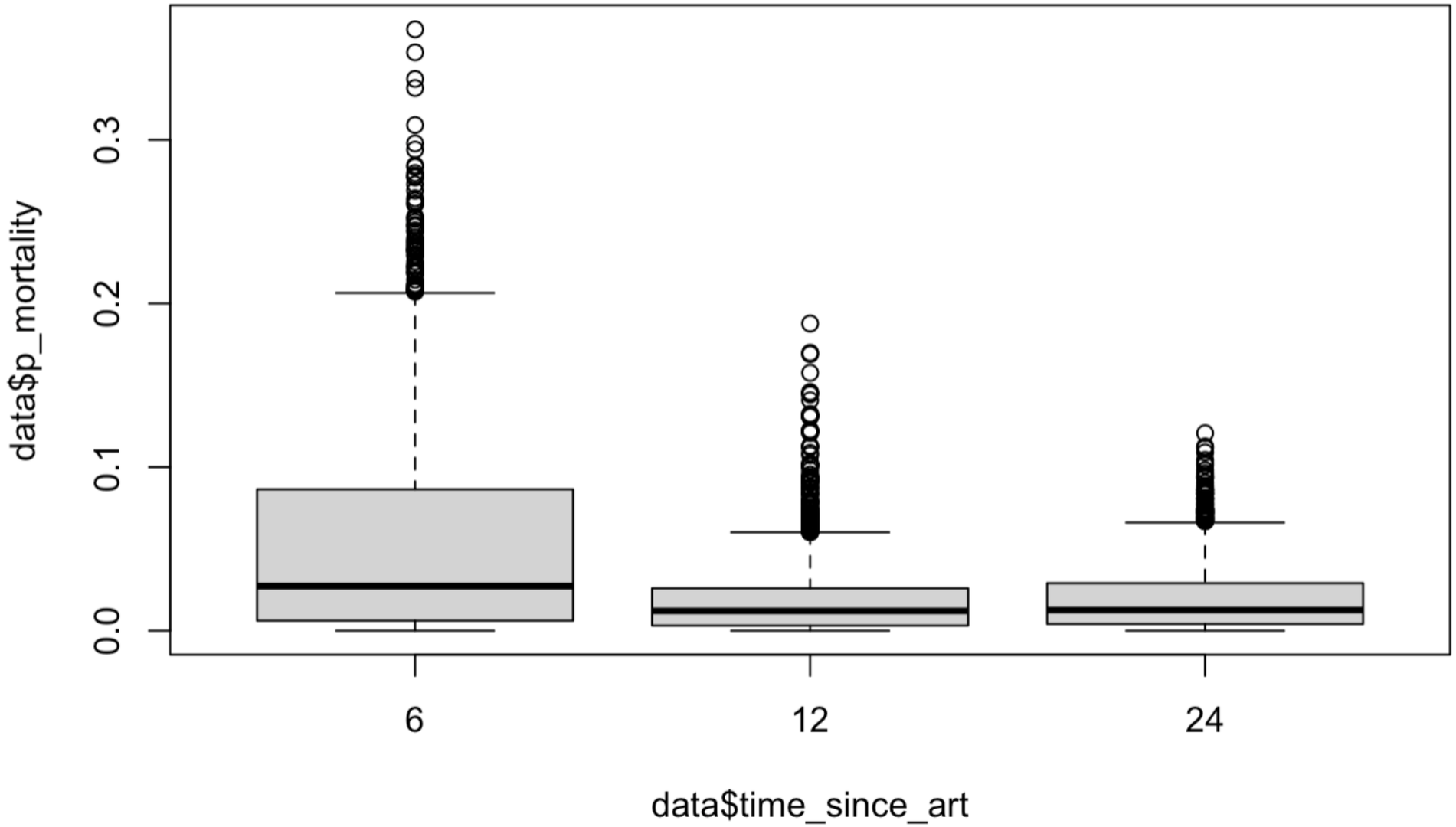
- Visualisation - regression outputs, goodness-of-fit, PCA, tSNE

Descriptive Analysis

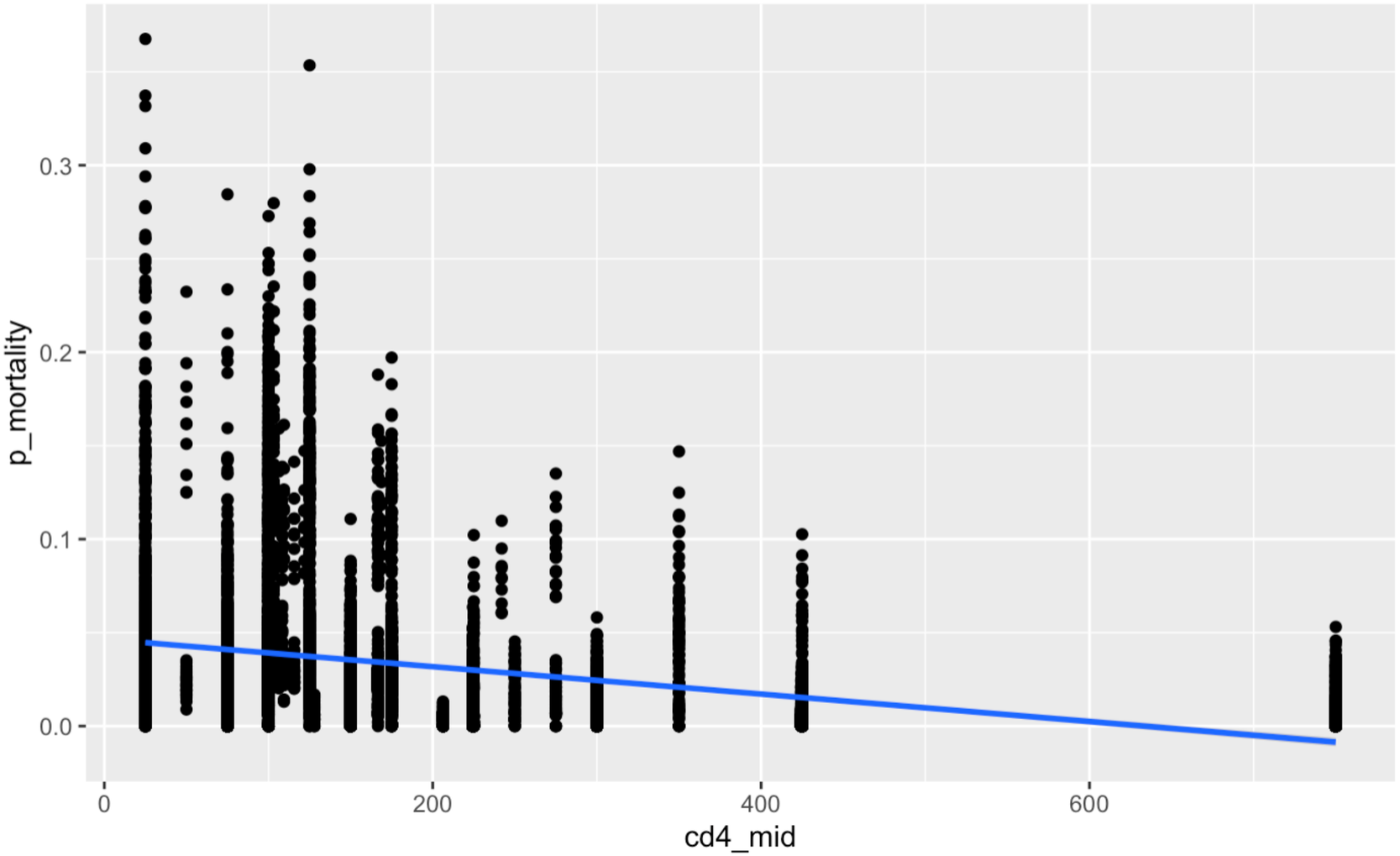
Characteristic	N = 9,334[1]
Sex	
female	4,662 (50%)
male	4,672 (50%)
Age groups	
15_25	1,866 (20%)
25_35	1,866 (20%)
35_45	1,866 (20%)
45_55	1,866 (20%)
55_100	1,870 (20%)
GBD Superregions	
high	4,770 (51%)
other	1,390 (15%)
ssa	3,174 (34%)
Time since ART inititation in months	
6	3,282 (35%)
12	3,045 (33%)
24	3,007 (32%)
Conditional probability of HIV mortality	0.02 (0.00, 0.04)
Unknown	20
Average CD4 count	150 (100, 300)

[1] n (%); Median (IQR)

Descriptive Analysis



Descriptive Analysis



Clustering for prediction

- Unsupervised machine learning method
- Grouping of data into a certain number of categories based on characteristics
- Many types of clustering algorithm
- Clustering types used in this analysis: k-means, k-medoids, hierarchical agglomerative, model-based

Example: K-means

- Most commonly used clustering method

- Define: number of centroids (cluster centres)

Pseudo-code:

1. Initiatilalise k random centroids
2. For each data point, find nearest centroid, assign
3. Move each centroid to mean of points assigned to that centroid
4. Repeat until convergence

Other clustering methods

- K-medoids - similar to K means, but different error measurement, uses datapoints as initial cluster centers
- Hierarchical agglomerative - attempts to find hierarchies in data, uses a dissimilarity matrix, very slow!
- Model-based - data assumed to be generated from a finite number of models, of which optimal model is chosen based on BIC

Cluster accuracy computation

- Depends on orientation of data

  Eg: True = (1, 2, 3, 3, 3)
  Pred = (3, 2, 1, 1, 1)

- Standard accuracy: 20%, fair accuracy: 80%

- Formation of bipartite graph

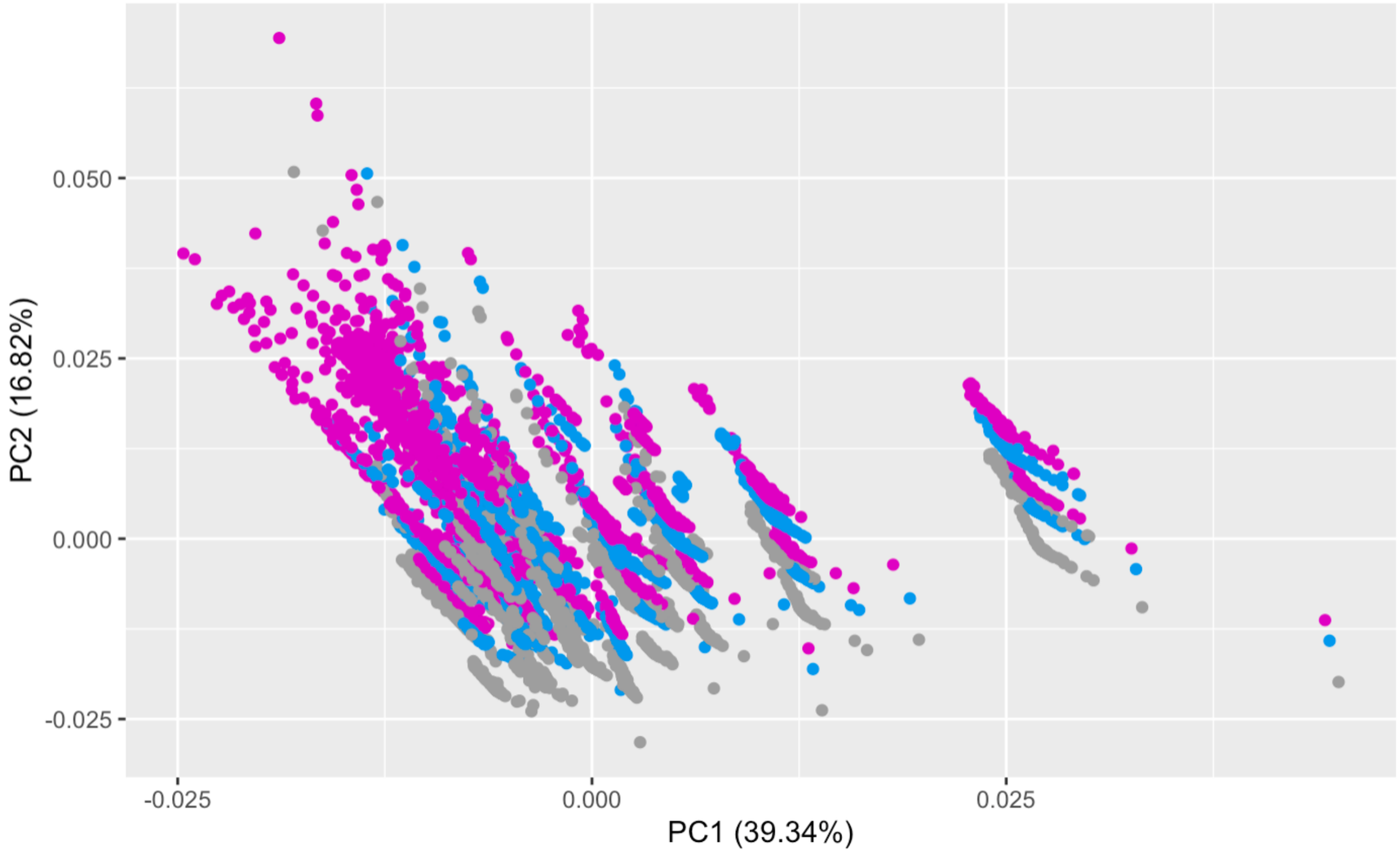
- Solved using Hungarian algorithm to compute accuracy

Clustering: Results & Conclusion

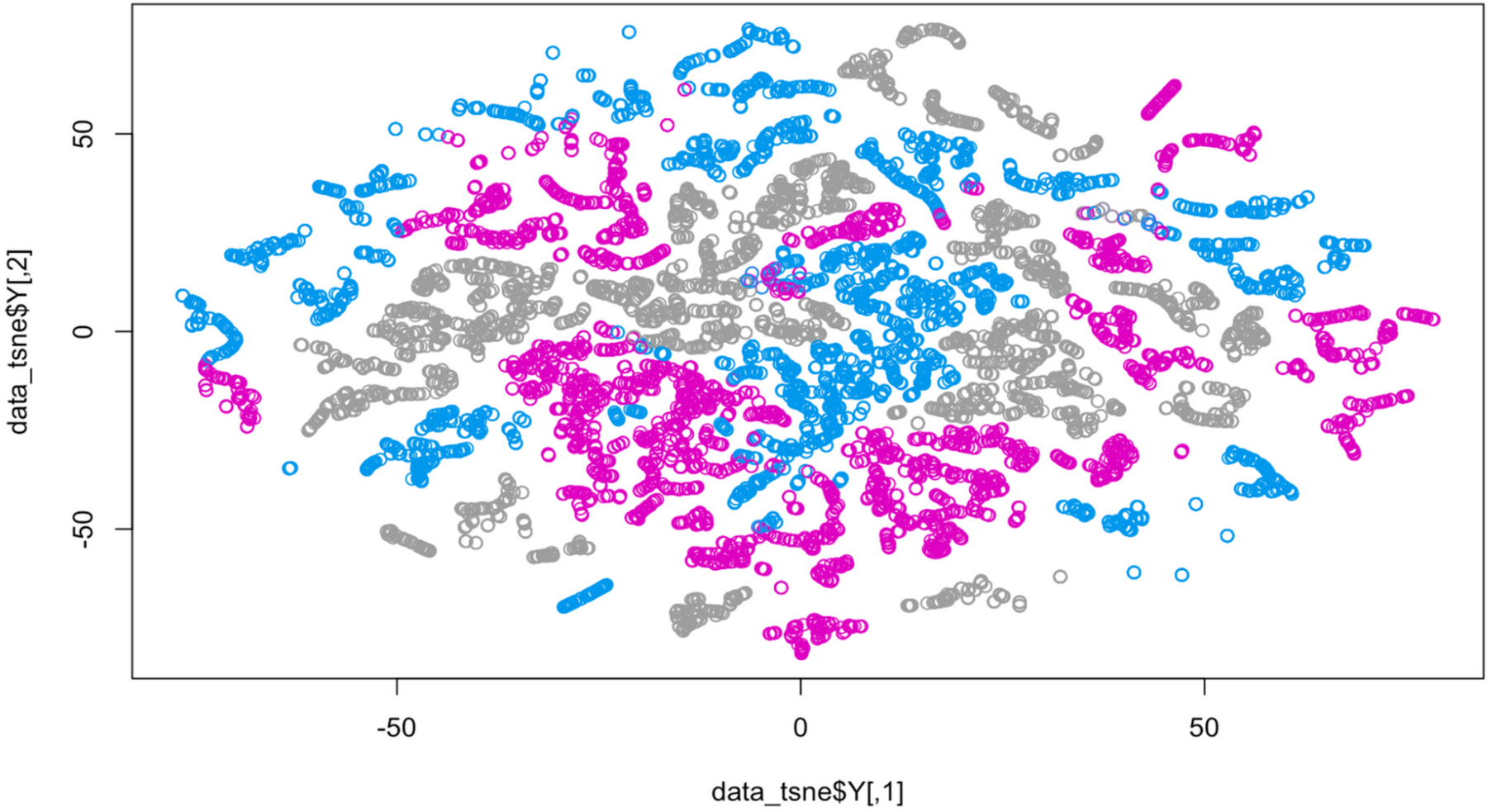
```
Mean k-means predictive accuracy is: 0.4288729
Mean k-medoids predictive accuracy is: 0.4933255
Mean agglomerative predictive accuracy is: 0.3514035
Mean model-based predictive accuracy is: 0.5170559>
```

- Reached above 50%
- Better to use supervised ML
- Yield higher accuracy with neural networks, support vector machines, ensemble methods

Visualisation with PCA



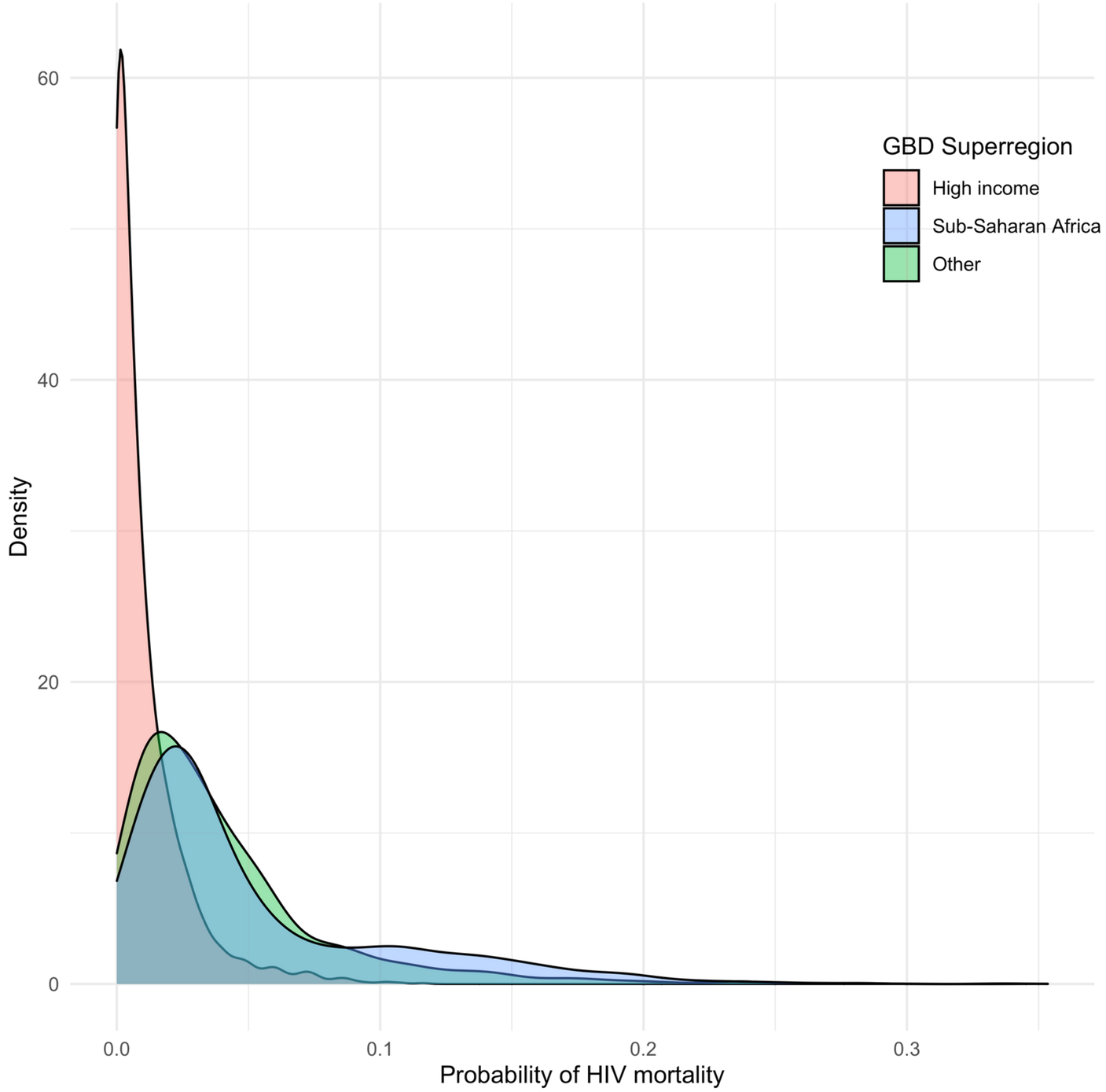
Visualization with tSNE



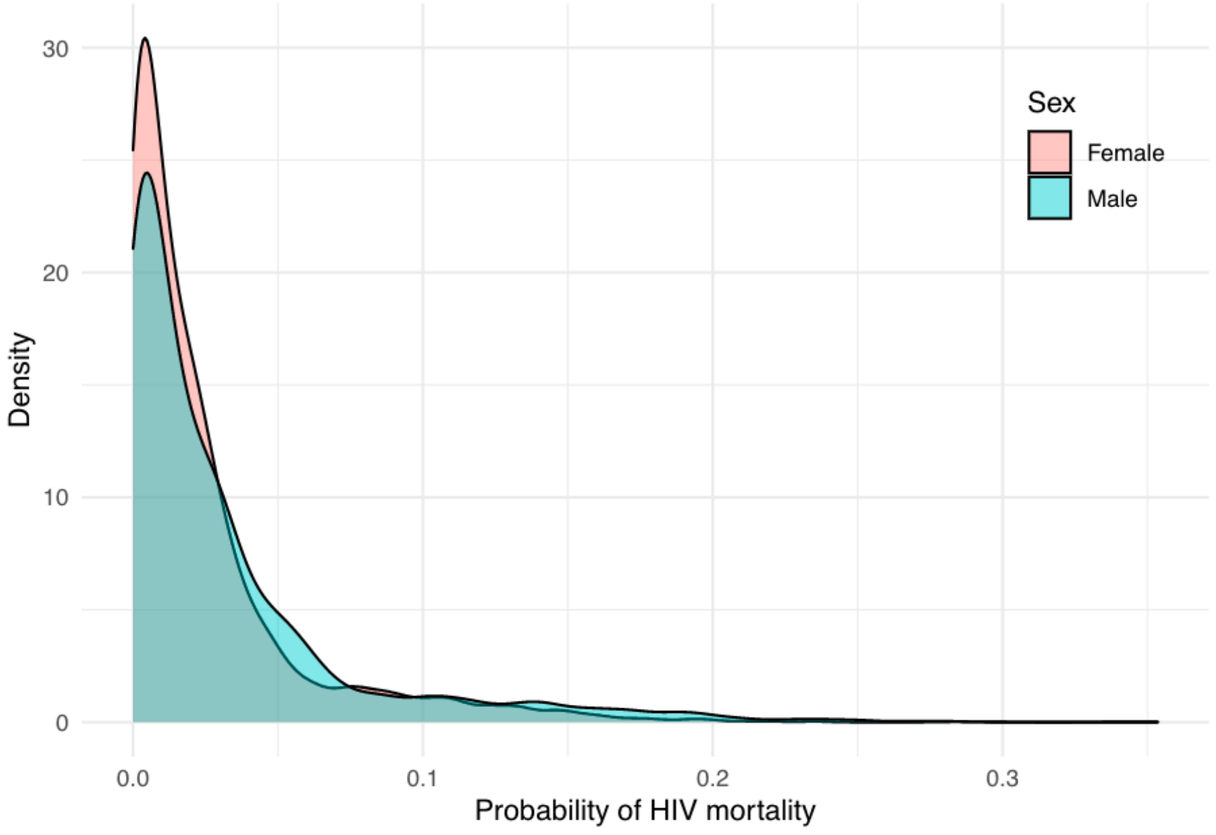
Time since ART initiation and HIV mortality

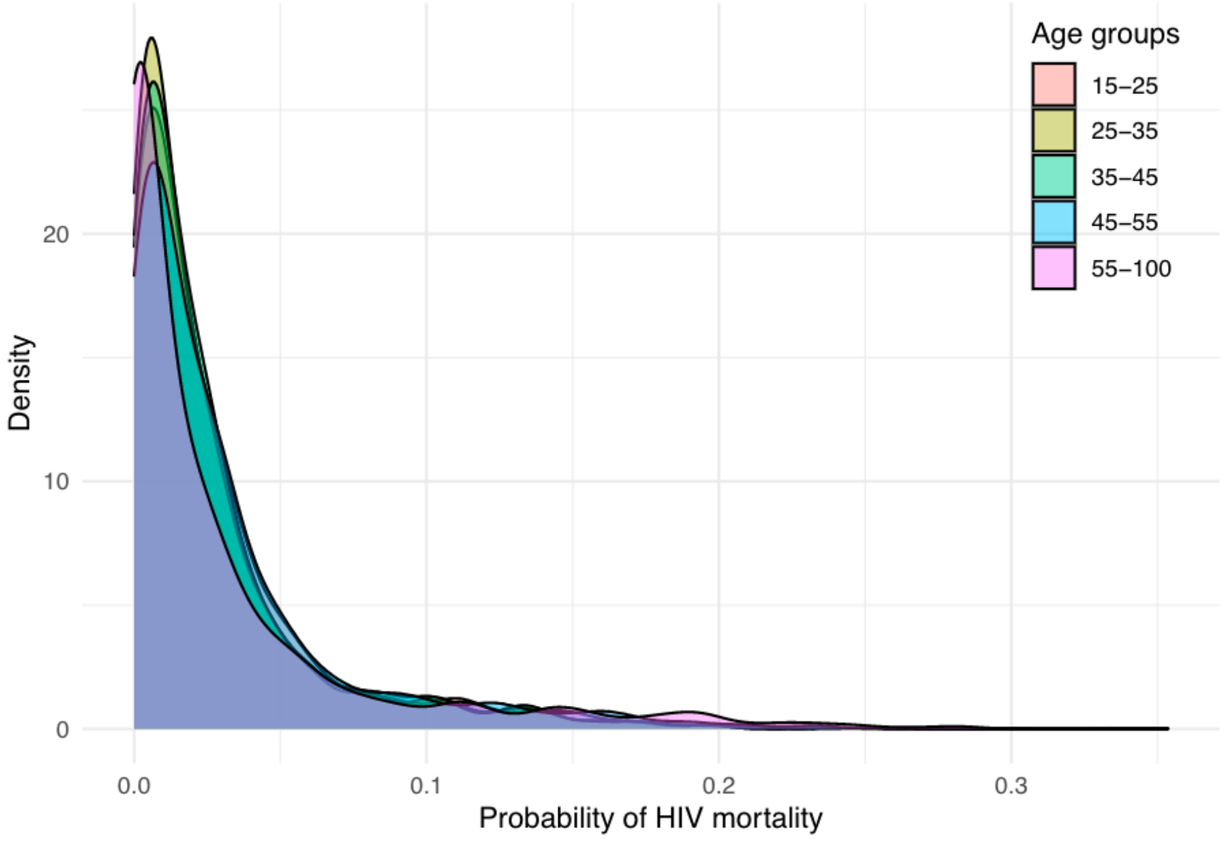
`Model: logit_mortality ~ time_since_art + cd4_mid + sex + age_group + study_length`

- Assocation between time since initiation of ART and mortality of HIV
- Unadjusted and adjusted weighted least squares linear regression models
- Inclusion of random effects for study ID



Let's have a look at the outcome


Sex
Female
Male
Density
Probability of HIV mortality
0
10
20
30
0.0
0.1
0.2
0.3


Age groups
15–25
25–35
35–45
45–55
55–100
Density
Probability of HIV mortality
0
10
20
0.0
0.1
0.2
0.3

Results

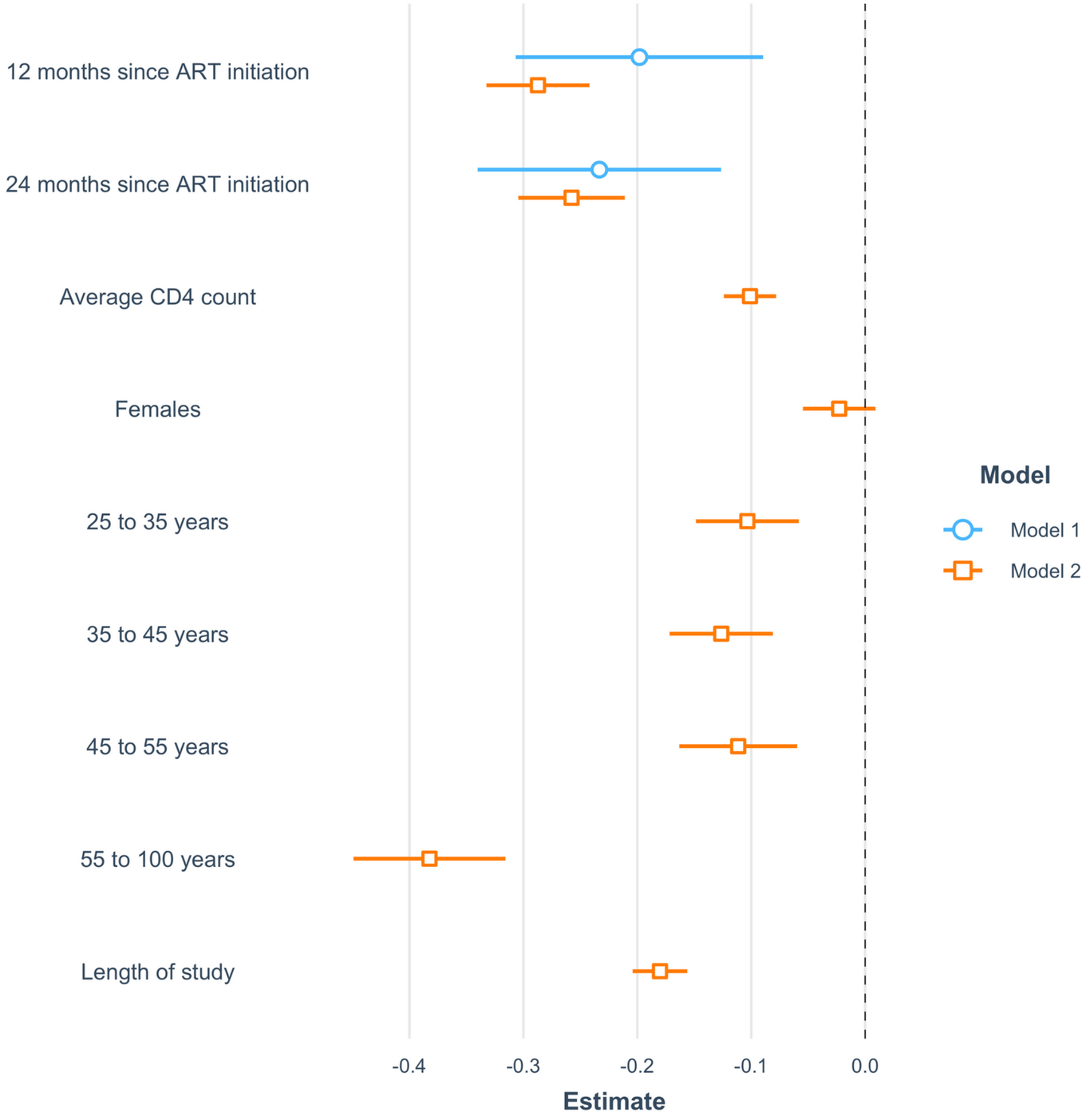
Results of unadjusted regression model

	Logit(Probability of Mortality)
	logit_mortality
time_since_art12	-0.198***
	(0.012)
time_since_art24	-0.233***
	(0.011)
Constant	-3.102***
	(0.009)
Observations	8,954
R^2	0.046
Adjusted R^2	0.046
Residual Std. Error	60.734 (df = 8951)
F Statistic	218.199*** (df = 2; 8951)
Note:	*p<0.1; **p<0.05; ***p<0.01

Results of adjusted regression model

	Logit(Probability of Mortality)
	logit_mortality
time_since_art12	-0.287***
	(0.008)
time_since_art24	-0.258***
	(0.008)
cd4_mid	-0.001***
	(0.00002)
sexmale	-0.023***
	(0.007)
age_group25_35	-0.103***
	(0.012)
age_group35_45	-0.126***
	(0.012)
age_group45_55	-0.111***
	(0.014)
age_group55_100	-0.382***
	(0.013)
study_length	-0.033***
	(0.001)
Constant	-2.452***
	(0.013)
Observations	8,954
R^2	0.582
Adjusted R^2	0.581
Residual Std. Error	40.241 (df = 8944)
F Statistic	1,382.183*** (df = 9; 8944)
Note:	*p<0.1; **p<0.05; ***p<0.01

Forest Plot


12 months since ART initiation
24 months since ART initiation
Average CD4 count
Females
25 to 35 years
35 to 45 years
45 to 55 years
55 to 100 years
Length of study
-0.4
-0.3
-0.2
-0.1
0.0
Estimate
Model
Model 1
Model 2

AIC and BIC tests

$$AIC\ (M) = D(M) + 2 \times |M|$$

-7364.462

$$BIC\ (M) = D(M) + \log(n) \times |M|$$

-7314.763

Thank you for listening!