Distribution Agreement

Signature:

	04/7/2023
Logan Gerig	Date

Clustering the Liver Measures of Women Living with HIV

By

Logan Gerig

Master of Science of Public Health

Department of Biostatistics and Bioinformatics

____

Christina Mehta, PhD, MSPH

(Thesis Advisor)

____

Cecile Lahiri, MD, MS

(Thesis Reader)

Clustering the Liver Measures of Women Living with HIV

By

Logan Gerig

B.S., Youngstown State University, 2021

Thesis Committee Chair: Christina Mehta, PhD, MSPH

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics & Bioinformatics

2023

Abstract

Clustering the Liver Measures of Women Living with HIV

By Logan Gerig

Background: Non-alcoholic fatty liver disease is more prevalent amongst those living with HIV compared to the general population (Maurice et al., 2017). Our previous work has found that three commonly used non-invasive liver measures, APRI, FIB-4, and NFS, showed conflicting results in quantifying the degree of liver fibrosis in women living with HIV (WWH) over an extended period (Yu et al., 2022). Clustering, an unsupervised machine learning technique, can be used to partition trajectories into homogeneous discrete groups where they can be compared amongst each other (Teuling et al., 2021).

Objectives: Compare five longitudinal clustering algorithms on WWH's liver trajectories to see how they perform with respect to observational data that is subject to unequal follow-up; explore the clusters identified by the best performing method; and compare these results to those identified by cluster results from Fibroscan data.

Methods: Data from the Women's Interagency HIV Study (WIHS) used in our previous work had all three liver measures clustered using: longitudinal K-Means (KML), growth-curve modeling into K-Means (GCKM), group-based trajectory modeling (GBTM), generalized linear mixed modeling assuming normal mixture in random effects (GLMM), and anchored k-medoids. The best performing method's clusters were explored to discover features associated with cluster membership. Cross-sectional, Fibroscan data was clustered using K-Means and had their subsequent clusters compared with the longitudinal ones.

Results: GBTM was the best performing method for cross-validation and clinical interpretably with a cluster solution of five, five, and six clusters for APRI, FIB-4, and NFS. Little correlation was found between the features examined and the clusters identified. Furthermore, cluster membership was inconsistent among the three liver measurements, with all three showing discordance with the two Fibroscan-identified clusters.

Conclusions: Issues such as convergence and extensive imputation were encountered for several of the longitudinal clustering methods, suggesting that more flexible methods such as GBTM should be developed. The clustering identified by GBTM indicated a lack of latent variables responsible for all three liver measurement trajectories. Finally, the observed inconsistency between the three liver measurement clusters and the Fibroscan cluster suggests that clinicians should exercise caution when assessing liver health in WWH.

Clustering the Liver Measures of Women Living with HIV

By

Logan Gerig

B.S., Youngstown State University, 2021

Thesis Committee Chair: Christina Mehta, PhD., MSPH

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health in
Biostatistics & Bioinformatics
2023

1. Introduction

Human immunodeficiency virus, better known as HIV, is the virus responsible for the development of acquired immunodeficiency syndrome (AIDs) and is a major public health crisis both domestically and globally. Since the 1980s, HIV/AIDS has been categorized as one of the deadliest epidemics in human history. HIV functions by effectively destroying CD4+ T cells, cells pivotal to fighting infections. Thus, their resulting destruction can lead to severe medical complications and even death (U.S. Department of Health and Human Services, *National Institute of Allergy and Infectious Diseases*).

Since the 1980s, extensive progress has been made regarding the treatment and prevention of HIV. This success can be attributed to the development and distribution of antiretroviral therapies (ARTs), which effectively limit the spread and progression of the virus (Zhao et al., 2022).

One subset of ARTs, integrase strand transfer inhibitors (INSTIs), has quickly become an effective tool in combating HIV, functioning to block the replication of the virus (Smith et al., 2021). After its release in 2007, INSTIs have become a mainstay as the newest form of first-line therapy in treatment-naïve people living with HIV (PLWH). There are now five INSTIs that have been introduced, namely: raltegravir, elvitegravir, dolutegravir, bictegravir, and cabotegravir, which are all FDA-approved. While effective, a growing body of evidence is showing a positive association between INSTI use and clinically significant weight gain (Scarsi et al., 2020).

This increased weight gain and the associated increase in body mass index (BMI) appeared to exceed that associated with other ART regimens, such as non-nucleoside reverse transcriptase inhibitors (NNRTIs). Little to no data from clinical trials exist regarding understanding weight gain among those treated with an INSTI regimen. Thus, the metabolic causes and resulting impacts of this weight gain remain largely unknown to clinicians and researchers (Sax et al., 2019). Furthermore, women living with HIV (WWH), when switched to an INSTI regimen, are particularly at risk for significant increases in body weight, as well as waist circumference (Kerchberger et al., 2019).

One consequence of such weight gain is the onset of non-alcoholic fatty liver disease (NAFLD). NAFLD can lead to severe liver scarring (fibrosis) which can result in liver failure. Previous research has found that both short and long-term weight gain increases the risk of NAFLD and significant fibrosis of liver tissue (Wijarnpreecha et al., 2022).

NAFLD has been estimated to have a prevalence of 35% among PLWH compared to 25% in the general population (Maurice et al., 2017). Furthermore, other research has unearthed a potential association between the uptake of INSTIs for PLWH and NAFLD. Stavudine, elvitegravir, and raltegravir use were associated with NAFLD presence. This same study also found a higher prevalence of fatty liver buildup (hepatic steatosis) and weight gain has been found to be significantly higher in those taking these INSTIs than in those with other treatment regimens. Thus, there exist clinical concerns regarding the uptake of INSTIs in PLWH and the onset of hepatic steatosis and liver fibrosis (Kirkegaard-Klitbo et al., 2021).

The tools used by clinicians to evaluate the degree of fibrosis vary. Clinicians oftentimes utilize non-invasive measures to avoid invasive and complication-prone liver biopsies. Three commonly utilized non-invasive measures include NAFLD fibrosis score (NFS), Fibrosis-4 (FIB-4), and aminotransferase (AST)/platelet ratio index (APRI). The formula for these scores is as follows:

1) $$NFS = -1.675 + 0.037 * age\ (year) + 0.094 * BMI\ (kg/m^2) + 1.13 * \frac{IFG}{diabetes\ (yes=1, no=0)} + 0.99 * \frac{AST}{ALT} - 0.013 * platelet\ count\ (\text{x } 10^9/\ L) - 0.66 * albumin(g/dL)$$

2) $$FIB4 = \frac{age*AST}{platelet\ count*\ \sqrt{ALT}}$$

3) $$APRI = \frac{\frac{AST}{40}}{platelet\ counts\ (x10^9/L)} * 100$$

For all three, a higher score indicates more liver fibrosis (Amernia, et al., 2021 & Angulo et al., 2007). These scores can be useful in discerning liver morbidity and mortality events related to NAFLD (Lee et al., 2021). Fibroscans, which are considered the best alternative to quantifying liver fibrosis aside from invasive biopsies, can also quantify the liver fibrosis (Afdhal, 2012).

Our previous work investigated the effects of an INSTI regimen on WWH using these liver measures. Unexpectedly, we observed discrepancies in the clinical conclusions of the three liver measures. While FIB-4 and APRI had minimal changes in the INSTI cohort, NFS diverged by showing a larger increase over time in WWH (Yu et al., 2022). This suggests that these non-invasive liver measures may not provide the same conclusions.

Such divergence has clinical implications. We see this in our previous work using the same study population that obtained a Fibroscan, where we found a higher odds of hepatic steatosis within one year of starting INSTI's relative to the control cohort (Lahiri et al., 2023). As a result, APRI, FIB-4, and NFS could lead to misleading insights into the relative liver health of WWH. This also suggests that there may exist some potential unknown population characteristics that influence the degree of NAFLD.

A method that can be used to acquire insight into these characteristics is clustering. Clustering, which is an exploratory unsupervised machine learning technique, results in the grouping of data points with heterogeneous characteristics. This is oftentimes done with the intention of gauging the underlying structure to which the data may conform (Pham et al., 2005). Within recent years, there has been an increased incidence of using techniques such as clustering in medical research for prognostic purposes (Alashwal et al., 2019).

Longitudinal data provides an avenue for clustering, not just data points, but patients' trajectories over time. Because common model approaches focus on the mean trend, between-subject variability may not be adequately captured. This is an issue that can be resolved through longitudinal clustering. Algorithms such as longitudinal K-means (KML) and group-based trajectory modeling (GBTM) allow for the subsequent clustering of patients based on their trajectories. These resulting clusters can then be examined to see the summary characteristics of patients contained within each cluster. This technique can provide researchers and clinicians with valuable diagnostic information.

While useful, there exist several limitations that can be encountered when performing longitudinal clustering. For instance, there are no standardized methods for conducting and evaluating cluster analysis nor for discerning which method is superior. Additionally, the performance of these methods may vary in the presence of the issues associated with observational data, such as missing visits or unequal follow-up periods.

Research has found conflicting performances of varying longitudinal clustering methods such as KML and GBTM (Teuling et al., 2021). Thus, there remains a lack of insight into the viability of methods in certain contexts. Most published applications oftentimes lack clear rationales for a chosen method. Furthermore, current research has uncovered contradictory results when comparing the efficacy of two or more clustering methods (Pham et al., 2005).

This paper is a secondary data analysis investigating the effects of the uptake of INSTIs in women living with HIV and subsequent NAFLD progression using cluster analysis. We first explore the utility of five different clustering methods in the presence of longitudinal observational data. This is important, given observational data such as the data used in this study are prone to trajectories with unequal visits and missing values.

To the best of our knowledge, there lacks literature that addresses the discrepancies between WWH and NAFLD using clustering methods.

Using cross-validation, we compared the resulting cluster solutions for KML, growth curve modeling into K-Means (GCKM), GBTM, generalized linear mixed models assuming a normal random mixture (GLMM), and anchored K-Medoids. These five methods reflect the diverse algorithms available for clustering longitudinal data. The method that showed the best results for the cross-validation and contained the most clinically interpretable results was then used to explore the attributes of the clusters across the three liver measures.

We then compare these cluster results to that of the clustering of the more clinically informative cross-sectional Fibroscan data. With this, we inquire which liver measure’s cluster assignments and attributes align with the Fibroscan clusters. All of this with the intention of providing clinicians and researchers with valuable information on the behavior of these non-invasive liver measures.

2. Methods

2.1 Study Population

Our study population consisted of participants from the Women's Interagency HIV Study (WIHS), who had biologic and behavioral data collected at visits that occurred every six months. The Women's Interagency HIV study was established in 1993 and is the biggest ongoing longitudinal cohort study of WWH and at-risk women in the U.S. Data is collected from WIHS participants through methods including surveys, medical examinations, and specimen collection. The sites of these visits included the following: Atlanta, GA; Birmingham, AL/Jackson, MS combined site; Chapel Hill, NC; Chicago, IL; Miami, FL; New York City, NY; Los Angeles, CA; San Francisco, CA; and Washington DC (Bacon et al., 2005 and Barkan et al., 1988).

Eligibility criteria follows that of our parent study where WWH that had untreated viral hepatitis, consumed more than 12 drinks a week, and had metabolic or autoimmune chronic diseases were excluded (Yu et al., 2022). 872 virologically suppressed WWH that either remained on non-INSTI ART or switched to or added an INSTI to their ART were used in subsequent analysis. Visits examined encompassed the baseline, which was prior to INSTI switch, to post-switch, which occurred after the second visit.

These participants' observations were collected between the years of 2007-2020. To ensure consistency in the time frame of follow-up visits, those that occurred prior to three or after nine months of the previous visit were excluded. Consequently, two observations were removed due to occurring less than three months from the previous visit, and 342 observations were removed due to occurring past nine months since the previous visit. Thus, there were 5,631 observations in total. The number of follow-up visits ranged from one to fifteen.

To enable cross-validation, a training and testing set reflecting a 70-30 split was randomly formed with a seed to ensure reproducibility. This resulted in 610 participants in the training set and 262 participants in the testing set. These two sets were clustered using the five methods described in 2.2. The results of these clustering methods were compared to determine the best performing method. This method was then used on the combined dataset to explore our second research question. All clustering methods employed utilized seeds to ensure reproducibility of results.

Fibroscan data that had been collected between 2014-2018 in 254 study participants was also used for cross-sectional clustering. These observations are limited to those who obtained a valid Fibroscan after the switch visit to INSTIs (Price et al., 2022). Values of hepatic steatosis via controlled attenuation parameter (CAP) and fibrosis via liver stiffness (LS) were jointly clustered. The cluster results were then explored to address our objectives.

2.2 Individual Cluster Methods

2.2.1 KML

K-means is a non-parametric, distance-based method that when used with longitudinal data becomes longitudinal k-means (KML). It functions by performing an expectation-maximization algorithm. As such, every observation is assigned to a prespecified number of clusters. From here, the algorithm iteratively partitions the observations into the prespecified number of clusters until the minimum distance within the cluster and maximum distance between cluster is achieved (i.e., convergence). The method requires that subjects contain an equal number of observations that are aligned in time. Because KML assumes equal variance, outliers can distort the resulting clusters (Genolini & Falissard, 2011).

Because of KML's complete data requirements, several data transformation procedures were performed. First, the trimmed APRI and FIB-4 scores were used to limit the effects of heterogeneous variance on the partitioning. These were derived by trimming 2.5% off the upper and lower bounds of these two measures. The upper and lower bound for APRI were 0.6 and 0.09 respectively. For FIB-4, the upper and lower bound used were 2.37 and 0.4 respectively. This resulted in 264 APRI and FIB-4 observations being trimmed.

Along with this, clustering was also limited to up to seven visits. This is due to both the dramatic drop-off in visits about the fifth visit and to ensure a long enough time frame to observe changes in liver health following INSTI use (Figure 1). This resulted in 718 observations from the training and 278 observations from the testing sets being excluded. However, no participants were removed from either set. The KML package automatically performs linear interpolation for missing values in the middle of a trajectory and last observation carried forward (LOCF) and first observation carried backwards (FOCB) for missing values at the end and start of a trajectory respectively (Genolini & Falissard, 2011).

2.2.2 GCKM

Modeling trajectories via a growth curve model and subsequent clustering of the subject's random effects using k-means is referred to as GCKM (Twisk & Hoekstra, 2012). In this method, the fixed effects (i.e., the overall trajectory) is modeled. From here, each subject's deviance from the fixed effects (i.e., random effects) is then clustered using the k-means algorithm. The random effects are assumed to be multivariately normally distributed with mean zero. They are also assumed to have an unstructured variance-covariance matrix and uncorrelated measurement error that is also independently and normally distributed with mean zero and common variance (Den Teuling et al., 2021).

The growth curve model is estimated via maximum likelihood estimation. The subject specific random effects are also estimated using the best linear unbiased predictors. From here, the random effects are clustered via the k-means algorithm. GCKM, like KML requires complete trajectories. However, these trajectories can be unequal in follow-up length (Den Teuling et al., 2021).

GCKM's inability to function in the presence of missing values meant that 96 training and 51 testing sets missing NFS values were imputed using the means of that specific participant's trajectory. This imputation method was chosen due to its relative simplicity. However, this comes at the potential expense of distorting the overall shape of patients' trajectories. In addition, due to the dramatic drop of follow-up among participants, clustering was also limited to up to seven visits, as was done for KML clustering.

2.2.3 GBTM

Group-based trajectory modeling (GBTM) models longitudinal data via homogeneous clusters, akin to that of k-means (Teuling et al., 2021). Consequently, subjects are represented solely by their corresponding cluster trajectory (Nagin & Odgers, 2010). These trajectories are modeled using a parametric model such as gaussian or gamma. These can be considered multilevel models, where the clusters are non-parametric random effects. The model parameters and corresponding clusters are then estimated via likelihood maximization (Teuling et al., 2021). Consequently, the GBTM method enables the use of domain knowledge through distributional assumptions in a relatively easy to interpret model (Den Teuling et al., 2021).

GBTM's inability to function in the presence of missing values meant that 96 training and 51 testing sets missing NFS values were imputed using the means of the participant's trajectory. With APRI and FIB-4 showing a non-zero right skewed distribution, a gamma distribution was assumed for both the APRI and FIB-4 models. A gaussian distribution was assumed for NFS.

2.2.4 GLMM

Generalized linear mixture modeling clustering (GLMM) can cluster participants by using a normal mixture in the random effects (Pan et al., 2020). GLMM works by modeling trajectories via a mixture of gaussian models. While these mixture models share the same gaussian distribution, they each contain different coefficients (Den Teuling et al., 2021). These models are estimated through likelihood maximization via the Monte Carlo expectation-maximization algorithm (Huang et al., 2014).

In essence, the model seeks to estimate cluster matrix containing the probability of cluster membership given the coefficients from the gaussian model for that subject. This is done iteratively, with the cluster membership probabilities conditioned on the model parameters being estimated and then vice versa. This is done until the resulting increase in likelihood is adequately low. Consequently, GLMM's are vastly more computationally intensive than that of GBTMs or GCKMs (Den Teuling et al., 2021).

2.2.5 Anchored K-Medoids

Anchored K-Medoids follow along the same process as that of K-means, but with slight modifications. The modifications given an ordinary least squares regression is first fitted to each participant's trajectory. Afterwards, the initial cluster means are selected amongst the regression slopes (anchors) rather than random initial values as in K-means. The point of this is to reduce the effect of outliers or drastic short-term fluctuations in trajectories (Adepeju et al., 2021).

From here, the algorithm continues to run iteratively to reduce not the squared error but rather the sum of dissimilarities between observations and the center of the respective cluster (medoids) rather than the average. This is done until convergence is obtained, like that of KML. The use of medoids rather than means enables k-medoids to be more robust to heterogeneity in variance and lead to more balanced cluster solutions (Adepeju et al., 2021).

For the same justification for KML and GCKM, clustering was limited to up to a participant's seventh visit. This resulted in 718 observations from the training and 278 observations from the testing sets being excluded. Because of the unequal number of follow-up visits and complete data requirements, linear interpolation for missing values was performed for trajectories containing missing visits.

2.3 Evaluating cluster methods

Because these methods are being compared with a non-synthetic dataset, the relative accuracy of these methods cannot be ascertained. This is due to the number and composition of the population's clusters being unknown. However, relative precision and clinical interpretability can be used to compare the relative efficacy of these five methods. Part of this efficacy is whether the results from the cluster methods are consistent between the training and testing sets in addition to whether these methods reach convergence.

For the purposes of this study, consistency can be interpreted as similar number and behavior of subsequent clusters obtained for the two sets. If so, this suggests that such a method is robust to differences in sample sizes and that the cluster results are more precise. Another criterion used to compare the relative effectiveness of the clustering methods is whether the cluster results are clinically meaningful and relevant. For our purposes, clusters that comprise less than 5% of the study population are deemed to be clinically irrelevant, as < 5% presents too small of a portion of the population to deem clinical value.

To determine the appropriate number of clusters for a solution, a mix of metrics is used in conjunction with the elbow method. This method is conducted by plotting the quality metrics and identifying the relative elbow of the curve. It is worth noting that deciding on the elbow of a plot is a relatively subjective process. In the case that more than one metric was used to evaluate a specific method, the elbow point that corresponded to the lowest number of clusters was used. A variety of metrics can be used for this method (Teuling et al., 2021).

One such type of metric is the traditional information criterion of AIC and BIC. These indicators seek to strike a balance between the relative fit of a method with that fit's complexity penalized. As such, a lower AIC or BIC corresponds to a better solution (Teuling et al., 2021).

For the purposes of this study, these two metrics are used for KML, GCKM, and GBTM. Moreover, the log-likelihood can be plotted and used in a similar manner. This metric is used in conjunction with AIC and BIC for both KML and GCKM (Teuling et al., 2021). As for GLMM, three different metrics are used for obtaining a cluster solution. These include the weighted residual sum of squares, a measure of deviation, where a smaller value corresponds to a better solution.

Mean squared error and entropy are the other metrics used for GLMM, where lower values for both indicate a better performing solution (Teuling et al., 2021). As for the anchored K-Medoids method, the Calinski-Harabasz score will be used to assess the appropriate number of clusters. This metric measures the within cluster variance against the variance between clusters (Teuling et al., 2021).

2.4 Clustering Fibroscan

In addition to the longitudinal clustering approaches employed, a cross-sectional dataset containing a select number of WIHS participants with Fibroscan visits was also clustered. This was done with the intention of exploring the characteristics of the identified clusters and whether these clusters align with the ones identified by the other liver measures.

For this, the traditional k-means method was employed and the average silhouette width (ASW) with the elbow method. This metric quantifies the similarity in a subject's clustered values within a cluster to the relatedness of the other clusters. A higher ASW corresponds to a greater solution (Teuling et al., 2021). To prevent outliers from influencing the results, LS was log-transformed and then scaled with CAP to be jointly clustered.

2.5 Software

All analysis was performed in R version 4.2.2 and RStudio version 3.0 (R Core Team, 2022 & Posit Team, 2023). The implementation of clustering methods KML, GCKM, GBTM, and GLMM was done using version 1.5.0 of the latrend package (Teuling, 2022). Anchored K-Medoids was evaluated by using the akmedoids package version 1.3.0 (Adepeju et.al, 2021). K-means clustering was performed using the factoextra version 1.0.3 package (Kassambara and Mundt, 2020). Imputation that was not performed automatically by KML was done via Simpuation package version 0.2.8 (van der Loo M, 2022).

3. Results

3.1 Summary of Baseline Characteristics

Table 1: Baseline Demographics

Variable	**N = 872**[1]
Age at visit	47 (41, 53)
Racial/Ethnic Group	
White Non-Hispanic	101 (12%)
Black Non-Hispanic	535 (61%)
Hispanic	204 (23%)
Other Non-Hispanic	32 (3.7%)
Body Mass Index (kg/m^2)	30 (25, 36)
Drinks Per Week	
Abstainer	563 (65%)
0-7	279 (32%)
> 7	22 (2.5%)
Unknown	8
Waist Circumference (cm)	97 (86, 110)
Unknown	3
Education Level	
< High School degree	296 (34%)
High School degree	252 (29%)
> High School degree	318 (37%)
Unknown	6

[1] Median (Q1, Q3); n (%)

872 WWH were followed over varying follow-up periods with their key liver indicators measured. 12% were White Non-Hispanic, 61% were Black Non-Hispanic, 23% were Hispanic, and 3.7% were Other Non-Hispanic. Overall, the median age was 47 years (Q1 = 41, Q3 = 53) among the participants at baseline. 65% abstained from alcohol, with only 2.5% self-reported consuming more than seven drinks a week. The median BMI at baseline was 30 kg/m^2 (25, 36) and the median waist circumference was 97 cm (86, 110). 34% had less than a high school level of education at baseline, with 29% and 37% having a high school degree or more, respectively (Table 1).

Table 2: Baseline Clinical Characteristics

Variable	**N = 872**[1]
NFS Score	-1.76 (-2.59, -0.76)
Unknown	6
FIB-4 Score	0.92 (0.69, 1.24)
APRI Score	0.20 (0.15, 0.27)
% CD4 positive cells (helpers)	35 (29, 42)
Unknown	5
Hierarchical ART type at visit	
1 PI	443 (51%)
2 NNRTI	413 (47%)
3 Other	16 (1.8%)
Study Group	
Control	549 (63%)
INSTI	323 (37%)

[1] Median (Q1, Q3); n (%); PI = Protease Inhibitor; NNRTI = Non-Nucleoside Reverse Transcriptase Inhibitor

The median NFS, FIB-4, and APRI scores at baseline were -1.76 (-2.59, -0.76), 0.92 (0.69, 1.24), and 0.20 (0.15, 0.27) respectively. The % of CD4 cells also had an overall median of 35% (29, 42) at baseline. 63% were a part of the control cohort and 37% were a part of the INSTI cohort (Table 2).

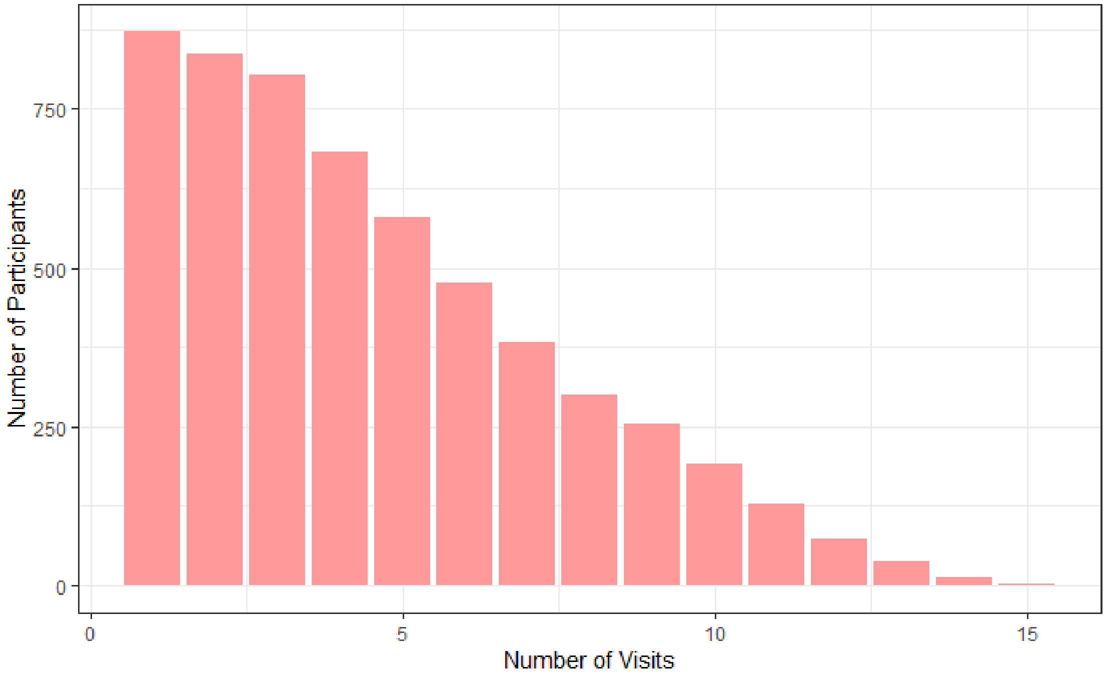


Figure 1: Distribution of Visits Among 872 WIHS Participants

The number of completed visits drops off around the fifth visit. This trend continues until there are about no participants having completed their 15th follow-up visit (Figure 1).

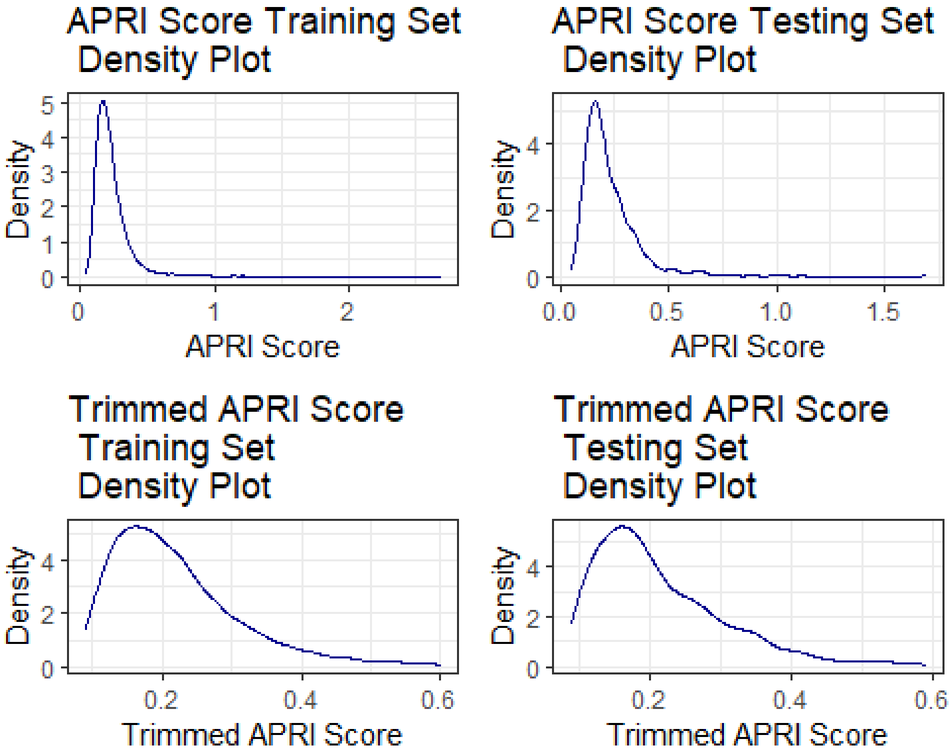


Figure 2: Density Plots of APRI and Trimmed APRI Scores Among the Training and Testing Sets

As seen in the plots of Figure 2, the distribution of the APRI score between both the training and testing sets is right skewed and non-zero. This skewness is alleviated when 5% of the total APRI observations are trimmed, as shown by bottom two plots of the figure. Regardless, the distribution of APRI is consistent between the two sets (Figure 2).

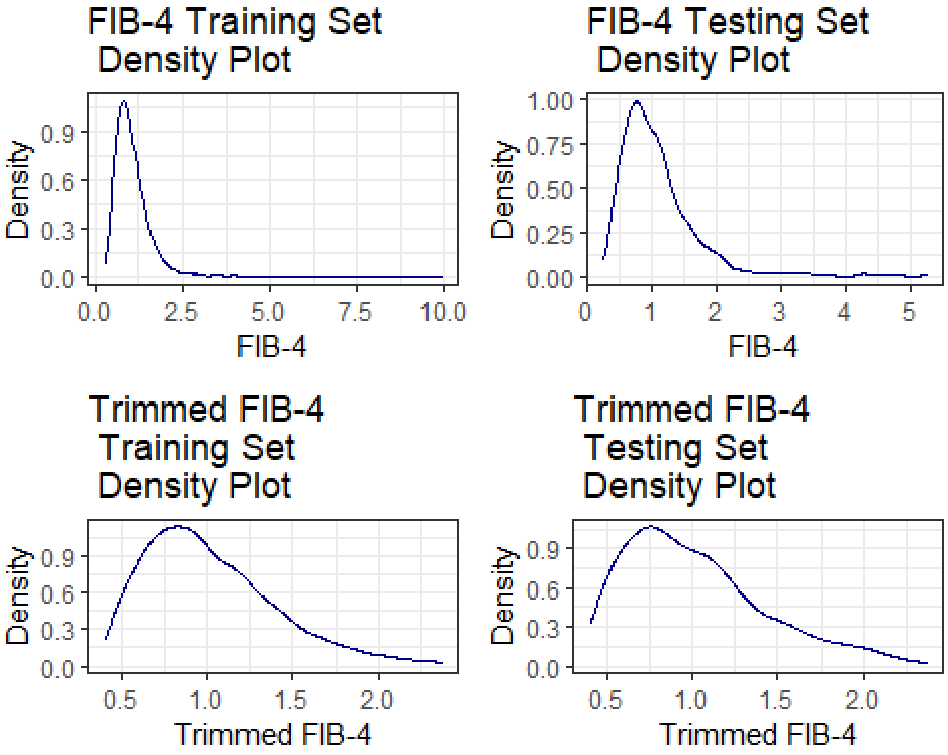


Figure 3: Density Plots of FIB-4 and Trimmed FIB-4 Scores Among the Training and Testing Sets

As seen in the plots of Figure 3, the distribution of FIB-4 between both the training and testing sets is right skewed and non-zero. This skewness is ameliorated when 5% of the total FIB-4 observations are trimmed, shown by bottom two plots of the figure. Regardless, the distribution of FIB-4 is also consistent between the two sets (Figure 3).

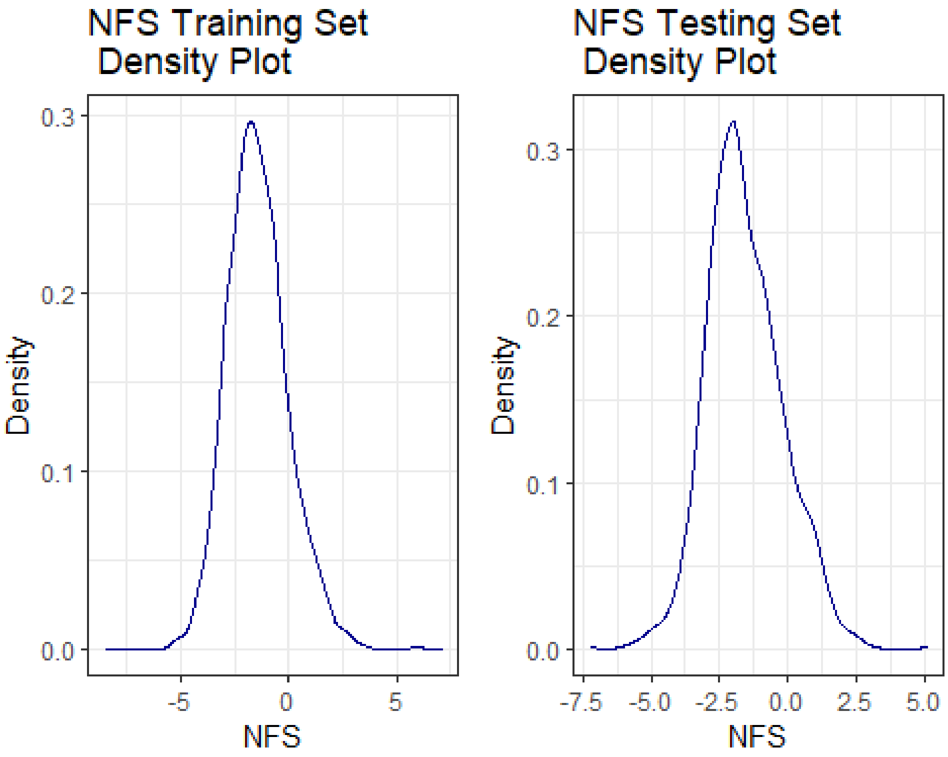


Figure 4: Density Plots of NFS Scores Among the Training and Testing Sets

Figure 4 displays the distribution of NFS between the training and testing sets. NFS is normally distributed within both the training and testing sets and has a range that encompasses zero, unlike APRI and FIB-4 (Figure 4).

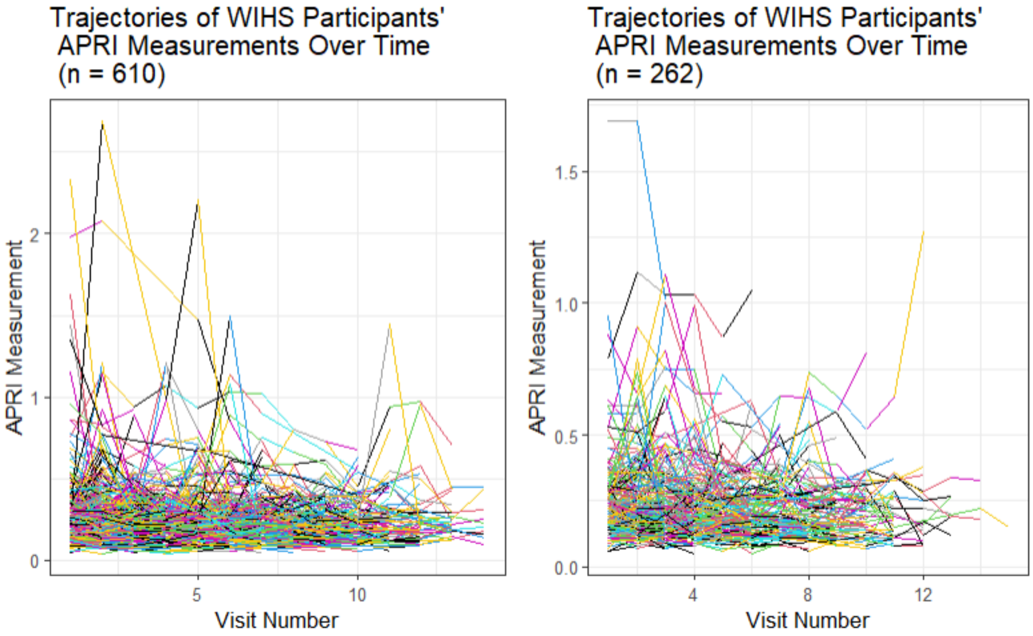


Figure 5: Trajectories of APRI Scores Among the Training (left) and Testing Set (right)

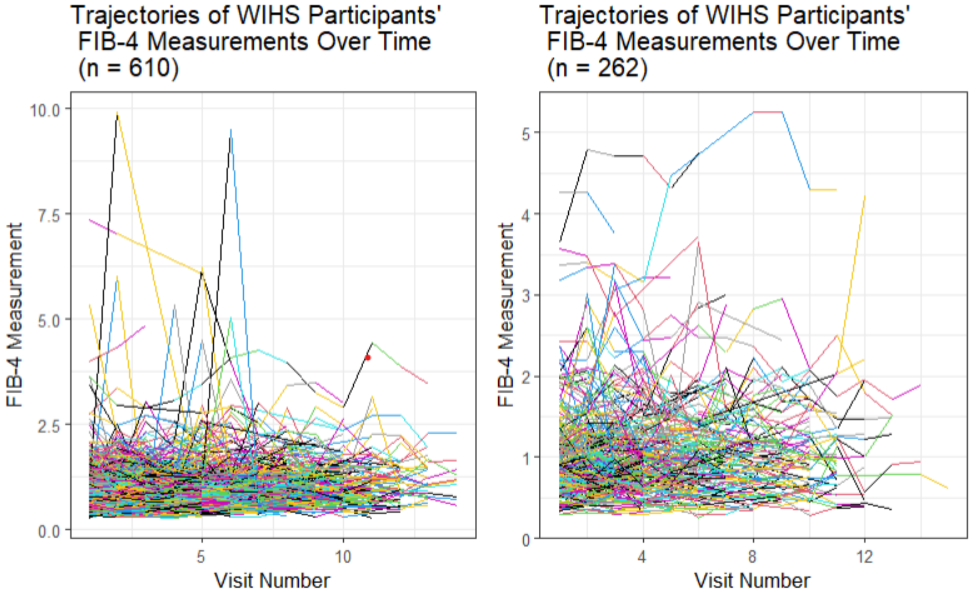


Figure 6: Trajectories of FIB-4 scores Among the Training (left) and Testing Set (right)

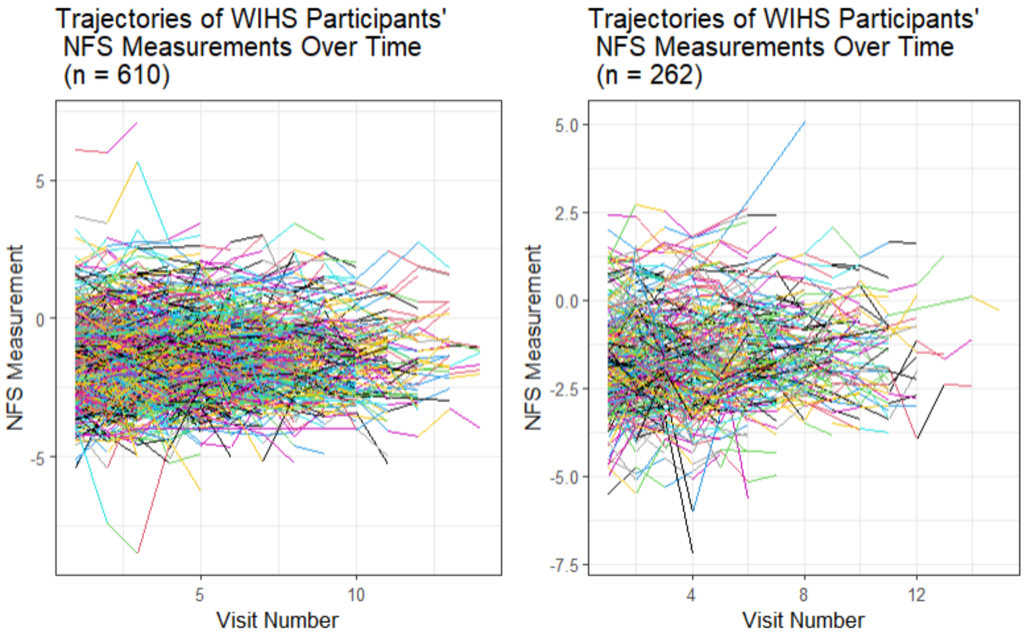
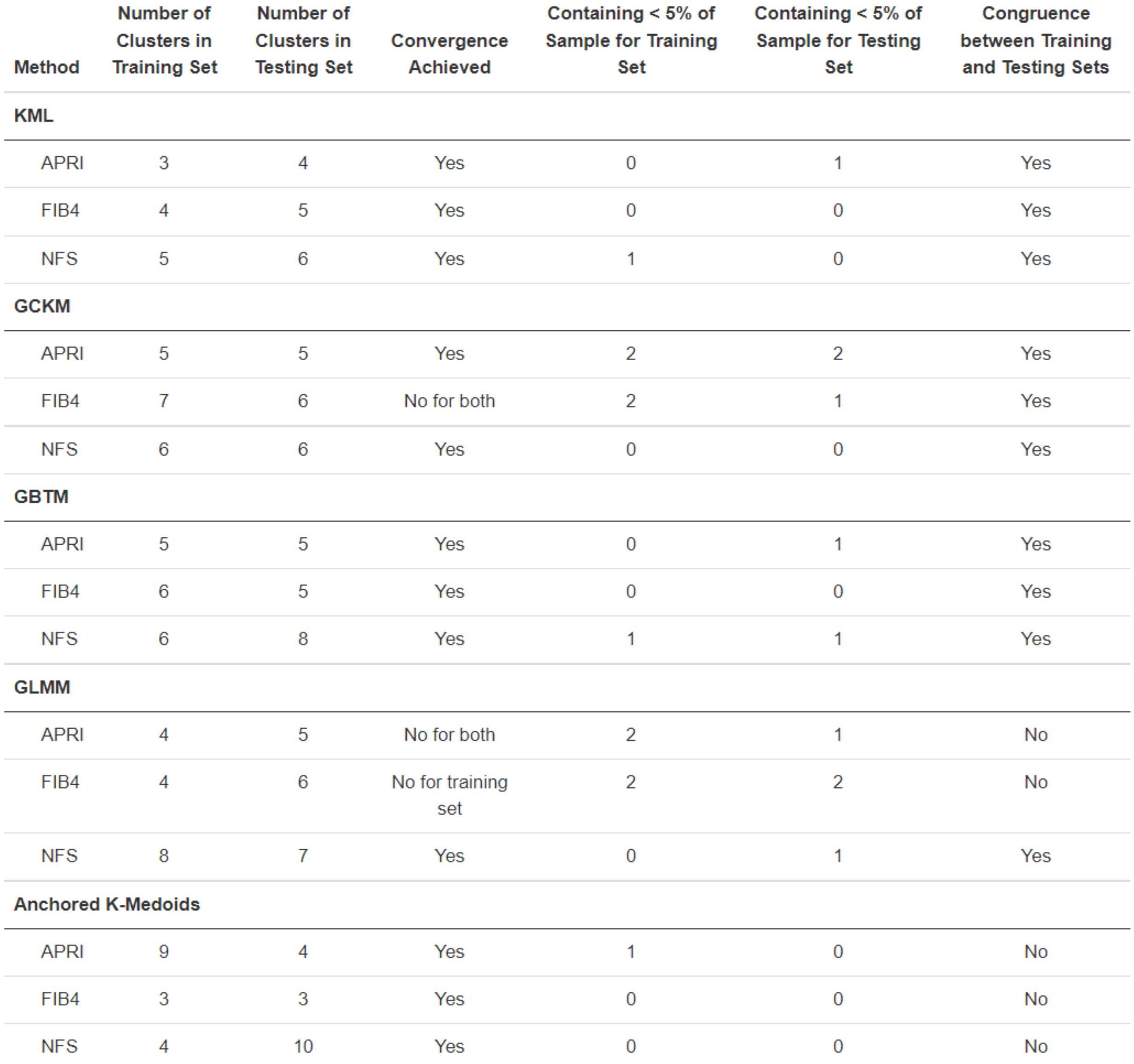


Figure 7: Trajectories of NFS scores Among the Training (left) and Testing Set (right)

The overall shape of the trajectories appears consistent between the two sets, aside from outliers in both sets for APRI and FIB-4 (Figures 5-7).

3.2 Comparing the results of the five methods

Table 3: Overall Results of Longitudinal Clustering Methods



Method	Number of Clusters in Training Set	Number of Clusters in Testing Set	Convergence Achieved	Containing < 5% of Sample for Training Set	Containing < 5% of Sample for Testing Set	Congruence between Training and Testing Sets
KML						
APRI	3	4	Yes	0	1	Yes
FIB4	4	5	Yes	0	0	Yes
NFS	5	6	Yes	1	0	Yes
GCKM						
APRI	5	5	Yes	2	2	Yes
FIB4	7	6	No for both	2	1	Yes
NFS	6	6	Yes	0	0	Yes
GBTM						
APRI	5	5	Yes	0	1	Yes
FIB4	6	5	Yes	0	0	Yes
NFS	6	8	Yes	1	1	Yes
GLMM						
APRI	4	5	No for both	2	1	No
FIB4	4	6	No for training set	2	2	No
NFS	8	7	Yes	0	1	Yes
Anchored K-Medoids						
APRI	9	4	Yes	1	0	No
FIB4	3	3	Yes	0	0	No
NFS	4	10	Yes	0	0	No

Table 3 summarizes the overall discrepancies and issues encountered for the five clustering methods. Please refer to the Appendix for the individual results for each clustering method.

Overall, the GLMM method showed the most issues with convergence. Convergence issues were also observed for FIB-4 with the GCKM method. GBTM, the other model-based clustering algorithm observed no issues with convergence. However, this method also returned empty clusters for the training sets. Because of their non-parametric nature, anchored k-medoids and longitudinal k-means are both guaranteed to converge. NFS never experienced any convergence issues for all five methods employed (Table 3).

The biggest discrepancy between the number of estimated clusters between the training and testing sets was observed with the anchored k-medoids approach, with a difference of five and six clusters for APRI and NFS, respectively. No method yielded an identical number of estimated clusters between the two sets. However, the GCKM method was the closest. Both the GLMM method and the GCKM method produced the largest number of clusters containing < 5% of participants (Table 3).

In terms of congruence in the average trajectories of the clusters between the two sets, GBTM performed the best with all three indicators matching the trends observed in both sets. Anchored k- medoids performed the worst, with all three liver indicators showing discordance in the trends between the two sets. APRI consistently showed discordance between the two sets for almost every method aside from the GBTM (Table 3).

Because of the agreement between the training and testing sets and lack of convergence issues and limited number of data imputation, the GBTM method was selected to cluster the combined 872 WIHS participants. This was done to address the question as to whether there would be agreement among the clusters for the three liver indicators.

3.3 Final GBTM Cluster Analysis

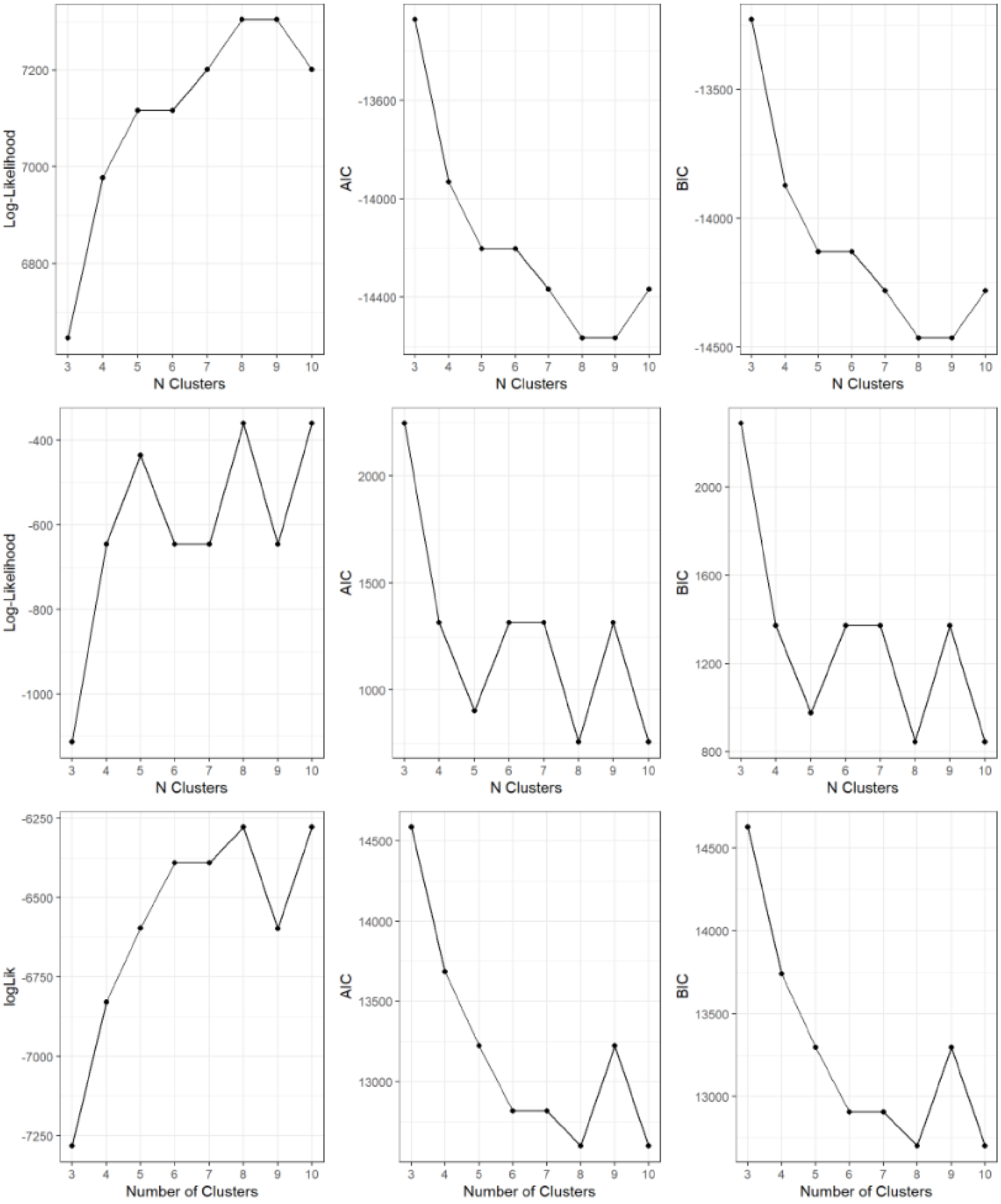


Figure 8: Quality Criteria for APRI, FIB-4, and NFS Group Based Trajectory Modeling Clustering (all 872 participants)

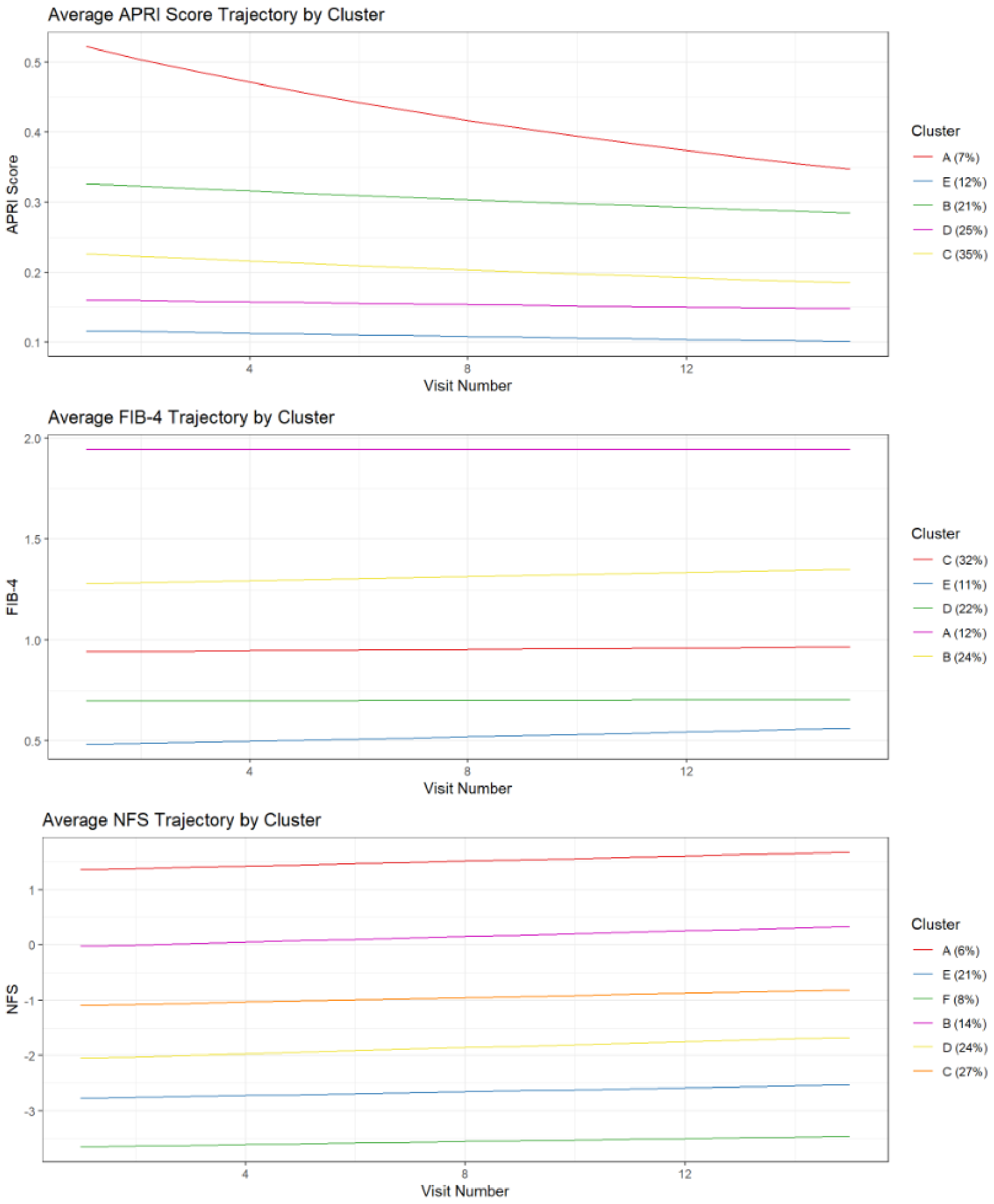


Figure 9: Average APRI, FIB-4, and NFS Trajectories by Cluster over Visits (all 872 participants)

GBTM's inability to function in the presence of missing values meant that 147 participants' missing NFS values were imputed using the means of the participant's trajectory.

As was the case for the training and testing sets, the distributions used for the GBTM models are the same. APRI and FIB-4 were modeled assuming a gamma distribution and NFS was modeled assuming a normal distribution.

Based on the plotted quality criteria and the corresponding elbow method, an optimal number of clusters chosen for APRI, FIB-4, and NFS were five, five, and six respectively (Figure 8). Participants APRI trajectories were partitioned into clusters containing 7, 12, 21, and 25% of participants. FIB-4 trajectories were partitioned into clusters containing 32, 11, 22, 12, and 24% of participants. NFS trajectories were partitioned into clusters containing 6, 21, 8, 14, 24, and 27% of participants. None of the clusters for the three liver indicators contained < 5% of participants (Figure 9).

Cluster A for APRI shows a slight decrease and remains the highest average APRI value among the five clusters. The four other clusters show a slight downward trend over time. All five clusters identified by FIB-4 show a slight increase over time in the average FIB-4 value. The same is observed for the six NFS clusters, with all six having a slightly higher trend than that of the FIB-4 clusters (Figure 9). These findings are consistent with those observed in both the training and testing sets for GBTM (Figures A.10 and A.12).

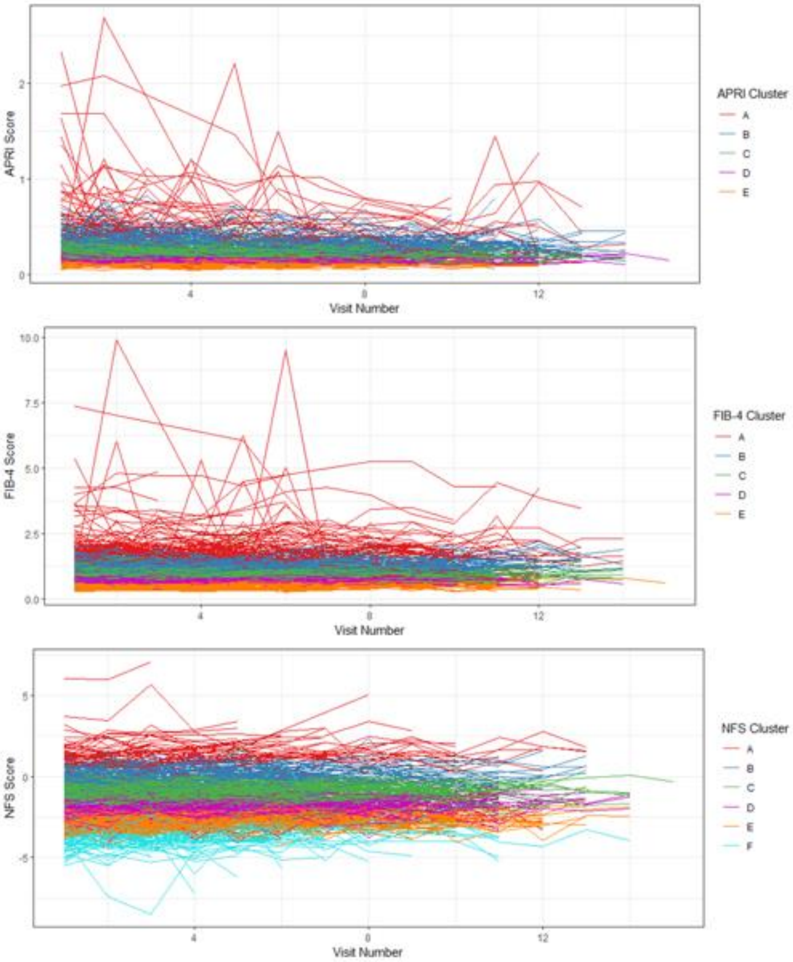
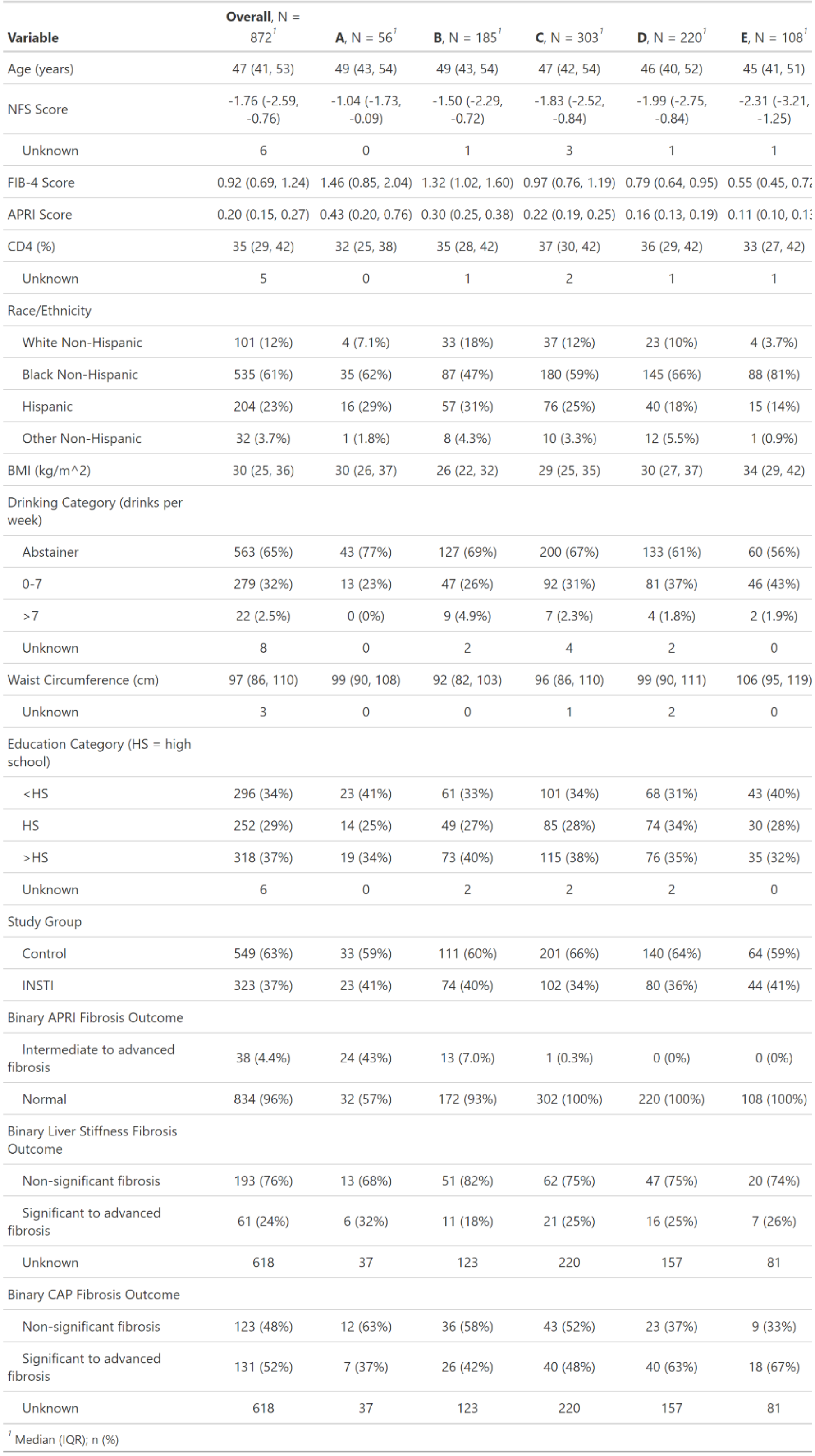


Figure 10: APRI, FIB-4, and NFS Trajectories by Clustering Result over Visits (all 872 participants)

Figure 10 plots the trajectory of all patients' APRI, FIB-4, and NFS scores over the follow-up period with the corresponding cluster assignments indicated by the colors above. To ensure clarity, cluster names are alphabetical letters (A-F), with A corresponding to the cluster with the highest average trajectory and E/F the lowest. The clusters possess little overlap amongst each other, as indicated by the figure above. Additionally, these clusters never cross each other, meaning, participants identified in cluster A were likely to have the highest average APRI score throughout the follow-up period (Figure 10).

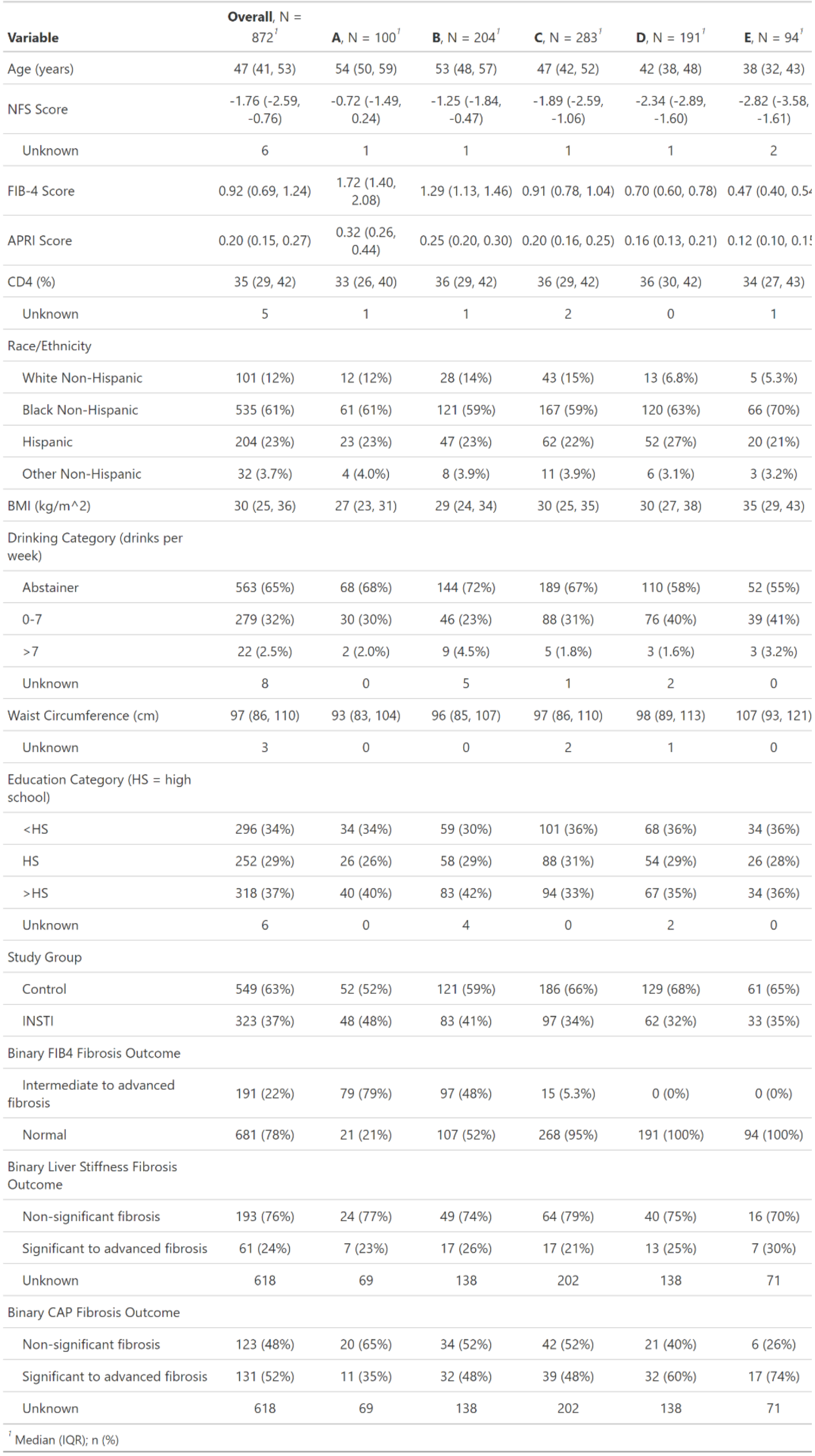
Table 4: Baseline Clinical and Demographic Variables by APRI Cluster Assignment



Variable	**Overall**, N = 872^1	**A**, N = 56^1	**B**, N = 185^1	**C**, N = 303^1	**D**, N = 220^1	**E**, N = 108^1
Age (years)	47 (41, 53)	49 (43, 54)	49 (43, 54)	47 (42, 54)	46 (40, 52)	45 (41, 51)
NFS Score	-1.76 (-2.59, -0.76)	-1.04 (-1.73, -0.09)	-1.50 (-2.29, -0.72)	-1.83 (-2.52, -0.84)	-1.99 (-2.75, -0.84)	-2.31 (-3.21, -1.25)
Unknown	6	0	1	3	1	1
FIB-4 Score	0.92 (0.69, 1.24)	1.46 (0.85, 2.04)	1.32 (1.02, 1.60)	0.97 (0.76, 1.19)	0.79 (0.64, 0.95)	0.55 (0.45, 0.7
APRI Score	0.20 (0.15, 0.27)	0.43 (0.20, 0.76)	0.30 (0.25, 0.38)	0.22 (0.19, 0.25)	0.16 (0.13, 0.19)	0.11 (0.10, 0.1
CD4 (%)	35 (29, 42)	32 (25, 38)	35 (28, 42)	37 (30, 42)	36 (29, 42)	33 (27, 42)
Unknown	5	0	1	2	1	1
Race/Ethnicity						
White Non-Hispanic	101 (12%)	4 (7.1%)	33 (18%)	37 (12%)	23 (10%)	4 (3.7%)
Black Non-Hispanic	535 (61%)	35 (62%)	87 (47%)	180 (59%)	145 (66%)	88 (81%)
Hispanic	204 (23%)	16 (29%)	57 (31%)	76 (25%)	40 (18%)	15 (14%)
Other Non-Hispanic	32 (3.7%)	1 (1.8%)	8 (4.3%)	10 (3.3%)	12 (5.5%)	1 (0.9%)
BMI (kg/m^2)	30 (25, 36)	30 (26, 37)	26 (22, 32)	29 (25, 35)	30 (27, 37)	34 (29, 42)
Drinking Category (drinks per week)						
Abstainer	563 (65%)	43 (77%)	127 (69%)	200 (67%)	133 (61%)	60 (56%)
0-7	279 (32%)	13 (23%)	47 (26%)	92 (31%)	81 (37%)	46 (43%)
>7	22 (2.5%)	0 (0%)	9 (4.9%)	7 (2.3%)	4 (1.8%)	2 (1.9%)
Unknown	8	0	2	4	2	0
Waist Circumference (cm)	97 (86, 110)	99 (90, 108)	92 (82, 103)	96 (86, 110)	99 (90, 111)	106 (95, 119)
Unknown	3	0	0	1	2	0
Education Category (HS = high school)						
<HS	296 (34%)	23 (41%)	61 (33%)	101 (34%)	68 (31%)	43 (40%)
HS	252 (29%)	14 (25%)	49 (27%)	85 (28%)	74 (34%)	30 (28%)
>HS	318 (37%)	19 (34%)	73 (40%)	115 (38%)	76 (35%)	35 (32%)
Unknown	6	0	2	2	2	0
Study Group						
Control	549 (63%)	33 (59%)	111 (60%)	201 (66%)	140 (64%)	64 (59%)
INSTI	323 (37%)	23 (41%)	74 (40%)	102 (34%)	80 (36%)	44 (41%)
Binary APRI Fibrosis Outcome						
Intermediate to advanced fibrosis	38 (4.4%)	24 (43%)	13 (7.0%)	1 (0.3%)	0 (0%)	0 (0%)
Normal	834 (96%)	32 (57%)	172 (93%)	302 (100%)	220 (100%)	108 (100%)
Binary Liver Stiffness Fibrosis Outcome						
Non-significant fibrosis	193 (76%)	13 (68%)	51 (82%)	62 (75%)	47 (75%)	20 (74%)
Significant to advanced fibrosis	61 (24%)	6 (32%)	11 (18%)	21 (25%)	16 (25%)	7 (26%)
Unknown	618	37	123	220	157	81
Binary CAP Fibrosis Outcome						
Non-significant fibrosis	123 (48%)	12 (63%)	36 (58%)	43 (52%)	23 (37%)	9 (33%)
Significant to advanced fibrosis	131 (52%)	7 (37%)	26 (42%)	40 (48%)	40 (63%)	18 (67%)
Unknown	618	37	123	220	157	81

^{1} Median (IQR); n (%)

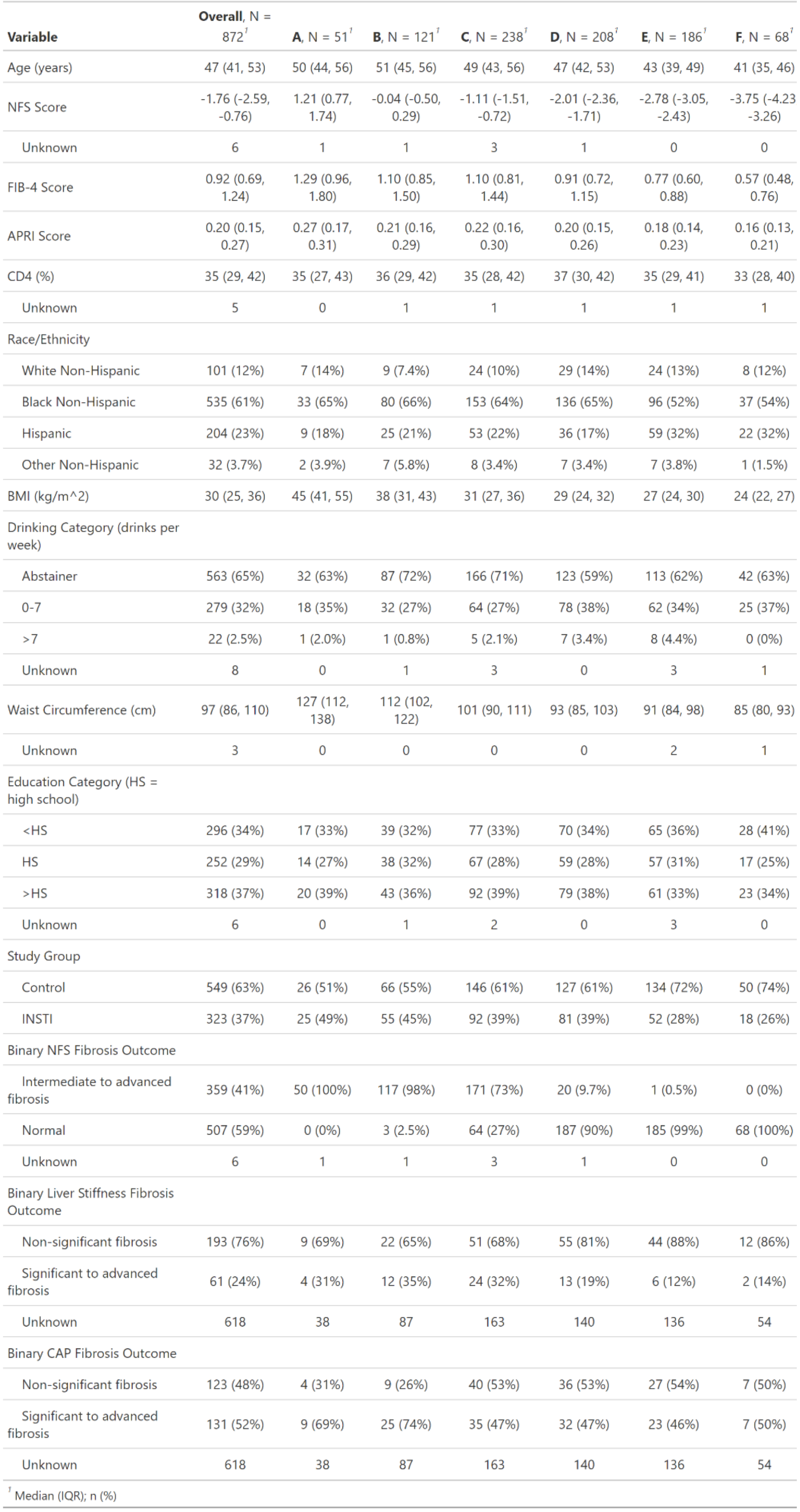
Table 5: Baseline Clinical and Demographic Variables by FIB-4 Cluster Assignment



Variable	**Overall**, N = 872[1]	**A**, N = 100[1]	**B**, N = 204[1]	**C**, N = 283[1]	**D**, N = 191[1]	**E**, N = 94[1]
Age (years)	47 (41, 53)	54 (50, 59)	53 (48, 57)	47 (42, 52)	42 (38, 48)	38 (32, 43)
NFS Score	-1.76 (-2.59, -0.76)	-0.72 (-1.49, 0.24)	-1.25 (-1.84, -0.47)	-1.89 (-2.59, -1.06)	-2.34 (-2.89, -1.60)	-2.82 (-3.58, -1.61)
Unknown	6	1	1	1	1	2
FIB-4 Score	0.92 (0.69, 1.24)	1.72 (1.40, 2.08)	1.29 (1.13, 1.46)	0.91 (0.78, 1.04)	0.70 (0.60, 0.78)	0.47 (0.40, 0.54
APRI Score	0.20 (0.15, 0.27)	0.32 (0.26, 0.44)	0.25 (0.20, 0.30)	0.20 (0.16, 0.25)	0.16 (0.13, 0.21)	0.12 (0.10, 0.15
CD4 (%)	35 (29, 42)	33 (26, 40)	36 (29, 42)	36 (29, 42)	36 (30, 42)	34 (27, 43)
Unknown	5	1	1	2	0	1
Race/Ethnicity						
White Non-Hispanic	101 (12%)	12 (12%)	28 (14%)	43 (15%)	13 (6.8%)	5 (5.3%)
Black Non-Hispanic	535 (61%)	61 (61%)	121 (59%)	167 (59%)	120 (63%)	66 (70%)
Hispanic	204 (23%)	23 (23%)	47 (23%)	62 (22%)	52 (27%)	20 (21%)
Other Non-Hispanic	32 (3.7%)	4 (4.0%)	8 (3.9%)	11 (3.9%)	6 (3.1%)	3 (3.2%)
BMI (kg/m^2)	30 (25, 36)	27 (23, 31)	29 (24, 34)	30 (25, 35)	30 (27, 38)	35 (29, 43)
Drinking Category (drinks per week)						
Abstainer	563 (65%)	68 (68%)	144 (72%)	189 (67%)	110 (58%)	52 (55%)
0-7	279 (32%)	30 (30%)	46 (23%)	88 (31%)	76 (40%)	39 (41%)
>7	22 (2.5%)	2 (2.0%)	9 (4.5%)	5 (1.8%)	3 (1.6%)	3 (3.2%)
Unknown	8	0	5	1	2	0
Waist Circumference (cm)	97 (86, 110)	93 (83, 104)	96 (85, 107)	97 (86, 110)	98 (89, 113)	107 (93, 121)
Unknown	3	0	0	2	1	0
Education Category (HS = high school)						
<HS	296 (34%)	34 (34%)	59 (30%)	101 (36%)	68 (36%)	34 (36%)
HS	252 (29%)	26 (26%)	58 (29%)	88 (31%)	54 (29%)	26 (28%)
>HS	318 (37%)	40 (40%)	83 (42%)	94 (33%)	67 (35%)	34 (36%)
Unknown	6	0	4	0	2	0
Study Group						
Control	549 (63%)	52 (52%)	121 (59%)	186 (66%)	129 (68%)	61 (65%)
INSTI	323 (37%)	48 (48%)	83 (41%)	97 (34%)	62 (32%)	33 (35%)
Binary FIB4 Fibrosis Outcome						
Intermediate to advanced fibrosis	191 (22%)	79 (79%)	97 (48%)	15 (5.3%)	0 (0%)	0 (0%)
Normal	681 (78%)	21 (21%)	107 (52%)	268 (95%)	191 (100%)	94 (100%)
Binary Liver Stiffness Fibrosis Outcome						
Non-significant fibrosis	193 (76%)	24 (77%)	49 (74%)	64 (79%)	40 (75%)	16 (70%)
Significant to advanced fibrosis	61 (24%)	7 (23%)	17 (26%)	17 (21%)	13 (25%)	7 (30%)
Unknown	618	69	138	202	138	71
Binary CAP Fibrosis Outcome						
Non-significant fibrosis	123 (48%)	20 (65%)	34 (52%)	42 (52%)	21 (40%)	6 (26%)
Significant to advanced fibrosis	131 (52%)	11 (35%)	32 (48%)	39 (48%)	32 (60%)	17 (74%)
Unknown	618	69	138	202	138	71

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

Table 6: Baseline Clinical and Demographic Variables by NFS Cluster Assignment



Variable	**Overall**, N = 872[1]	**A**, N = 51[1]	**B**, N = 121[1]	**C**, N = 238[1]	**D**, N = 208[1]	**E**, N = 186[1]	**F**, N = 68[1]
Age (years)	47 (41, 53)	50 (44, 56)	51 (45, 56)	49 (43, 56)	47 (42, 53)	43 (39, 49)	41 (35, 46)
NFS Score	-1.76 (-2.59, -0.76)	1.21 (0.77, 1.74)	-0.04 (-0.50, 0.29)	-1.11 (-1.51, -0.72)	-2.01 (-2.36, -1.71)	-2.78 (-3.05, -2.43)	-3.75 (-4.23 -3.26)
Unknown	6	1	1	3	1	0	0
FIB-4 Score	0.92 (0.69, 1.24)	1.29 (0.96, 1.80)	1.10 (0.85, 1.50)	1.10 (0.81, 1.44)	0.91 (0.72, 1.15)	0.77 (0.60, 0.88)	0.57 (0.48, 0.76)
APRI Score	0.20 (0.15, 0.27)	0.27 (0.17, 0.31)	0.21 (0.16, 0.29)	0.22 (0.16, 0.30)	0.20 (0.15, 0.26)	0.18 (0.14, 0.23)	0.16 (0.13, 0.21)
CD4 (%)	35 (29, 42)	35 (27, 43)	36 (29, 42)	35 (28, 42)	37 (30, 42)	35 (29, 41)	33 (28, 40)
Unknown	5	0	1	1	1	1	1
Race/Ethnicity							
White Non-Hispanic	101 (12%)	7 (14%)	9 (7.4%)	24 (10%)	29 (14%)	24 (13%)	8 (12%)
Black Non-Hispanic	535 (61%)	33 (65%)	80 (66%)	153 (64%)	136 (65%)	96 (52%)	37 (54%)
Hispanic	204 (23%)	9 (18%)	25 (21%)	53 (22%)	36 (17%)	59 (32%)	22 (32%)
Other Non-Hispanic	32 (3.7%)	2 (3.9%)	7 (5.8%)	8 (3.4%)	7 (3.4%)	7 (3.8%)	1 (1.5%)
BMI (kg/m^2)	30 (25, 36)	45 (41, 55)	38 (31, 43)	31 (27, 36)	29 (24, 32)	27 (24, 30)	24 (22, 27)
Drinking Category (drinks per week)							
Abstainer	563 (65%)	32 (63%)	87 (72%)	166 (71%)	123 (59%)	113 (62%)	42 (63%)
0-7	279 (32%)	18 (35%)	32 (27%)	64 (27%)	78 (38%)	62 (34%)	25 (37%)
>7	22 (2.5%)	1 (2.0%)	1 (0.8%)	5 (2.1%)	7 (3.4%)	8 (4.4%)	0 (0%)
Unknown	8	0	1	3	0	3	1
Waist Circumference (cm)	97 (86, 110)	127 (112, 138)	112 (102, 122)	101 (90, 111)	93 (85, 103)	91 (84, 98)	85 (80, 93)
Unknown	3	0	0	0	0	2	1
Education Category (HS = high school)							
<HS	296 (34%)	17 (33%)	39 (32%)	77 (33%)	70 (34%)	65 (36%)	28 (41%)
HS	252 (29%)	14 (27%)	38 (32%)	67 (28%)	59 (28%)	57 (31%)	17 (25%)
>HS	318 (37%)	20 (39%)	43 (36%)	92 (39%)	79 (38%)	61 (33%)	23 (34%)
Unknown	6	0	1	2	0	3	0
Study Group							
Control	549 (63%)	26 (51%)	66 (55%)	146 (61%)	127 (61%)	134 (72%)	50 (74%)
INSTI	323 (37%)	25 (49%)	55 (45%)	92 (39%)	81 (39%)	52 (28%)	18 (26%)
Binary NFS Fibrosis Outcome							
Intermediate to advanced fibrosis	359 (41%)	50 (100%)	117 (98%)	171 (73%)	20 (9.7%)	1 (0.5%)	0 (0%)
Normal	507 (59%)	0 (0%)	3 (2.5%)	64 (27%)	187 (90%)	185 (99%)	68 (100%)
Unknown	6	1	1	3	1	0	0
Binary Liver Stiffness Fibrosis Outcome							
Non-significant fibrosis	193 (76%)	9 (69%)	22 (65%)	51 (68%)	55 (81%)	44 (88%)	12 (86%)
Significant to advanced fibrosis	61 (24%)	4 (31%)	12 (35%)	24 (32%)	13 (19%)	6 (12%)	2 (14%)
Unknown	618	38	87	163	140	136	54
Binary CAP Fibrosis Outcome							
Non-significant fibrosis	123 (48%)	4 (31%)	9 (26%)	40 (53%)	36 (53%)	27 (54%)	7 (50%)
Significant to advanced fibrosis	131 (52%)	9 (69%)	25 (74%)	35 (47%)	32 (47%)	23 (46%)	7 (50%)
Unknown	618	38	87	163	140	136	54

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

The differences amongst the clusters identified by APRI are minimal to non-existent regarding demographics. Clusters A and B contain a majority of APRI fibrosis cases, with cluster A containing a 60-40 split of fibrosis and non-fibrosis. While no trend is observed in LS fibrosis outcome, there appears to be an upward trend in CAP fibrosis as you go from cluster A to cluster E, with cluster A having 37% of participants diagnosed with CAP fibrosis vs. 67% of cluster E (Table 4).

As for FIB-4, no noticeable trends across the clusters can be ascertained aside from the following. BMI and waist circumference show a slight upward trend as you move from cluster A to cluster E, with a median BMI of 27 (23, 21) and 35 kg/m^2 (29, 43) respectively. Clusters A and B have nearly all of the FIB-4 fibrosis cases, of which cluster A has an 80-20 split between fibrosis and non-fibrosis. As with APRI, there is a difference in the frequency of CAP fibrosis between clusters A and E, with 35% vs. 74% respectively. It is worth noting that clusters B-D, which contain most of the participants, do not show a discernable trend (Table 5).

The clusters assigned via NFS show discrepancies in their characteristics to those assigned by FIB-4 and APRI. NFS demonstrates a consistent downward trend of BMI and waist circumference from cluster A to cluster F with a median waist circumference of 127cm (112, 138) to 85cm (80, 93) respectively. While cluster A possesses a larger proportion of INSTI participants relative to the overall study population (49 vs. 37%), clusters C-E show minimal trends. Noticeably, clusters A and B are comprised almost solely of participants diagnosed with liver fibrosis based off their NFS score. Cluster C also has 70% of it's composition with NFS scores large enough for a fibrosis category. Clusters D-F contain less than 10% of these participants. There are slight to non-existent trends observed across the clusters for CAP and LS fibrosis categories (Table 6).

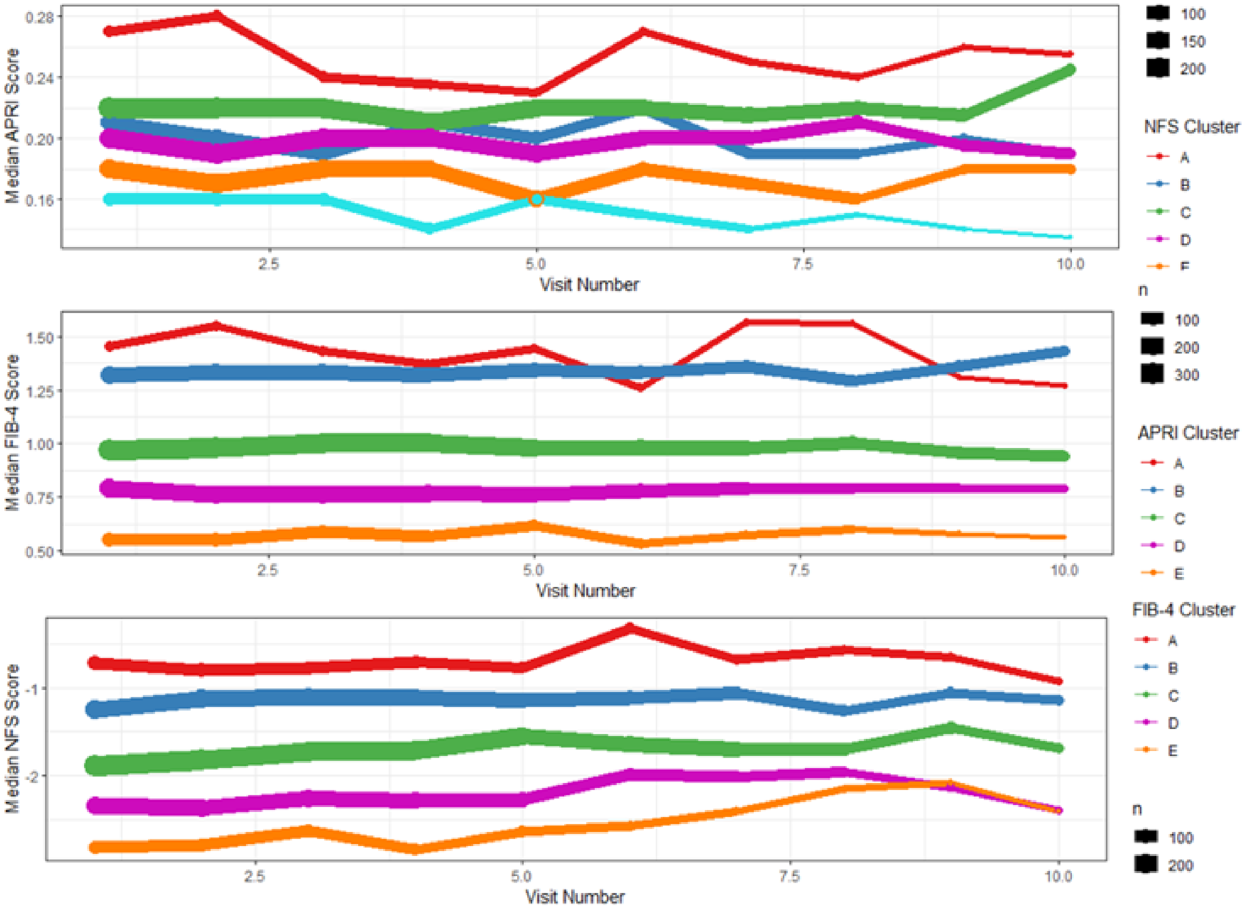


Figure 11: Pairwise of Median APRI, FIB-4, and NFS Trajectories by Clustering Result over Visits (all 872 participants)

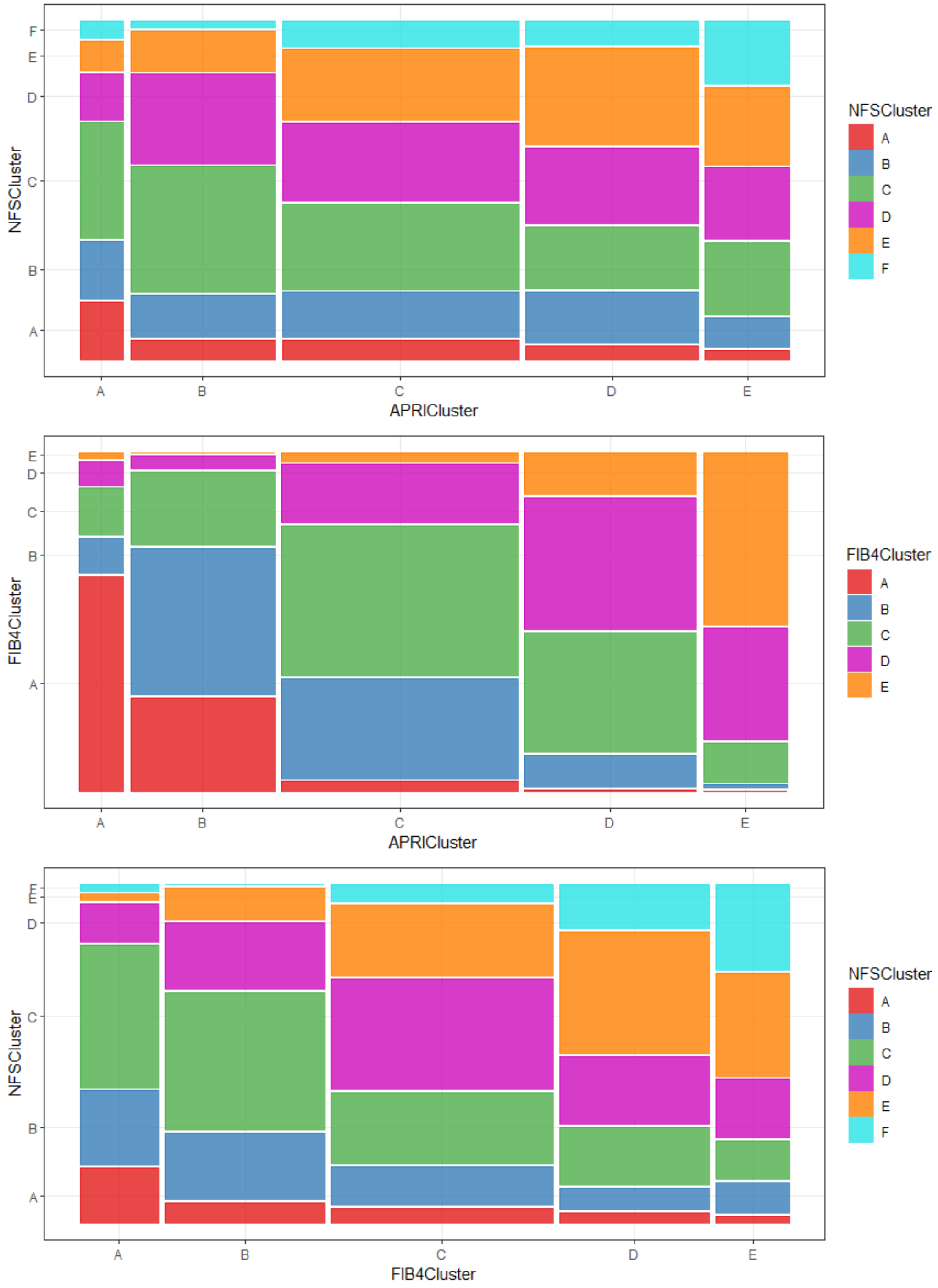


Figure 12: Mosaic Plots of APRI, FIB-4, and NFS Clusters Cross-Tabbed

The cluster assignments from APRI appeared to have less association with the trajectory of a patient’s NFS score, as indicated by the overlapping median APRI for NFS clusters B-D. This contrasts with the near perfect overlap in the median FIB-4 trajectories of the clusters identified by APRI and median NFS trajectories identified by FIB-4. Thus, the FIB-4 clusters appear more aligned with the APRI and NFS clusters (Figure 11).

While the median trajectories of the clusters align aside from APRI and NFS, cluster membership shows more discordance. This is most notable for APRI and NFS, where the NFS cluster assignments are spread out across the varying APRI clusters. The same can be seen for APRI and FIB-4, where cluster C for APRI comprises a large portion of all the FIB-4 clusters (Figure 12).

3.4 LIVRA Clustering

Table 7: Demographics of LIVRA Participants

Characteristic	**N = 254[1]**
Age at visit (years)	50 (44, 55)
% CD4 positive cells (helpers)	38 (32, 44)
Race/Ethnicity	
White Non-Hispanic	27 (11%)
Black Non-Hispanic	186 (73%)
Hispanic	27 (11%)
Other Non-Hispanic	14 (5.5%)
Body Mass Index (kg/m^2)	31 (27, 36)
Drinking Category (drinks per week)	
Abstainer	157 (62%)
0-7	92 (36%)
> 7	4 (1.6%)
Unknown	1
CAP (Db/m)	250 (215, 294)
LS (kPa)	5.40 (4.10, 6.88)
Level of Education (HS = high school)	
<HS	70 (28%)
HS	84 (33%)
>HS	100 (39%)
Study Group	
Control	134 (53%)
INSTI	120 (47%)

[1] Median (Q1, Q3); n (%)

There were 254 eligible participants who had obtained a Fibroscan visit. Three participants were excluded due to having biologically implausible CAP values of 0. 73% of LIVRA participants were Black non-Hispanic, 11% were White non-Hispanic, 11% were Hispanic and 5.5% were other non-Hispanic. The median age observed among the LIVRA participants was 50 years (44, 55). LIVRA participants had a median BMI of 31 kg/m^2 (27, 36). Additionally, LIVRA participants had a median CD4 % of 38% (32, 44). 98% of participants were either abstainers or consumed less than seven drinks a week. 28% had less than a high school level of education, while 33 and 39% had either a high school degree or more, respectively (Table 4). The median LS and CAP was 5.40 kPa (4.10, 6.88) and 250 Db/m (215, 294), respectively. Overall, 53% of participants were in the control group and 47% were in the INSTI study group (Table 4).

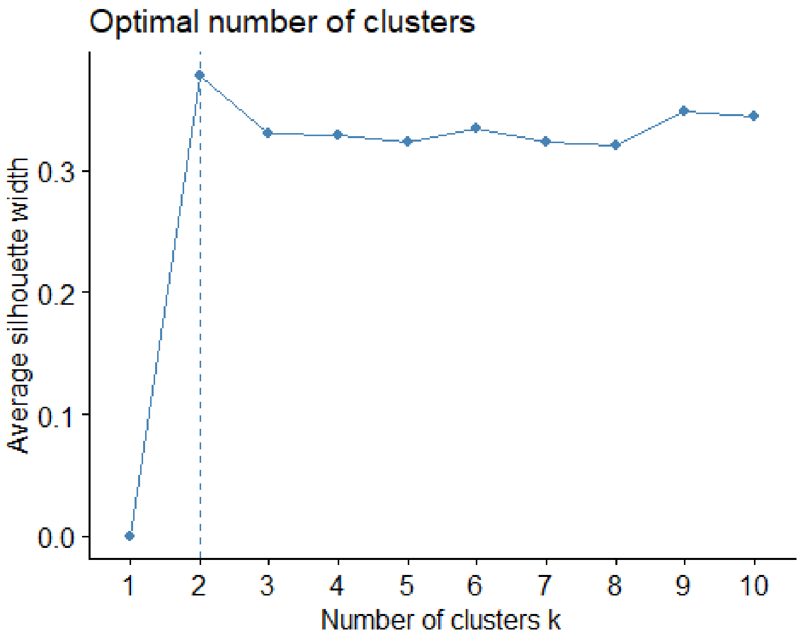


Figure 13: Average Silhouette Width Among Number of Clusters

Based on the plotted average silhouette width against the proposed number of clusters and the elbow method, an optimal solution of two clusters was chosen (Figure 13). The two-cluster solution partitioned participants into either cluster A or B, which contained 60 and 40% of participants respectively.

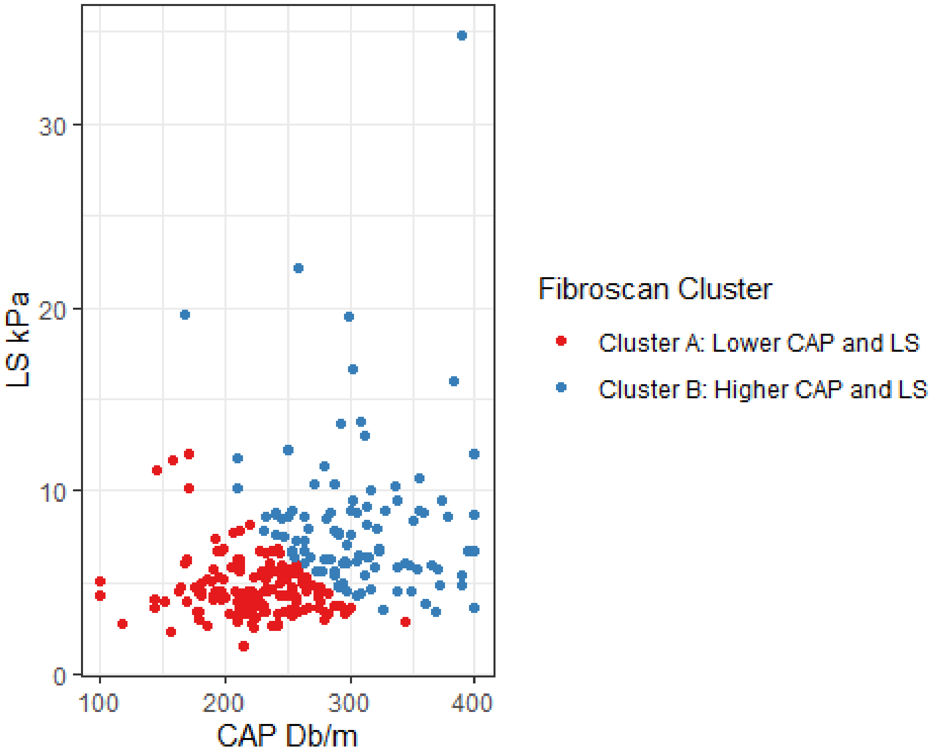
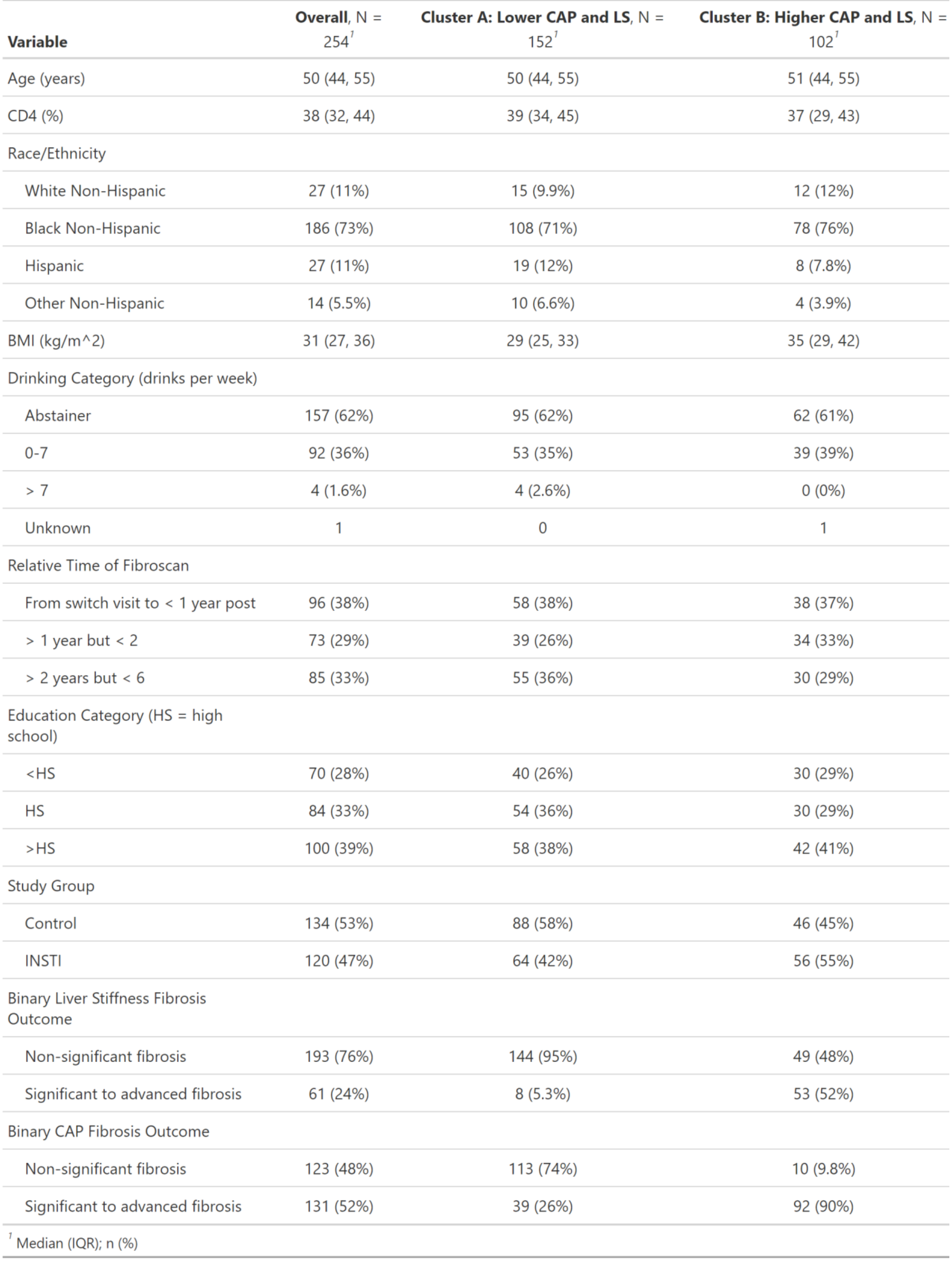


Figure 14: Values of LS vs. CAP by Cluster Assignment

Membership in cluster A appeared to be associated with a lower CAP and LS values than those within cluster B. However, this partition between the two outcomes is not perfect, as overlap exists between those having a high CAP also containing a lower LS value (Figure 14).

Table 8: Clinical and Demographic Variables by Fibroscan Cluster Assignment



Variable	**Overall**, N = 254[1]	**Cluster A: Lower CAP and LS**, N = 152[1]	**Cluster B: Higher CAP and LS**, N = 102[1]
Age (years)	50 (44, 55)	50 (44, 55)	51 (44, 55)
CD4 (%)	38 (32, 44)	39 (34, 45)	37 (29, 43)
Race/Ethnicity			
White Non-Hispanic	27 (11%)	15 (9.9%)	12 (12%)
Black Non-Hispanic	186 (73%)	108 (71%)	78 (76%)
Hispanic	27 (11%)	19 (12%)	8 (7.8%)
Other Non-Hispanic	14 (5.5%)	10 (6.6%)	4 (3.9%)
BMI (kg/m^2)	31 (27, 36)	29 (25, 33)	35 (29, 42)
Drinking Category (drinks per week)			
Abstainer	157 (62%)	95 (62%)	62 (61%)
0-7	92 (36%)	53 (35%)	39 (39%)
> 7	4 (1.6%)	4 (2.6%)	0 (0%)
Unknown	1	0	1
Relative Time of Fibroscan			
From switch visit to < 1 year post	96 (38%)	58 (38%)	38 (37%)
> 1 year but < 2	73 (29%)	39 (26%)	34 (33%)
> 2 years but < 6	85 (33%)	55 (36%)	30 (29%)
Education Category (HS = high school)			
<HS	70 (28%)	40 (26%)	30 (29%)
HS	84 (33%)	54 (36%)	30 (29%)
>HS	100 (39%)	58 (38%)	42 (41%)
Study Group			
Control	134 (53%)	88 (58%)	46 (45%)
INSTI	120 (47%)	64 (42%)	56 (55%)
Binary Liver Stiffness Fibrosis Outcome			
Non-significant fibrosis	193 (76%)	144 (95%)	49 (48%)
Significant to advanced fibrosis	61 (24%)	8 (5.3%)	53 (52%)
Binary CAP Fibrosis Outcome			
Non-significant fibrosis	123 (48%)	113 (74%)	10 (9.8%)
Significant to advanced fibrosis	131 (52%)	39 (26%)	92 (90%)

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

The two clusters identified by Fibroscan share a remarkably similar composition of racial and ethnic groups, education levels, drinking category, and ages. The time of Fibroscan attainment is also consistent between the two clusters (Table 8).

BMI appears to be slightly higher in cluster B than cluster A, with a median of 35 kg/m^2 (29, 42) compared to 29 kg/m^2 (25, 33) respectively. Additionally, cluster B appears to have a slightly higher proportion of INSTI participants relative to the overall sample (55% vs. 47%). Cluster A consists mostly of non-significant fibrosis outcome based on LS (95%). Cluster B, on the other hand, has a nearly 50-50 split. The partitioning of fibrosis outcomes between the two clusters is more defined for CAP defined fibrosis, with cluster B having 90% of its participants diagnosed with CAP fibrosis compared to just 26% in cluster A (Table 8).

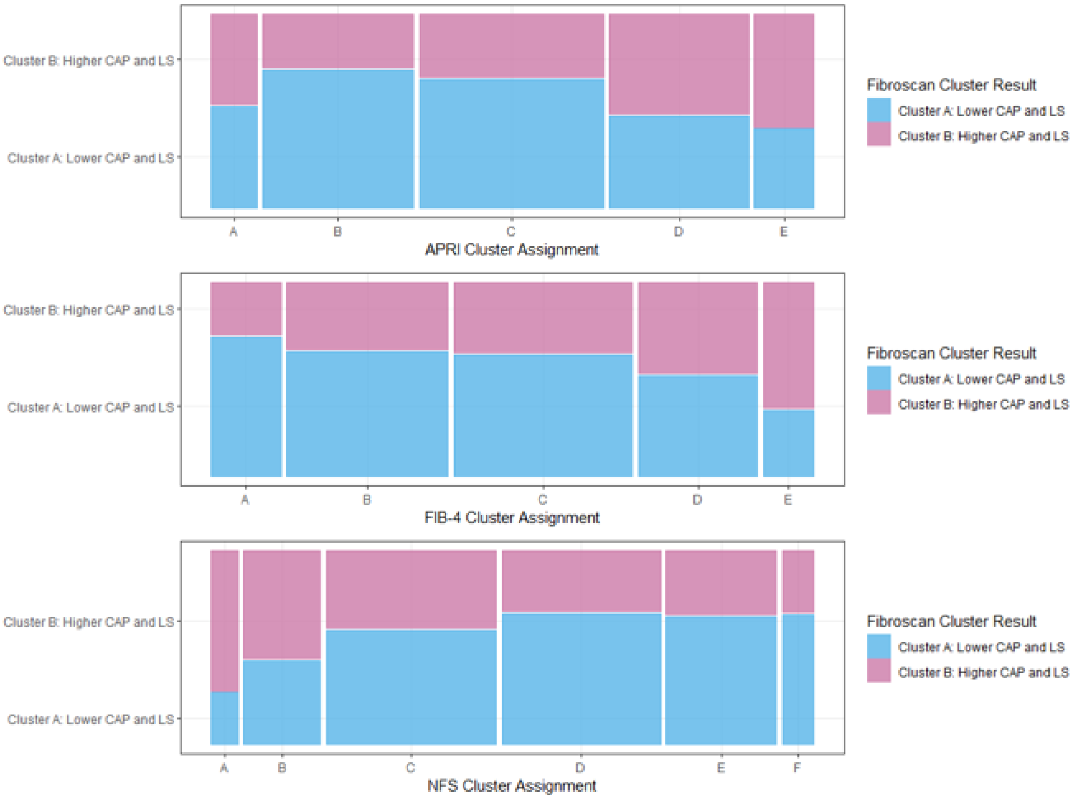


Figure 15: Mosaic Plots of APRI, FIB-4, and NFS Cluster Results Compared to Fibroscan Cluster Results

Across the liver measures, discrepancies between the identified clusters and the Fibroscan clusters exist. This is indicated by the presence of both Fibroscan clusters across all the clusters for each liver measure. The only notable exception is cluster A for NFS, which is mostly composed of participants with Fibroscan cluster B membership (Figure 15).

4. Discussion

4.1 Comparing longitudinal clustering methods

Three overarching objectives were addressed within this study. First, we compared several longitudinal clustering methods on the WIHS sub study to elucidate the effectiveness of these methods on real-world observational data. Second, we also utilized the clustering results to assess whether the relationship between a patient's liver scores (APRI, FIB-4, or NFS) were consistent with one another. Finally, we compared these results to the clustering results of a cross-sectional dataset of participants that received Fibroscan to further determine if INSTI use appeared to be associated with a higher liver fibrosis prevalence. worsening livers.

For our first objective, we compared a mix of different longitudinal clustering methods on three liver score trajectories. Cross-validation was performed by examining if the number of clusters identified were consistent between the training and testing sets. In this regard, four of the five methods performed well, with no more than two differences in clusters. Only the anchored k-medoids method performed poorly in this regard, with it estimating a difference of five clusters for APRI and six for NFS.

Additionally, we sought to see if the clusters identified in the training and testing sets possessed consistent behavior. If so, this indicated that such method was robust to sample sizes. In this regard, all the methods aside from the GLMM and anchored k- medoids performed well. These methods detected clusters that behaved similarly between the training and testing sets across three of the five methods applied (KML, GCKM, and GBTM).

Clinically speaking, clusters containing less than 5% of participants were deemed clinically irrelevant, as these reflected too small a portion of the sample. This was mostly an issue for those identified for APRI and FIB-4 by GCKM and GLMM, while the other methods employed did not experience this issue as drastically.

Convergence issues were observed with GLMM and GCKM, but not GBTM. They were most prominent for FIB-4 and APRI. NFS never ran into convergence issues.

Of the five methods, GBTM was selected as the best performing method based on both cross-validation and clinical relevance. KML had consistent results between the testing and training sets, with only a one cluster difference for all three liver indicators. However, GBTM's consistent results between the training and testing sets along with the larger portion of the data clustered outweighed the slightly better results observed for KML. Furthermore, while GCKM also had consistent results, it experienced convergence issues for FIB-4 and produced several clusters containing less than 5% of participants. GLMM also possessed significant convergence issues for both sets for APRI and the training set for FIB-4. Anchored k- medoids, which is guaranteed to converge, also required extensive data transformations and exclusion of observations like in KML and possessed the largest discrepancy between the training and testing sets.

From these results, it appears that the methods commonly employed for clustering longitudinal data are not always suitable for typical observational data. KML and anchored k- medoids, two methods that require complete data are unrealistic to use in most observational settings. In our case, hundreds of observations were discarded to prevent even more excessive data imputation due to the unequal follow-up periods among participants. Too much imputation is likely to hinder the true nature of patient's trajectories over time. This was also an issue for GCKM, which required the same data exclusion to obtain clinically plausible results. Unequal follow-up periods are a common feature of observational data, and thus pose a major limitation to the use of KML, GCKM, or anchored K- medoids for such data.

GLMM and GBTM, a parametric and semi-parametric method, did not require complete data. However, GLMM ran into significant convergence issues for APRI and FIB-4 trajectories. This is likely due to the heavily skewed nature of the two variables. Regardless, this suggests that the GLMM method is sensitive to violations in the normality assumption. This is further backed by NFS experiencing no convergence issues while also being normally distributed. GBTM allowed for the gamma distribution to be used in lieu of the gaussian for FIB-4 and APRI and thus likely avoided convergence issues. This suggests that researchers seeking to cluster data that is not normally distributed should avoid using the GLMM method and instead look to GBTM.

One interesting result observed was that for four of the five methods employed, the cluster results were consistent between the training and testing sets. Aside from anchored k- medoids, no clustering method yielded a discrepancy greater than two clusters between the training and testing set. Furthermore, the overall behavior of these clusters' liver score trajectories was mostly consistent between the two sets. This suggests that these methods are robust to smaller sample sizes, as they were able to detect a similar number of clusters. The same cannot be said for the anchored k- medoids method, which saw major differences between the two sets. This is likely due to the extensive data imputation required for this method to function.

An additional result observed was the consistency in the overall behavior of the clusters across the methods. In most cases, the clusters identified for APRI showed a negligible or slight decrease over time. The same was observed for FIB-4 and NFS, except for rather a slight increase in their respective values over time. This suggests two things. One being that the methods were all in agreement in the behavior of the assigned clusters for the three liver indicators. Second being in a clinical context that there may be some external or unmeasured factor that influences the health of patients' livers and the tested scores.

4.2 Results of Clustering APRI, FIB-4, and NFS

Based on comparing these five methods, GBTM was run again but with the combined dataset. The clusters identified for all three liver scores were well partitioned, with little overlap between the trajectories of the clusters. This was done to address our second objective, which was to determine if the clusters identified for the three liver scores would be consistent with each other. We found few demographic or clinical characteristics that varied between clusters per measure. The exception to this is BMI for NFS cluster membership, which showed that higher BMI appeared to correlate to worse NFS cluster membership. The same was observed for INSTI use, albeit less significantly.

Additionally, we found that while the clusters associated with higher APRI slightly corresponded to higher FIB-4, the same could not be said for NFS. Rather, clusters identified by NFS appeared to have little to no bearing on the resulting cluster assignment from APRI or FIB-4. This suggests that our identified clusters per liver measure are uninformative of the subsequent cluster membership for the other liver measures. It also suggests that APRI and FIB-4 share more similarities amongst each other than with NFS.

Interestingly, this discordance appears less prominent when examining the median liver measure trajectory per cluster. Rather, higher cluster membership for each liver measure corresponded to a higher median value for the other liver measures. This isn't as discernable for APRI and NFS, where clusters B-D for NFS had overlapping median APRI trajectories. This is important, given these clusters combined account for most of the participants. The clinical significance of this is less clear, given clusters A and B for all three liver measures comprise most of the observations with fibrosis for their respective liver measure. Thus, the choice of cut-off can greatly impact the interpretation of these results.

These results are consistent with our previous findings. The implications of this are that physicians should be wary in using these three liver scores interchangeably, as they may not lead to the same medical conclusion. While FIB-4 and APRI showed a strong relationship with each other, the clusters identified by NFS were slightly different than the ones identified by FIB-4 and APRI.

4.3 Fibroscan Clustering Results

To further explore this, we proceeded to cluster a cross-sectional Fibroscan visit. This was so that we could explore our third question, which was whether the clusters identified on LS and

CAP were also associated with higher cluster membership for the three liver measures. A two-cluster solution was identified, with a roughly 60-40 split into cluster A and B, respectively. Cluster B was found to be representative of patients with higher values of CAP and LS. It was also shown to appear to be related to being on INSTIs. Additionally, the partitioning done still possessed some overlap between LS and CAP fibrosis outcomes. Furthermore, these identified clusters showed little alignment with the clusters identified for all three liver measures.

The implications of this suggest that INSTI use may be associated with slightly worse liver outcomes. These results also imply that none of the three liver measures explored are informative. Thus, clinicians should utilize more than just a liver score or measurement when examining their patient's liver health.

4.4 Limitations

There are limitations in this study. The first being the fact that the data used in this study is observational, hampering our ability to ascertain causality. Because of the longitudinal observational nature of the data, there is also extensive loss to follow-up observed in participants at around the fifth visit. Loss of follow-up limits our ability to learn the behavior of patients' livers over time. These limitations are prevalent both in general and for clustering.

As for clustering, there are several well-known limitations. The first is that clustering is generally an exploratory procedure. Thus, the results obtained should be interpreted carefully, as they cannot be verified statistically. Because the true nature of liver trajectories in our population is unknown, assessing the correct number of clusters or shape is impossible. Thus, the method that we find performs the best in this context may not be the one that most accurately reflects the population. Additionally, ascertaining the number of clusters is a well-known issue in clustering. For our purposes we used a more subjective use of the elbow method to ensure clinical interpretability. This may come at the expense of potentially less accurate results.

4.5 Key Takeaways

Based on this study, several recommendations can be made. First, more methodological research into longitudinal clustering algorithms that are more robust to the issues common to observational data should be conducted. That way, researchers are not limited to methods that are prone to not converging. A second recommendation is that in the clinical context of this study, clinicians should be wary using APRI, FIB-4, and NFS interchangeably in practice for their patients. This also leads to the final recommendation, which is that more research into the characteristics and variables associated with liver health should be investigated. That way, clinicians are not reliant on metrics that may be inaccurate and thus can ensure the health of their patients.

References

Adepeju M, Langton S, Bannister J (2021). _akmedoids: Anchored Kmedoids for Longitudinal Data Clustering_. R package version 1.3.0, .

Den Teuling N (2022). latrend: A Framework for Clustering Longitudinal Data_. R package version 1.5.0, https://CRAN.R-project.org/package=latrend

Adepeju, M., Langton, S., & Bannister, J. (2021). Anchored K-medoids: A novel adaptation of K-medoids further refined to measure long-term instability in the exposure to crime. *Journal of Computational Social Science, 4*(2), 655–680. https://doi.org/10.1007/s42001-021-00103-1

Afdhal NH. Fibroscan (transient elastography) for the measurement of liver fibrosis. Gastroenterol Hepatol (N Y). 2012 Sep;8(9):605-7. PMID: 23483859; PMCID: PMC3594956.

Alashwal, H., El Halaby, M., Crouse, J. J., Abdalla, A., & Moustafa, A. A. (2019). The application of unsupervised clustering methods to alzheimer's disease. Frontiers in Computational Neuroscience, 13. https://doi.org/10.3389/fncom.2019.00031

Amernia, B., Moosavy, S. H., Banookh, F., & Zoghi, G. (2021). FIB-4, Apri, and AST/ALT ratio compared to fibroscan for the assessment of hepatic fibrosis in patients with non-alcoholic fatty liver disease in Bandar Abbas, Iran. BMC Gastroenterology, 21(1). https://doi.org/10.1186/s12876-021-02038-3

Angulo, P., Hui, J.M., Marchesini, G., Bugianesi, E., George, J., Farrell, G.C., Enders, F., Saksena, S., Burt, A.D., Bida, J.P., Lindor, K., Sanderson, S.O., Lenzi, M., Adams, L.A., Kench, J., Therneau, T.M. and Day, C.P. (2007), The NAFLD fibrosis score: A noninvasive system that identifies liver fibrosis in patients with NAFLD. Hepatology, 45: 846-854. https://doi.org/10.1002/hep.21496

Bacon, M. C., von Wyl, V., Alden, C., Sharp, G., Robison, E., Hessol, N., Gange, S., Barranday, Y., Holman, S., Weber, K., & Young, M. A. (2005). The women's interagency HIV study: An observational cohort brings clinical sciences to the bench. *Clinical and Vaccine Immunology, 12*(9), 1013–1019. https://doi.org/10.1128/cdli.12.9.1013-1019.2005

Barkan, S. E., Melnick, S. L., Preston-Martin, S., Weber, K., Kalish, L. A., Miotti, P., Young, M., Greenblatt, R., Sacks, H., & Feldman, J. (1998). The Women's Interagency HIV Study. WIHS Collaborative Study Group. Epidemiology (Cambridge, Mass.), 9(2), 117–125.

Den Teuling, N. G., Pauws, S. C., & van den Heuvel, E. R. (2021). A comparison of methods for clustering longitudinal data with slowly changing trends. *Communications in Statistics - Simulation and Computation*, *52*(3), 621–648. https://doi.org/10.1080/03610918.2020.1861464

Genolini, C., & Falissard, B. (2011). KML: A package to Cluster Longitudinal Data. *Computer Methods and Programs in Biomedicine*, *104*(3). https://doi.org/10.1016/j.cmpb.2011.05.008

Huang, H., Li, Y., & Guan, Y. (2014). Joint modeling and clustering paired generalized longitudinal trajectories with application to cocaine abuse treatment data. *Journal of the*

American Statistical Association, *109*(508), 1412–1424. https://doi.org/10.1080/01621459.2014.957286

Kassambara A, Mundt F (2020). _factoextra: Extract and Visualize the Results of Multivariate Data Analyses_. R package version 1.0.7, .

Kerchberger, A. M., Sheth, A. N., Angert, C. D., Mehta, C. C., Summers, N. A., Ofotokun, I., Gustafson, D., Weiser, S. D., Sharma, A., Adimora, A. A., French, A. L., Augenbraun, M., Cocohoba, J., Kassaye, S., Bolivar, H., Govindarajulu, U., Konkle-Parker, D., Golub, E. T., & Lahiri, C. D. (2019). Weight gain associated with integrase stand transfer inhibitor use in women. Clinical Infectious Diseases, 71(3), 593–600. https://doi.org/10.1093/cid/ciz853

Kirkegaard-Klitbo, D. M., Thomsen, M. T., Gelpi, M., Bendtsen, F., Nielsen, S. D., & Benfield, T. (2021). Hepatic steatosis associated with exposure to elvitegravir and raltegravir. Clinical Infectious Diseases, 73(3). https://doi.org/10.1093/cid/ciab057

Lahiri, C. D., Yu, M. A., Gerig, L. G., Mehta, C. C., Musonge-Effoe, J., Price, J. C., Tien, P. C., Spencer, A. B., Albrecht, S., Alcaide, M. L., Adimora, A. A., French, A. L., Augenbraun, M. H., Anastos, K., & Alvarez, J. A. (2023). LIVER STEATOSIS AND FIBROSIS IN WOMEN WITH HIV BY INTEGRASE INHIBITOR USE. *Conference on Retroviruses and Opportunistic Infections*. https://doi.org/https://www.natap.org/2023/CROI/croi_20.html

Lee, J., Vali, Y., Boursier, J., Spijker, R., Anstee, Q. M., Bossuyt, P. M., & Zafarmand, M. H. (2021). Prognostic accuracy of FIB-4, NAFLD fibrosis score and Apri for nafld-related events: A systematic review. Liver International, 41(2), 261–270. https://doi.org/10.1111/liv.14669

Maurice, J. B., Patel, A., Scott, A. J., Patel, K., Thursz, M., & Lemoine, M. (2017). Prevalence and risk factors of nonalcoholic fatty liver disease in HIV-monoinfection. AIDS, 31(11), 1621–1632. https://doi.org/10.1097/qad.0000000000001504

Nagin, D. S., & Odgers, C. L. (2010). Group-based trajectory modeling in clinical research. *Annual Review of Clinical Psychology*, *6*(1), 109–138. https://doi.org/10.1146/annurev.clinpsy.121208.131413

Pan, L., Li, Y., He, K., Li, Y., & Li, Y. (2020). Generalized linear mixed models with gaussian mixture random effects: Inference and Application. *Journal of Multivariate Analysis*, *175*, 104555. https://doi.org/10.1016/j.jmva.2019.104555

Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2005). Selection of k in k-means clustering. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 219(1), 103–119. https://doi.org/10.1243/095440605x8298

Posit team (2023). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA. URL http://www.posit.com

Price, J. C., Ma, Y., Kuniholm, M. H., Adimora, A. A., Fischl, M., French, A. L., Golub, E. T., Konkle-Parker, D., Minkoff, H., Ofotokun, I., Plankey, M., Sharma, A., & Tien, P. C. (2022). Human immunodeficiency virus is associated with elevated fibroscan–aspartate aminotransferase (FAST) score. *Clinical Infectious Diseases*, *75*(12), 2119–2127. https://doi.org/10.1093/cid/ciac337

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sax, P. E., Erlandson, K. M., Lake, J. E., Mccomsey, G. A., Orkin, C., Esser, S., Brown, T. T., Rockstroh, J. K., Wei, X., Carter, C. C., Zhong, L., Brainard, D. M., Melbourne, K., Das, M., Stellbrink, H.-J., Post, F. A., Waters, L., & Koethe, J. R. (2019). Weight gain following initiation of antiretroviral therapy: Risk factors in randomized comparative clinical trials. Clinical Infectious Diseases, 71(6), 1379–1389. https://doi.org/10.1093/cid/ciz999

Scarsi, K. K., Havens, J. P., Podany, A. T., Avedissian, S. N., & Fletcher, C. V. (2020). HIV-1 integrase inhibitors: A Comparative Review of Efficacy and Safety. Drugs, 80(16), 1649–1676. https://doi.org/10.1007/s40265-020-01379-9

Smith, S. J., Zhao, X. Z., Passos, D. O., Lyumkis, D., Burke, T. R., & Hughes, S. H. (2021). Integrase strand transfer inhibitors are effective Anti-HIV Drugs. Viruses, 13(2), 205. https://doi.org/10.3390/v13020205

Teuling, N. D., Pauws, S., & Heuvel, E. van den. (2021, November 10). *Clustering of longitudinal data: A tutorial on a variety of approaches*. arXiv.org. Retrieved March 27, 2023, from https://arxiv.org/abs/2111.05469

Twisk, J., & Hoekstra, T. (2012). Classifying developmental trajectories over time should be done with great caution: A comparison between methods. *Journal of Clinical Epidemiology*, *65*(10), 1078–1087. https://doi.org/10.1016/j.jclinepi.2012.04.010

U.S. Department of Health and Human Services. (n.d.). *HIV/AIDS*. National Institute of Allergy and Infectious Diseases. Retrieved March 27, 2023, from https://www.niaid.nih.gov/diseases-conditions/hivaids

van der Loo M (2022). _simputation: Simple Imputation_. R package version 0.2.8, https://CRAN.R project.org/package=simputation.

Wijarnpreecha K, Aby ES, Ahmed A, Kim D. The association of weight gain with nonalcoholic fatty liver disease and fibrosis detected by FibroScan in the United States. Ann Gastroenterol. 2022 Mar-Apr;35(2):194-202. doi: 10.20524/aog.2022.0687. Epub 2022 Feb 15. PMID: 35479585; PMCID: PMC8922259.

Yu, M. A., Gerig, L., Mehta, C. C., Yusef, O., Musonge-Effoe, J., Alvarez, J., Spence, A. B., Albrecht, S., Alcaide, M. L., Adimora, A. A., Abraham, A. G., French, A. L., Augenbraun, M., Anastos, K., Price, J. C., Tien, P. C., & Lahiri, C. D. (2022). 1276. noninvasive assessment of change in hepatic fibrosis following initiation of integrase inhibitors in women living with HIV. *Open Forum Infectious Diseases*, *9*(Supplement_2). https://doi.org/10.1093/ofid/ofac492.1107

Zhao, A. V., Crutchley, R. D., Guduru, R. C., Ton, K., Lam, T., & Min, A. C. (2022). A clinical review of HIV integrase strand transfer inhibitors (instis) for the prevention and treatment of HIV-1 infection. *Retrovirology*, *19*(1). https://doi.org/10.1186/s12977-022-00608-1

Appendix A.

A.1 KML

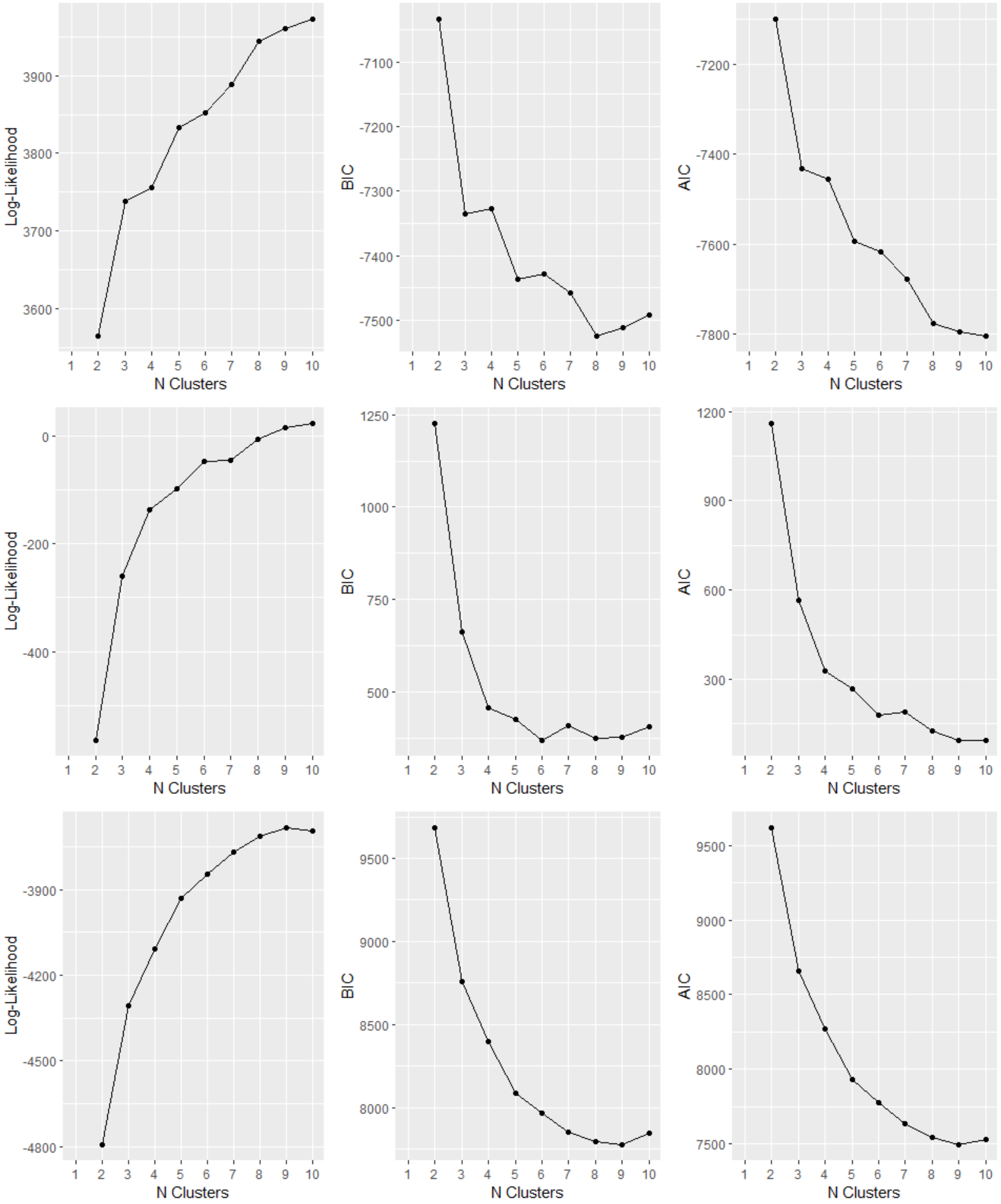


Figure A.1: Quality Criteria for Trimmed APRI, Trimmed FIB-4, and NFS Longitudinal K-Means Clustering Respectively (Training Set)

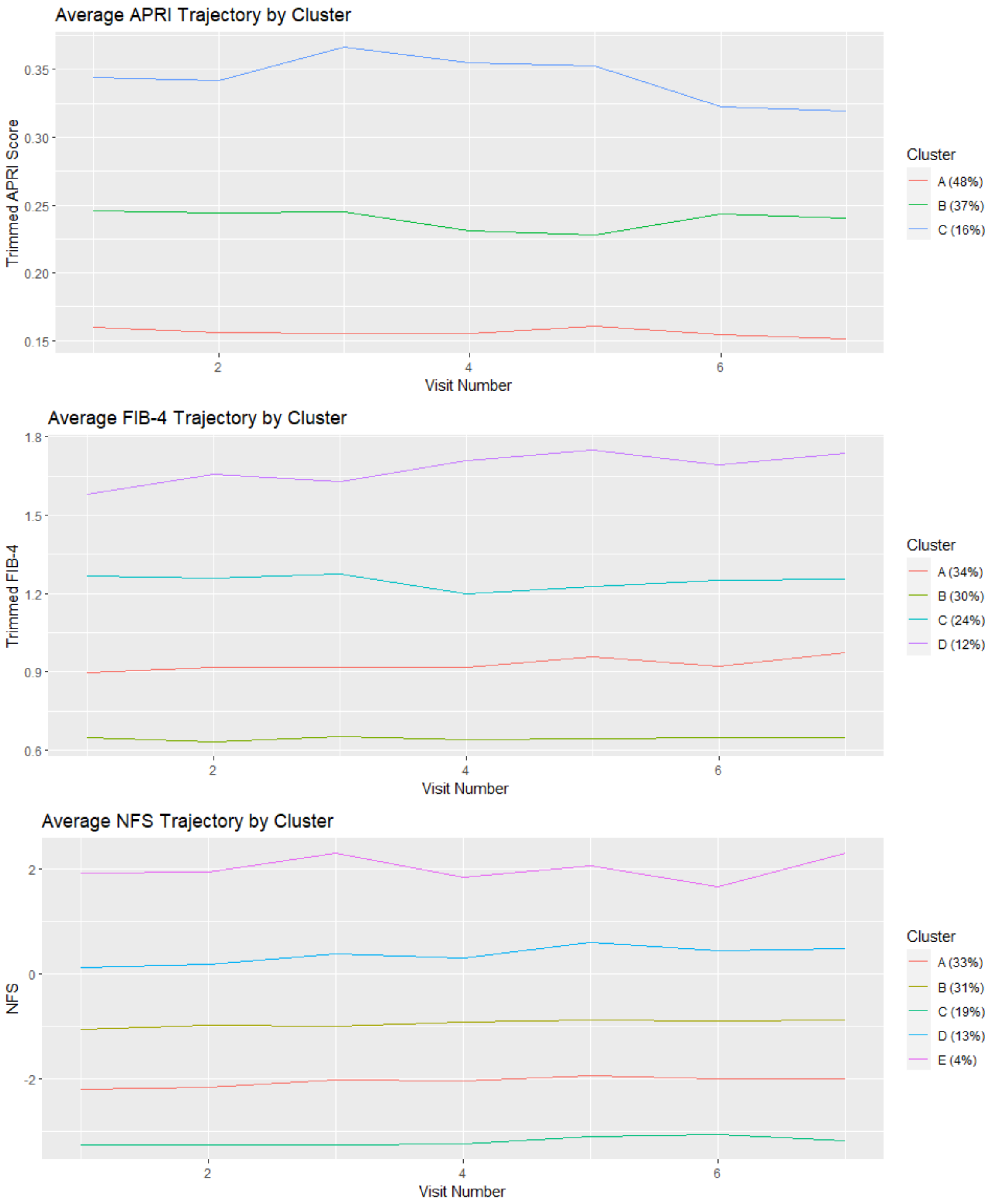


Figure A.2: Average APRI, FIB-4, and NFS Trajectories by Cluster over Visits (Training Set)

Based on the plotted quality criteria in Figure A.1 and the corresponding elbow method, an optimal number of clusters chosen for APRI, FIB-4, and NFS were three, four, and five respectively. Participants' trimmed APRI trajectories were partitioned into clusters containing 48, 37, and 16% of participants (Figure A.2). FIB-4 trajectories were partitioned into clusters containing 30, 34, 24, and 12% of participants. NFS trajectories were partitioned into clusters containing 33, 31, 19, 13, and 4% of participants respectively. Only cluster E for NFS contained < 5% of the overall participants (Figure A.2).

The average trimmed APRI trajectories do not cross or stray far from their value at baseline. Clusters D and E for FIB-4 and NFS show slight increases over the course of the seven visits (Figure A.2).

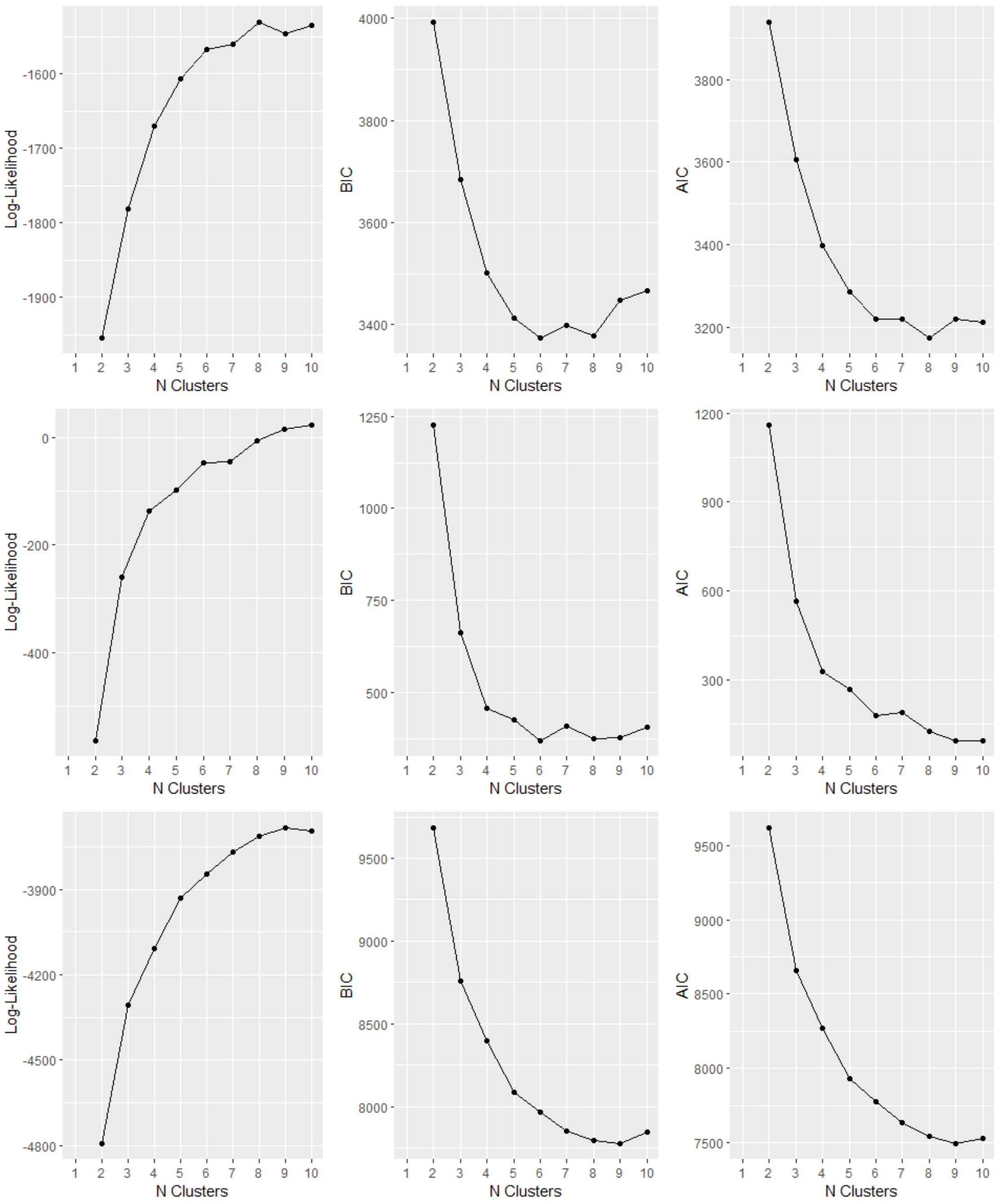


Figure A.3: Quality Criteria for Trimmed APRI, Trimmed FIB-4, and NFS Longitudinal K-Means Clustering Respectively (Testing Set)

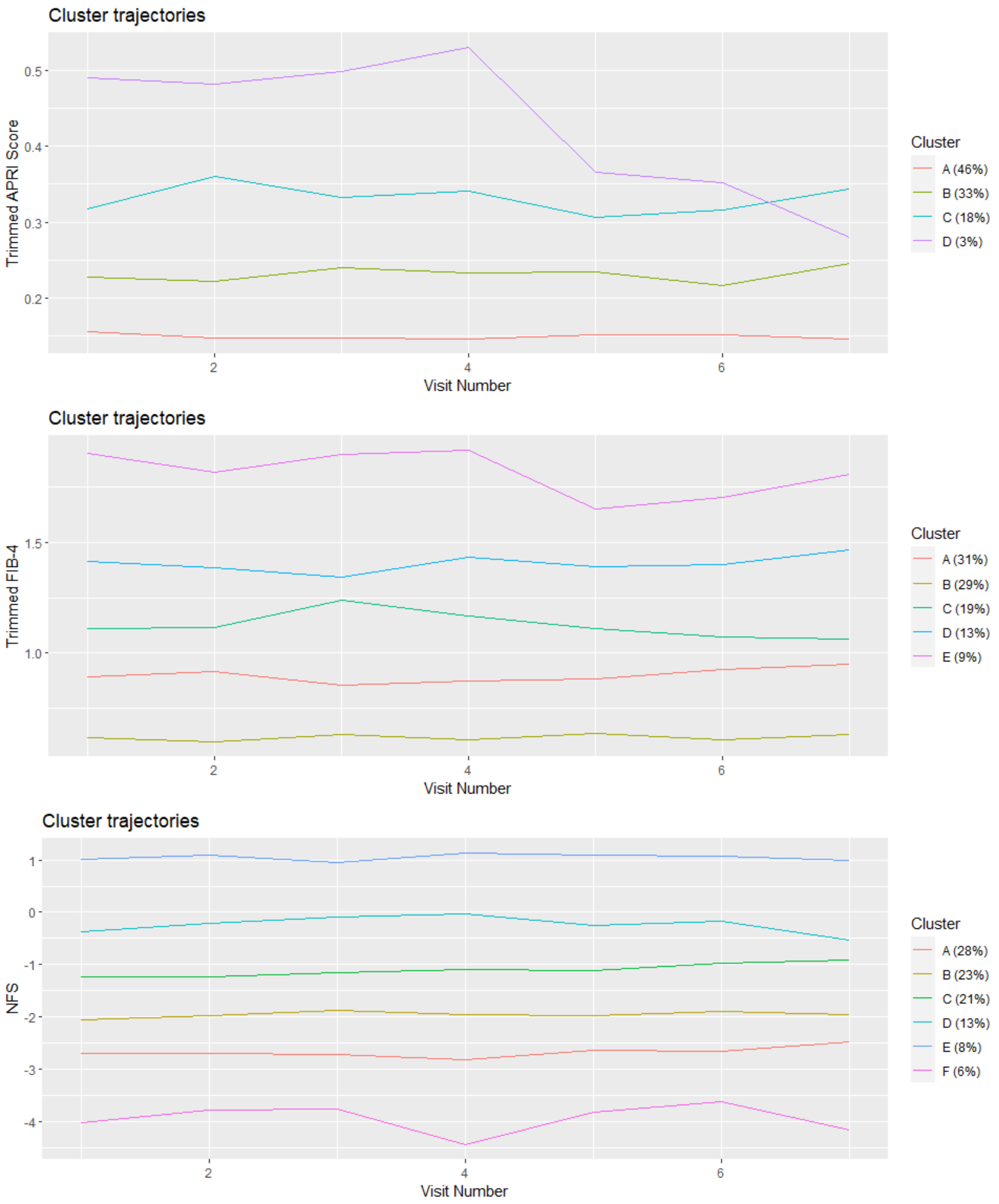


Figure A.4: Average APRI, FIB-4, and NFS Trajectories by Cluster over Visits (Testing Set)

Based on the plotted quality criteria in Figure A.3 and the corresponding elbow method, an optimal number of clusters chosen for APRI, FIB-4, and NFS were four, five, and six respectively. Participants trimmed APRI trajectories were partitioned into clusters containing 46, 33, 18, and 3% of participants. FIB-4 trajectories were partitioned into clusters containing

31, 29, 19, 13, and 9% of participants. NFS trajectories were partitioned into clusters containing 28, 23, 21, 13, 8, and 6% of participants. Only cluster D for trimmed APRI contained < 5% of participants (Figure A.3).

The average trimmed APRI score for cluster D crosses cluster C at about the 7th visit. Otherwise, the trajectories show little trend over time. Average trajectories for trimmed FIB-4 and NFS show a negligible trend over the course of the seven visits. These trends are consistent with those observed in the training set (Figure A.4).

A.2 GCKM

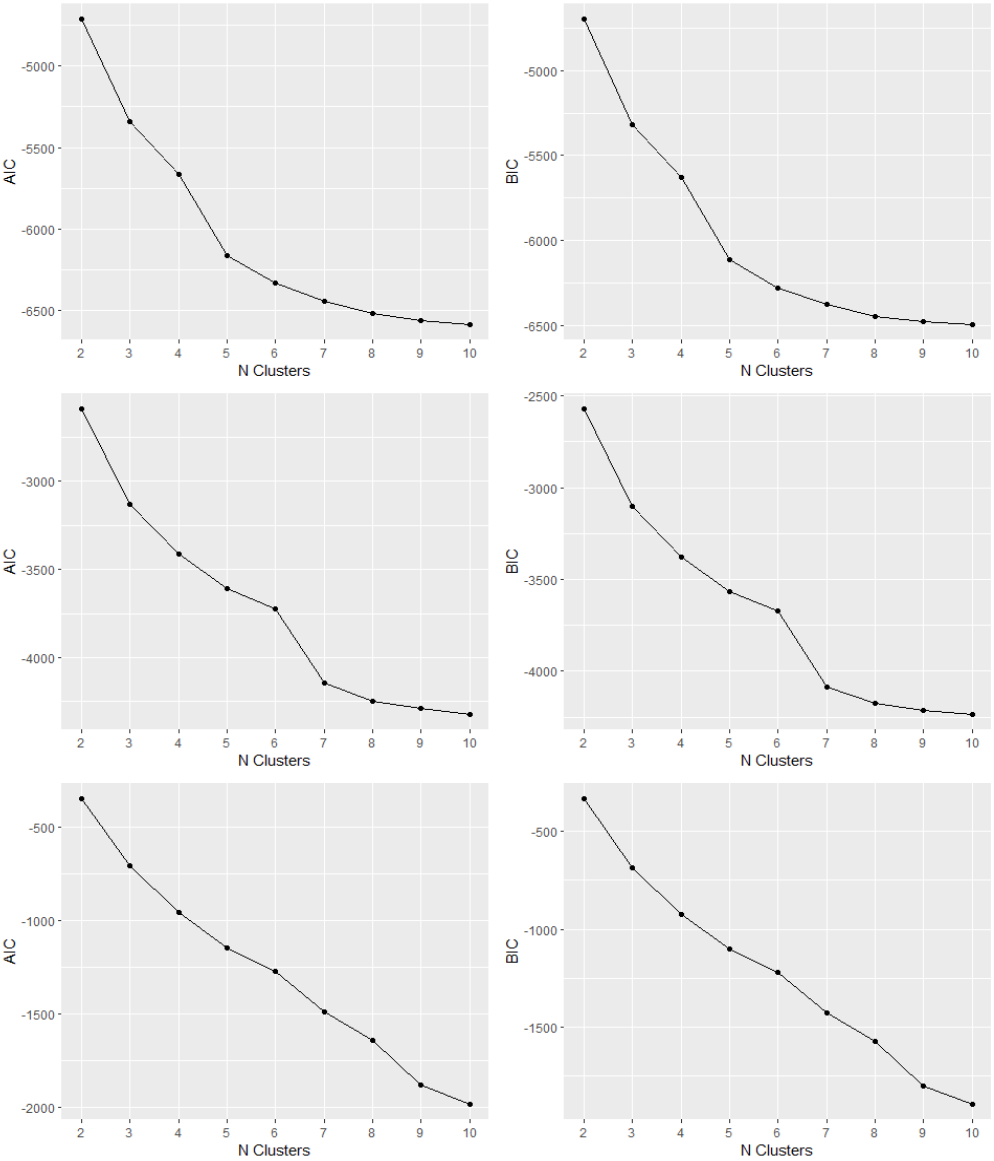


Figure A.5: Quality Criteria for APRI, FIB-4, and NFS Growth Curve K-means Clustering Respectively (Training Set)

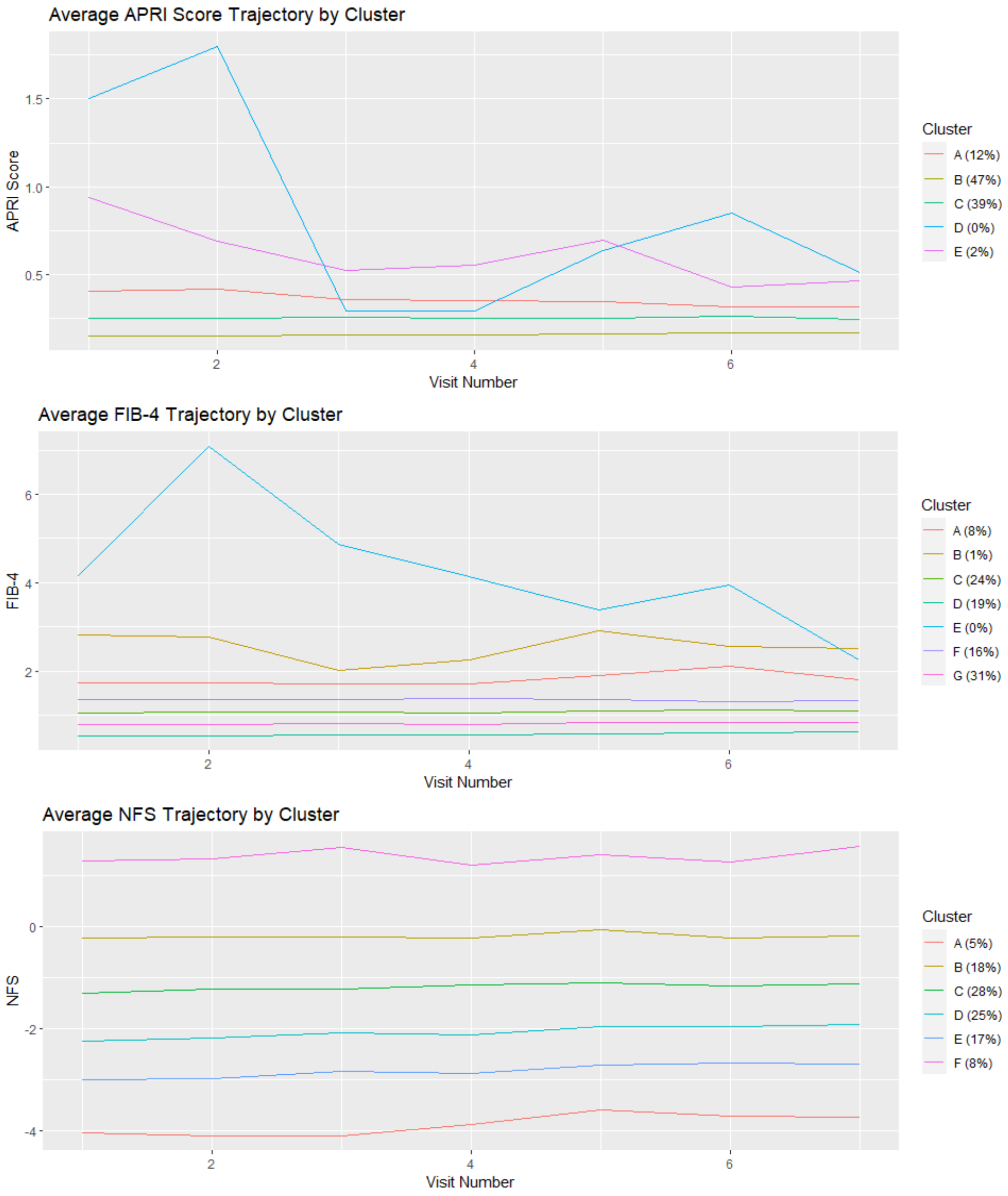


Figure A.6: Average APRI, FIB-4, and NFS Trajectories by Cluster over Visits (Training Set)

Convergence failure was observed for all FIB-4 cluster models. Convergence issues were not observed for any of the APRI or NFS models.

Based on the plotted quality criteria in Figure A.5 and the corresponding elbow method, an optimal number of clusters chosen for APRI, FIB-4, and NFS were five, seven, and six respectively. Participants APRI trajectories were partitioned into clusters containing 12, 47, 39, < 1%, and 2% of participants. FIB-4 trajectories were partitioned into clusters containing 8, 1, 24, 19, < 1%, 16, and 31% of participants. NFS trajectories were partitioned into clusters containing 5, 18, 28, 25, 17, and 8% of participants. Clusters D and E for APRI and B and E for FIB-4 contained < 5% of participants (Figure A.6).

The average APRI, FIB-4, and NFS show little trend or change for clusters that contain > 5% of participants. (Figure A.6).

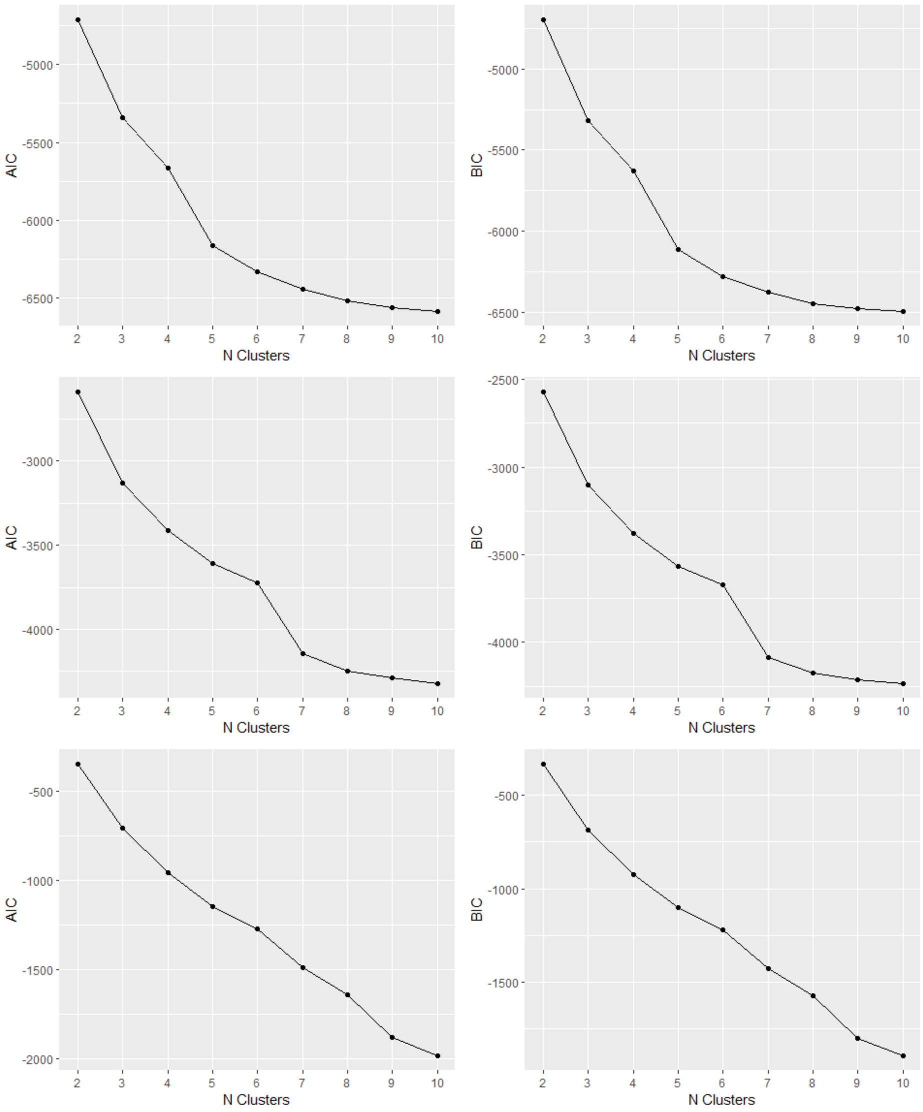


Figure A.7: Quality Criteria for APRI, FIB-4, and NFS Growth Curve K-means Clustering Respectively (Testing Set)

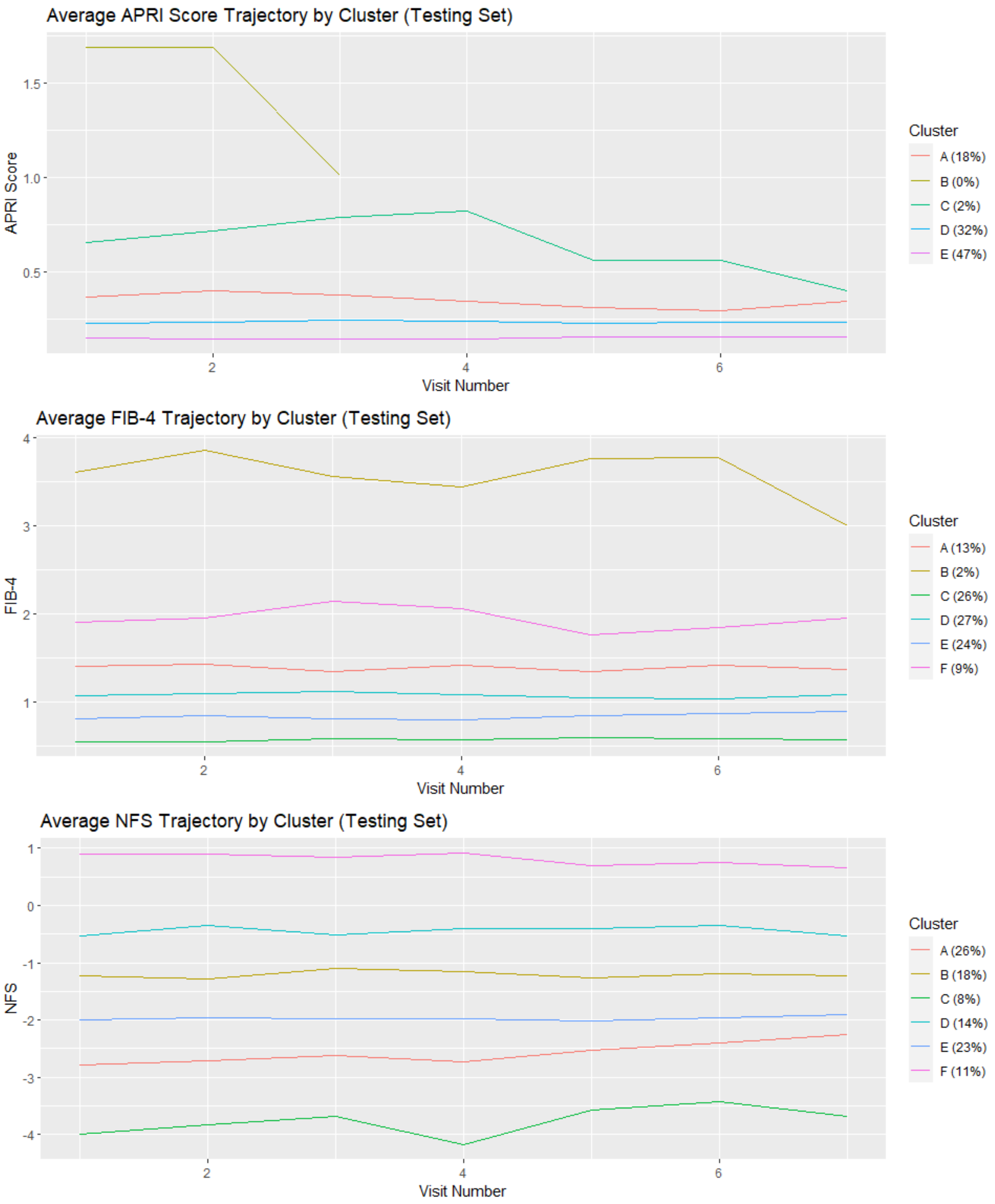


Figure A.8: Average APRI, FIB-4, and NFS Trajectories by Cluster over Visits (Training Set)

Based on the plotted quality criteria in Figure A.7 and the corresponding elbow method, an optimal number of clusters chosen for APRI, FIB-4, and NFS were five, six, and six respectively. Participants APRI trajectories were partitioned into clusters containing 18, < 1, 2, 32, and 47% of participants respectively. FIB-4 trajectories were partitioned into clusters containing 13, 2, 26, 27, 24, and 9% of participants respectively. NFS trajectories were partitioned into clusters containing 26, 18, 8, 14, 23, and 11% of participants respectively. Clusters B and C for APRI and B for FIB-4 contained < 5% of participants (Figure A.8).

The average APRI, FIB-4, and NFS show little trend or change for clusters that contain > 5% of participants. (Figure A.8). Cluster B for APRI contained < 1% of trajectories with follow-up time less than three visits. The trends observed between the two sets are consistent as they both indicate little change in the liver indicators over time (Figures A.6 and A.8).

A.3 GBTM

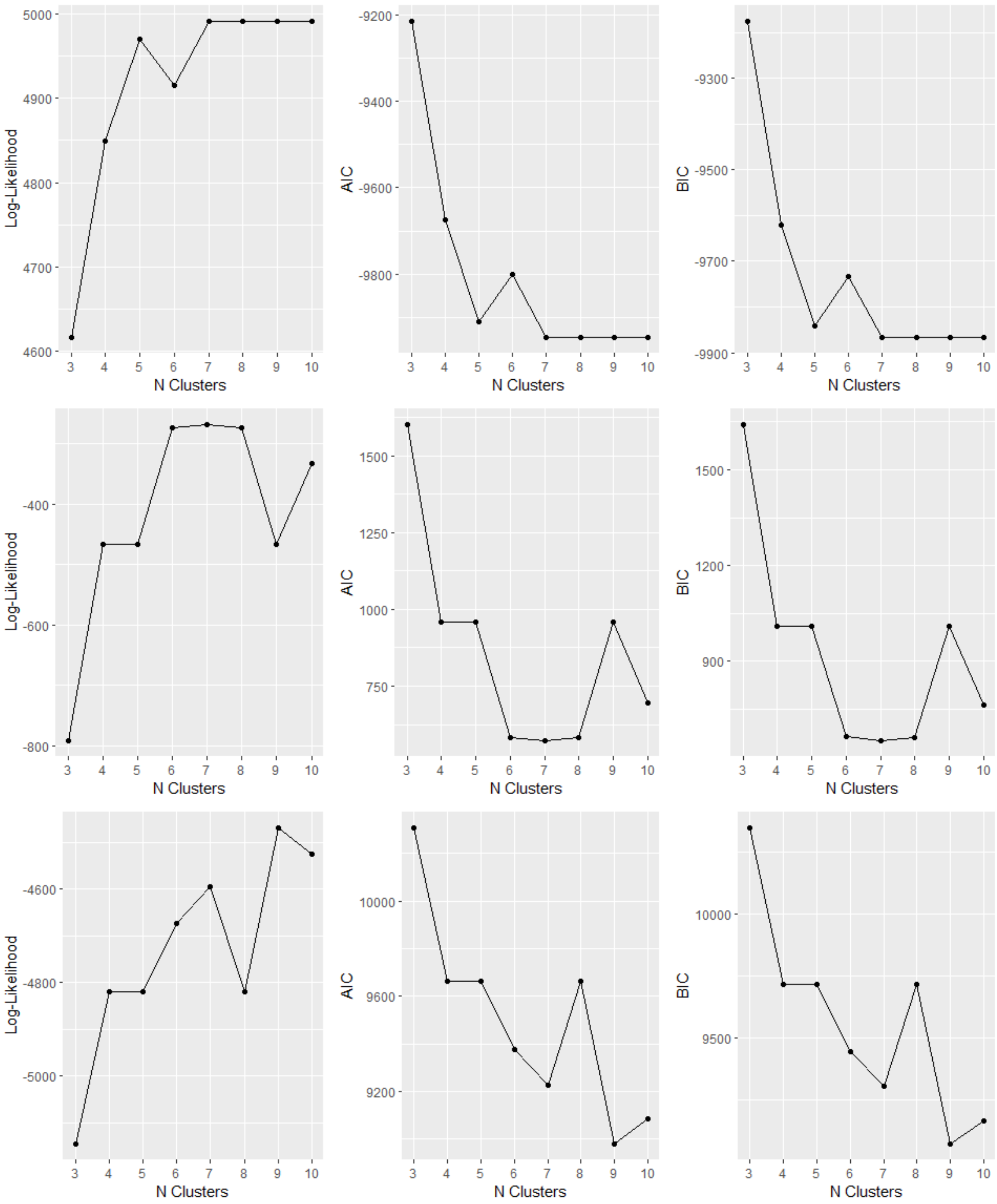


Figure A.9: Quality Criteria for APRI, FIB-4, and NFS Group Based Trajectory Modeling Clustering Respectively (Training Set)

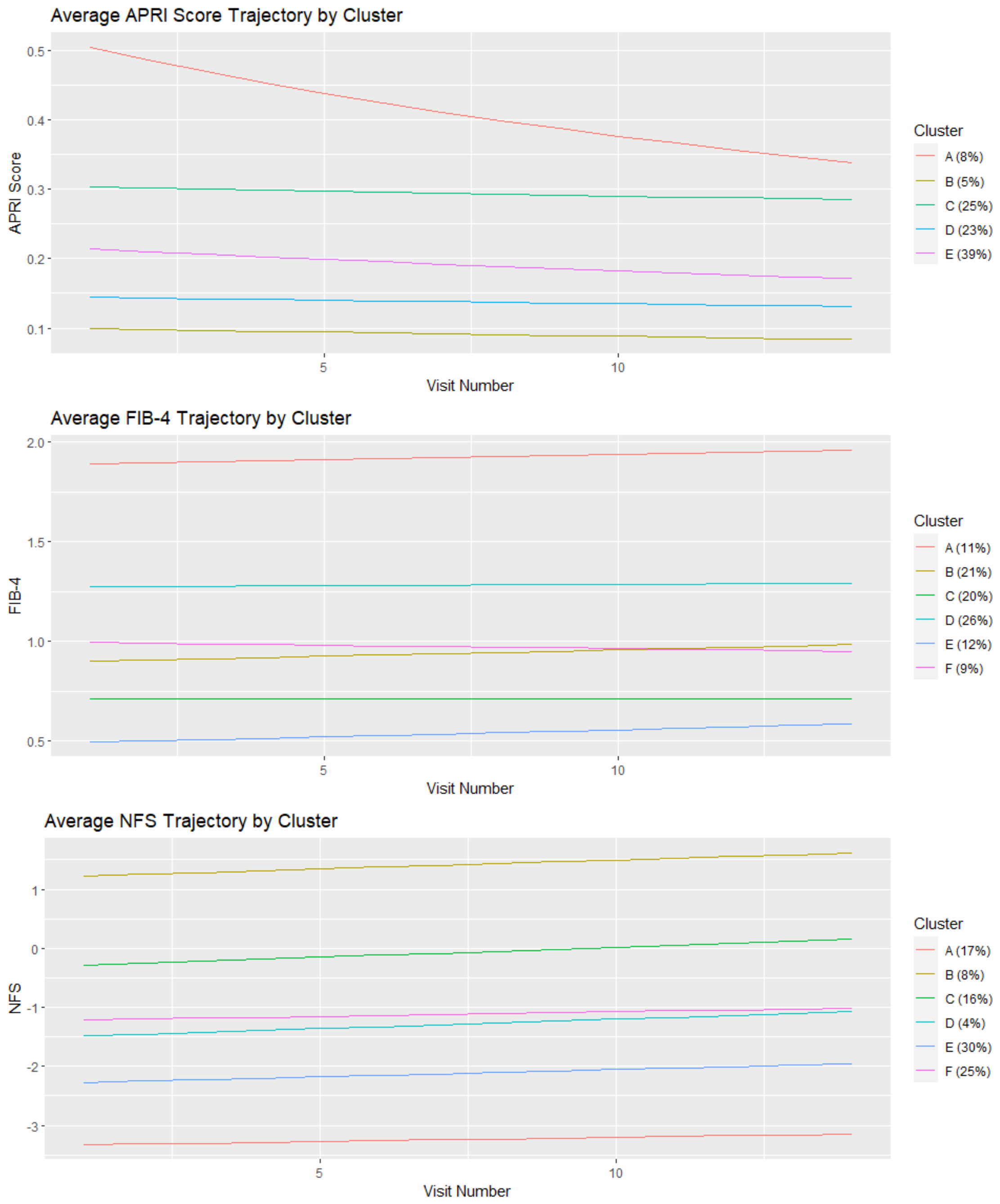


Figure A.10: Average APRI, FIB-4, and NFS Trajectories by Cluster over Visits (Training Set)

Based on the plotted quality criteria in Figure A.9 and the corresponding elbow method, an optimal number of clusters chosen for APRI, FIB-4, and NFS were five, six, and six respectively. Participants APRI trajectories were partitioned into clusters containing 8, 5, 25, 23, and 39% of participants. FIB-4 trajectories were partitioned into clusters containing 11, 21, 20, 26, 12, and 9% of participants. NFS trajectories were partitioned into clusters containing 17, 8, 16, 4, 30, and 25% of participants. Only cluster D for NFS contained < 5% of participants (Figure A.10).

The average APRI for participants in cluster A appears to decrease over the course of the visits. Other clusters show a slight decrease in their average APRI score, but none cross each other. The average FIB-4 score in cluster A shows an upward trend, with the other clusters showing a stagnant or slightly decreasing trend. The average NFS score among all clusters shows a slight upward trend over time (Figure A.10).

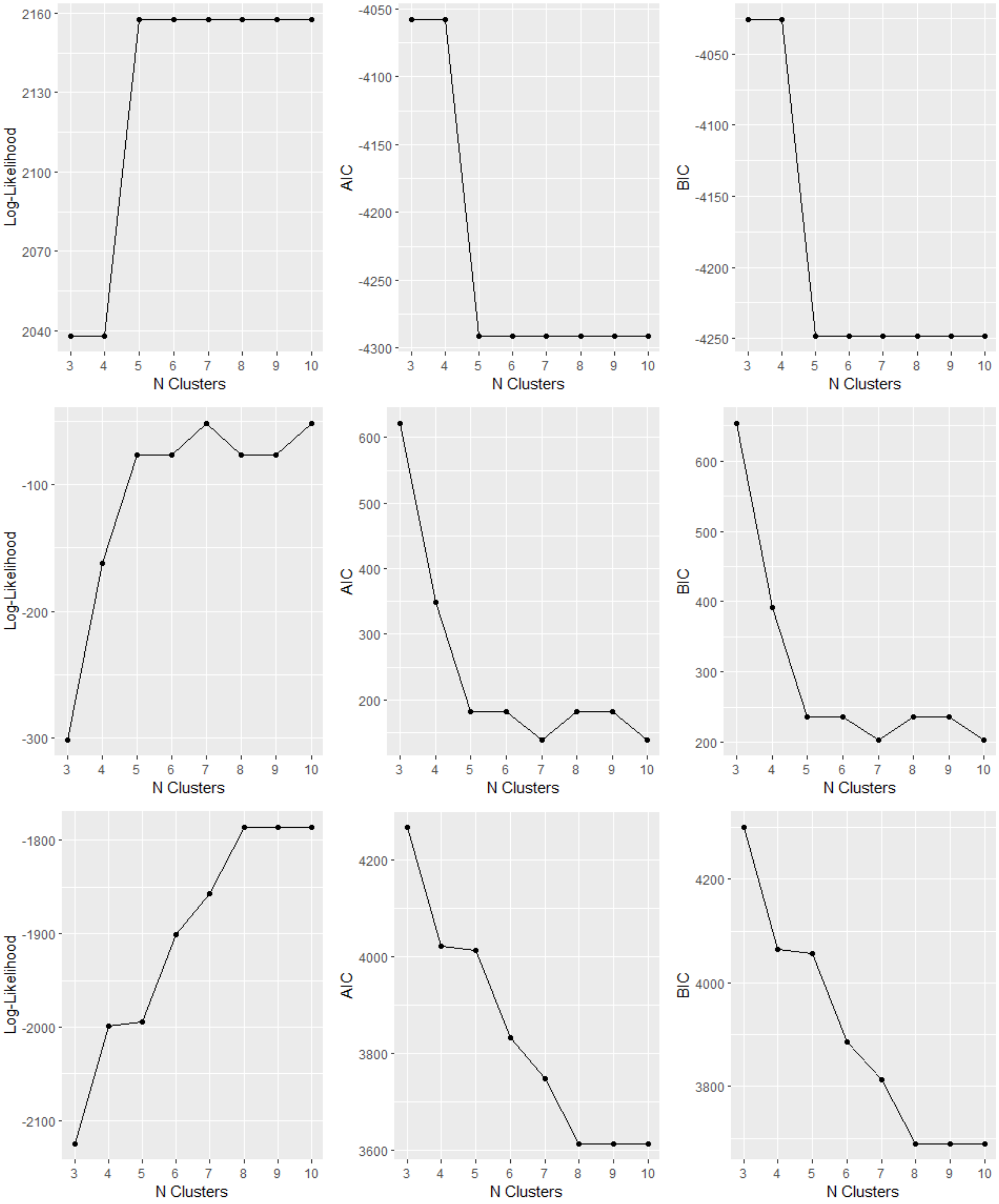


Figure A.11: Quality Criteria for APRI, FIB-4, and NFS Group Based Trajectory Modeling Clustering Respectively (Testing Set)

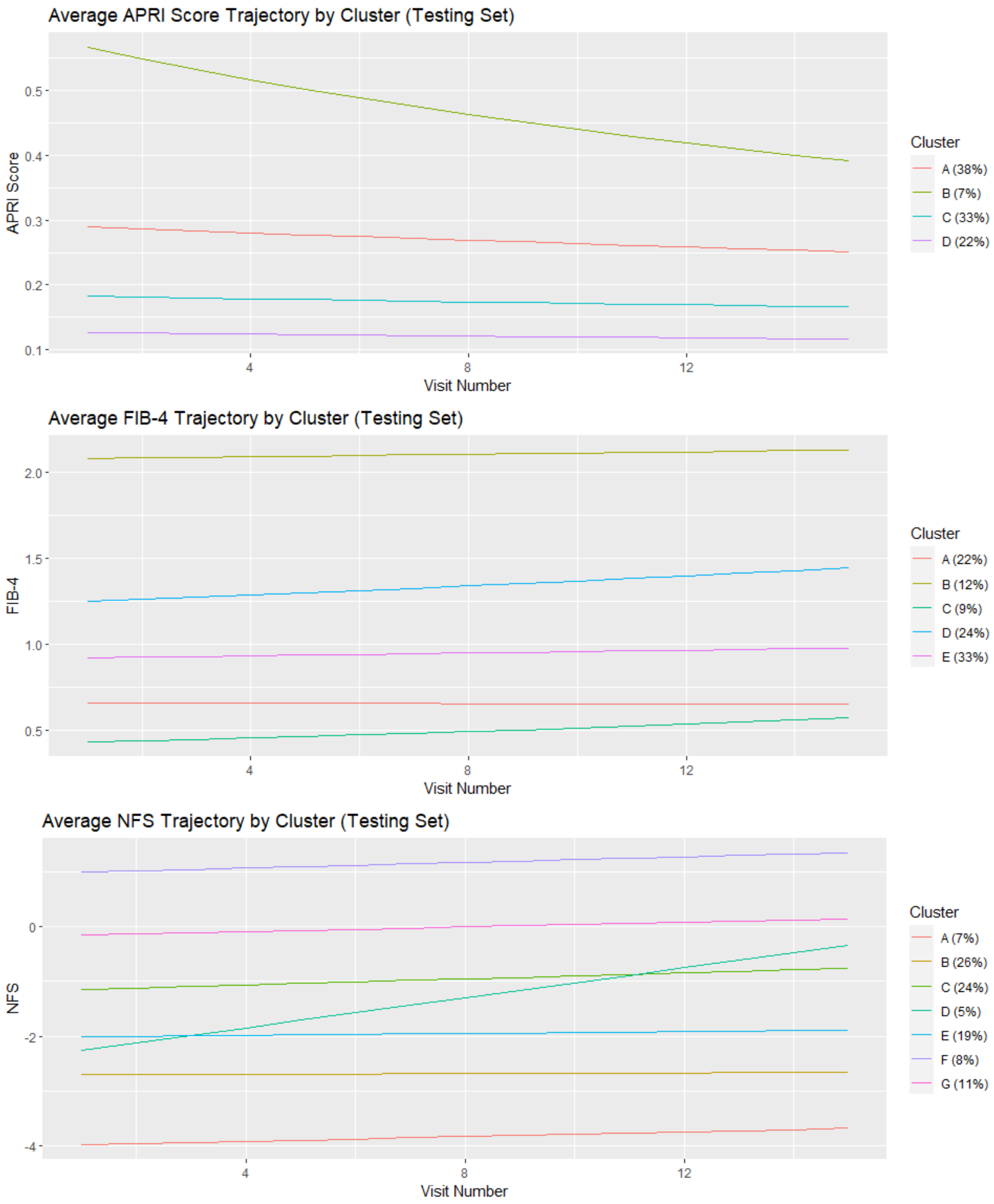


Figure A.12: Average APRI, FIB-4, and NFS Trajectories by Cluster over Visits (Testing Set)

Based on the plotted quality criteria in Figure A.11 and the corresponding elbow method, an optimal number of clusters chosen for APRI, FIB-4, and NFS were five, five, and eight respectively. Participants APRI trajectories were partitioned into clusters containing 38, 7, 33, and 22% of participants. FIB-4 trajectories were partitioned into clusters containing 22, 12, 9, 24, and 33% of participants. NFS trajectories were partitioned into clusters containing 7, 26, 24, 5, 19, 8, and 11% of participants. Both APRI and NFS yielded an empty cluster that contained none of the participants for each (Figure A.12)

The mean trends observed for the clusters in the training set are consistent with those of the testing set, with APRI’s clusters showing a slight downward trend and FIB-4 and NFS showing a slight upward trend (Figures A.10 and A.12).

A.4 GLMM

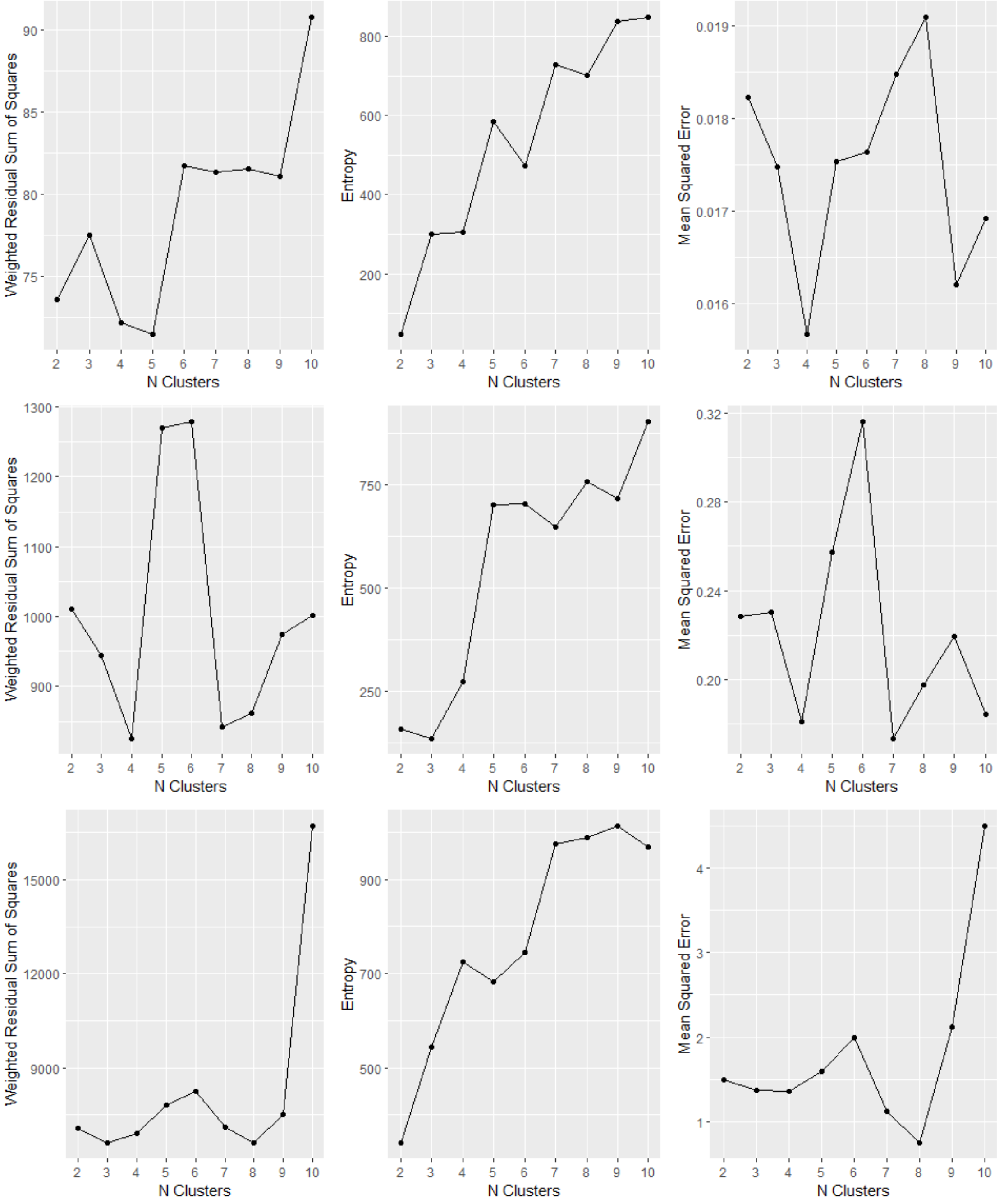


Figure A.13: Quality Criteria for APRI, FIB-4, and NFS Generalized Linear Mixed Model Clustering (Training Set)

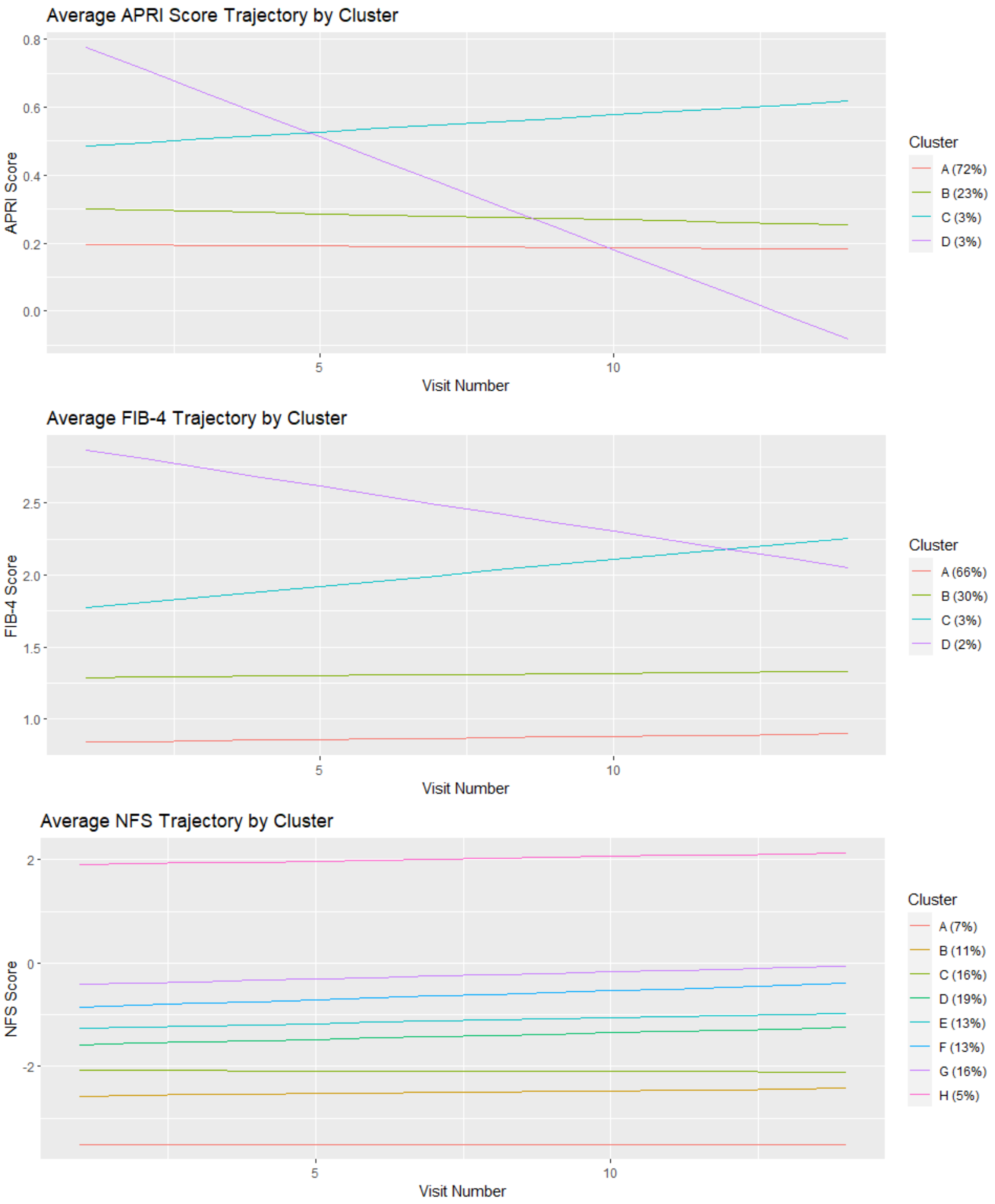


Figure A.14: Average APRI, FIB-4, and NFS Trajectories by Cluster over Visits (Training Set)

Convergence failure was observed for all cluster sizes tested for APRI and FIB-4 scores in the training set. The APRI models with the testing set also failed to converge for all cluster sizes tested. Convergence was observed for the FIB-4 testing set and both NFS sets.

Based on the plotted quality criteria in Figure A.13 and the corresponding elbow method, an optimal number of clusters chosen for APRI, FIB-4, and NFS were four, four, and eight respectively. Participants APRI trajectories were partitioned into clusters containing 72, 23, 3, and 3% of participants. FIB-4 trajectories were partitioned into clusters containing 66, 30, 3, and 2% of participants. NFS trajectories were partitioned into clusters containing 7, 11, 16, 19, 13, 13, 16, and 5% of participants. Clusters C and D for FIB-4 and APRI contained < 5% of participants (Figure A.14).

The average APRI for participants in clusters A and B, which compose around 95% of all participants, show a slight decrease over the course of the follow-up period. Clusters A and B for FIB-4 show the opposite, with their combined clusters containing about 95% of participants with a slight upward trend. All eight clusters for NFS show either no trend or a slight increase over time (Figure A.14).

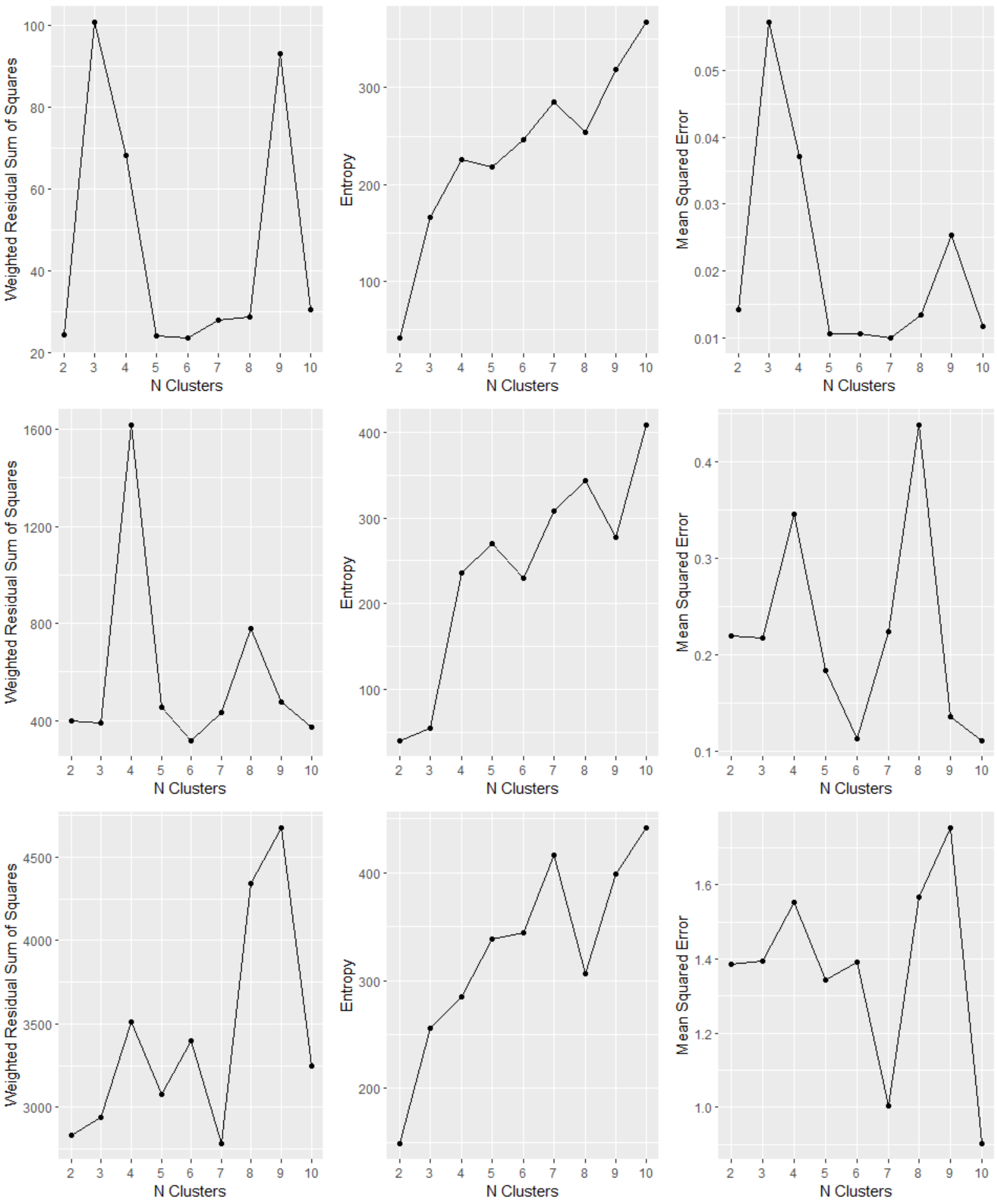


Figure A.15: Quality Criteria for APRI, FIB-4, and NFS Generalized Linear Mixed Model Clustering (Testing Set)

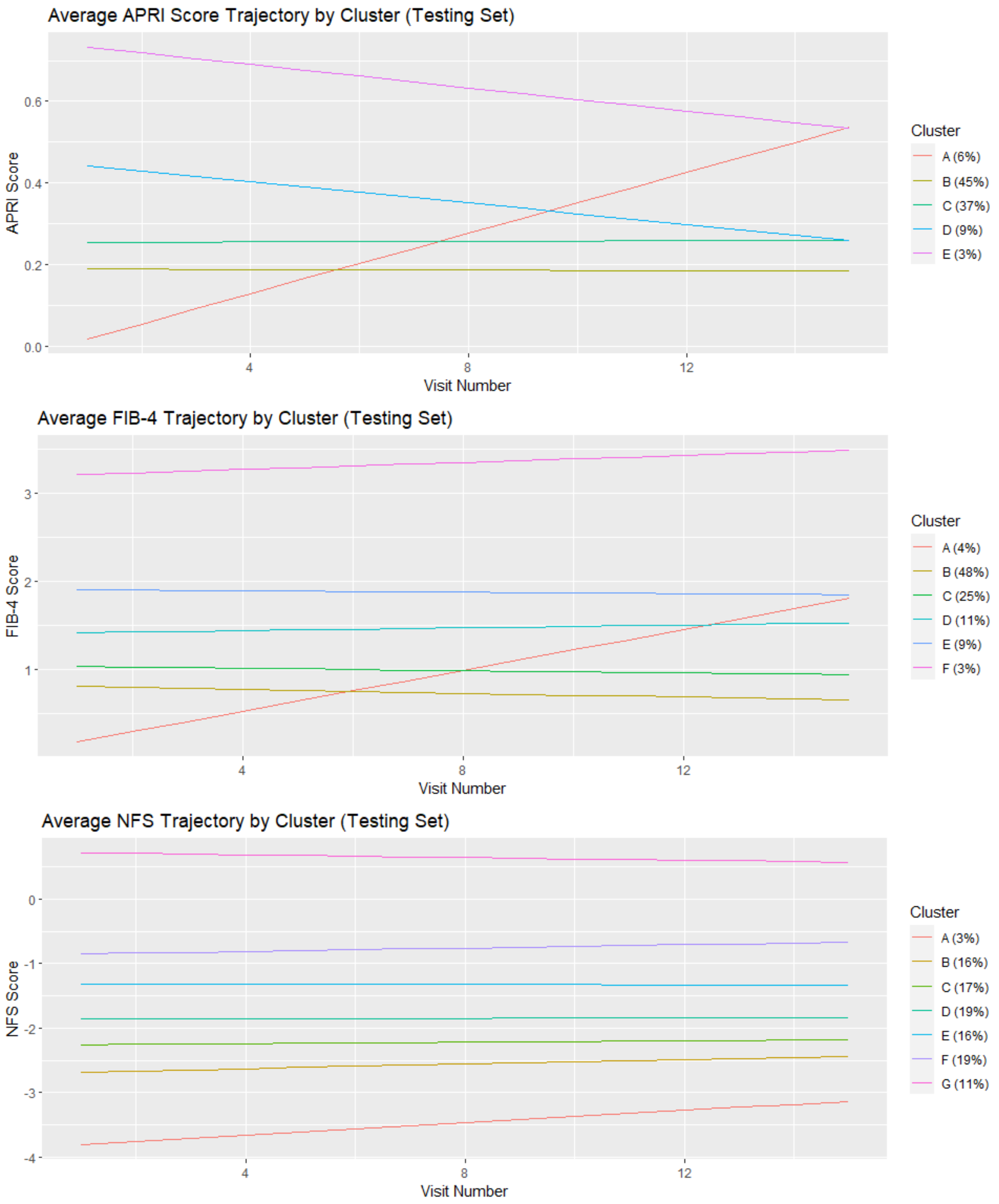


Figure A.16: Average APRI, FIB-4, and NFS Trajectories by Cluster over Visits (Testing Set)

Based on the plotted quality criteria in Figure A.16 and the corresponding elbow method, an optimal number of clusters chosen for APRI, FIB-4, and NFS were five, six, and seven respectively. Participants APRI trajectories were partitioned into clusters containing 6, 45, 37, 9, and 3% of participants. FIB-4 trajectories were partitioned into clusters containing 4, 48, 25, 11, 9, and 3% of participants. NFS trajectories were partitioned into clusters containing 3, 16, 17, 19, 16, 19, and 11% of participants. Cluster E for APRI, A and F for FIB-4, and A for NFS contained < 5% of participants (Figure A.16).

The average APRI for participants in all the clusters aside from A, which composes 6% of the participants, shows a slight decrease over the course of the follow-up period. Cluster A and F, which together compose around 7% of participants, show an upward trend over time. The other clusters show no or slight decrease over time. All the clusters for NFS show no trend or a slight upward increase over time (Figure A.16).

While the trends for the clusters for NFS appear consistent among the two sets, the clusters identified for APRI and FIB-4 show discordance (Figures A.14 and A.16). While none of the clusters identified in the training set for APRI show a sharp increase, cluster A for APRI in the testing set shows the opposite. Additionally, more participants are in a cluster that shows an increase in FIB-4 for the testing set than that of the training set (Figures A.14 and A.16).

A.5 Anchored K-Medoids

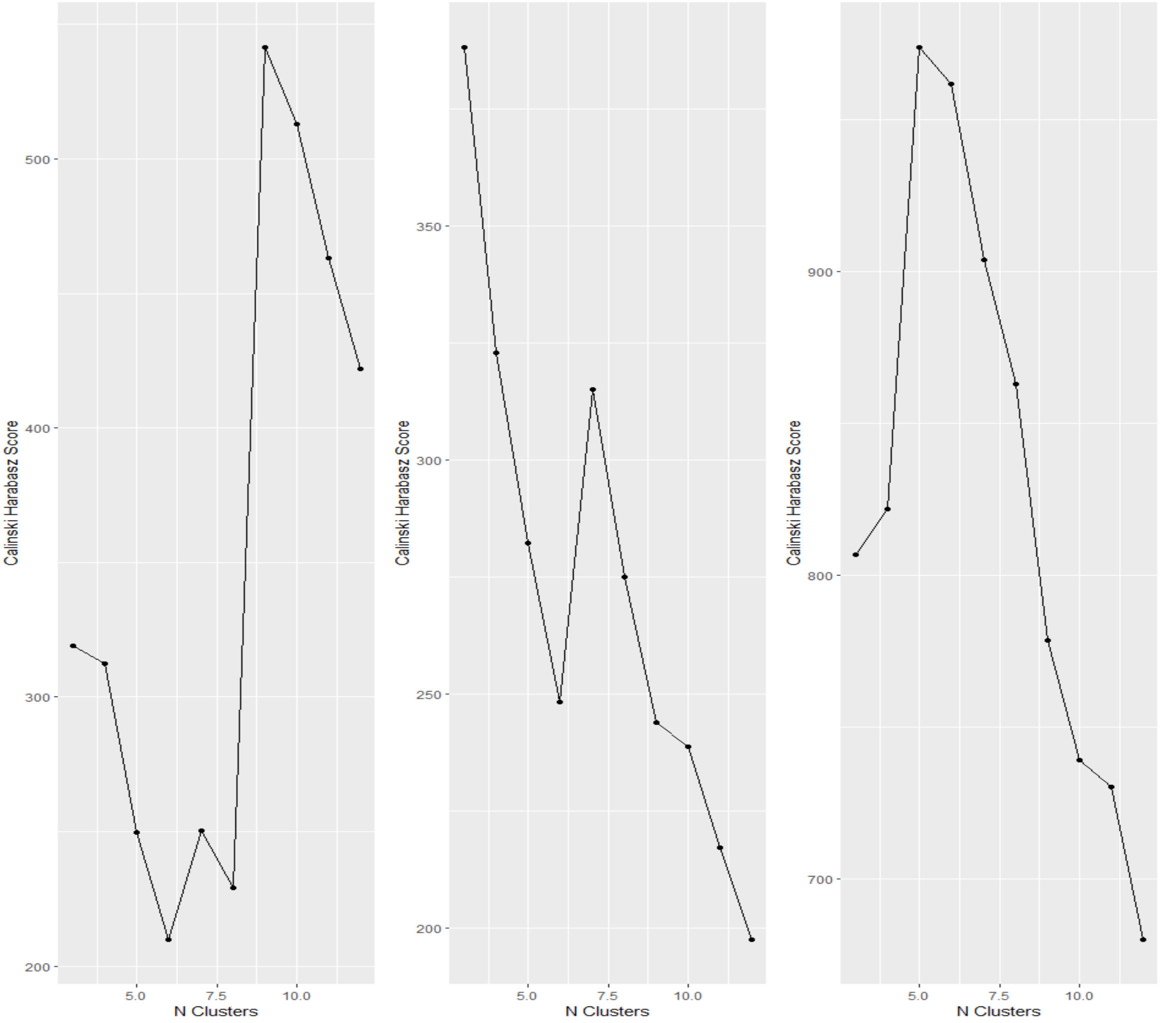


Figure A.17: Quality Criteria for APRI, FIB-4, and NFS Generalized Anchored K-Medoids Clustering (Training Set)

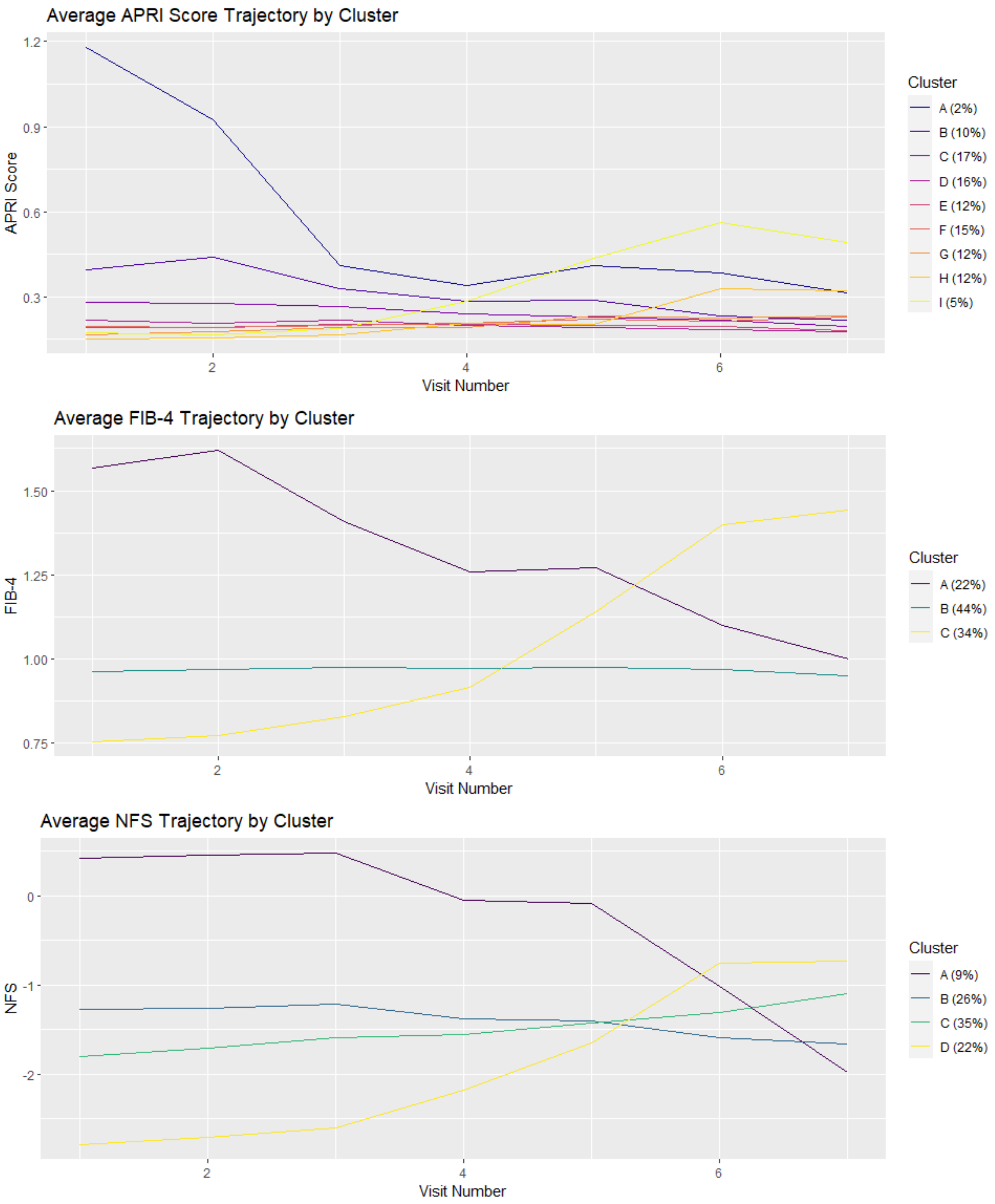


Figure A.18: Average APRI, FIB-4, and NFS Trajectories by Cluster over Visits (Training Set)

Based on the plotted quality criteria in Figure A.17 and the corresponding elbow method, an optimal number of clusters chosen for APRI, FIB-4, and NFS were nine, three, and four respectively. Participants APRI trajectories were partitioned into clusters containing 2, 10, 17, 16, 12, 15, 12, 12, and 5% of participants. FIB-4 trajectories were partitioned into clusters containing 22, 34, and 44% of participants. NFS trajectories were partitioned into clusters containing 9, 26, 35, and 22% of participants. Only cluster A for APRI contained < 5% of participants (Figure A.18).

The trajectory of the average APRI score among the clusters don’t appear to show much of a trend aside from Cluster A and I. Cluster C for FIB-4 shows an increase over time where it exceeds the decreasing cluster A at around the 5th visit and the stagnating cluster B at around the 4th. A similar trend of discordance is observed for the NFS clusters, where cluster A starts high and decreases rapidly at around the 5th visit while clusters B and C stagnate and cluster D increases (Figure A.18).

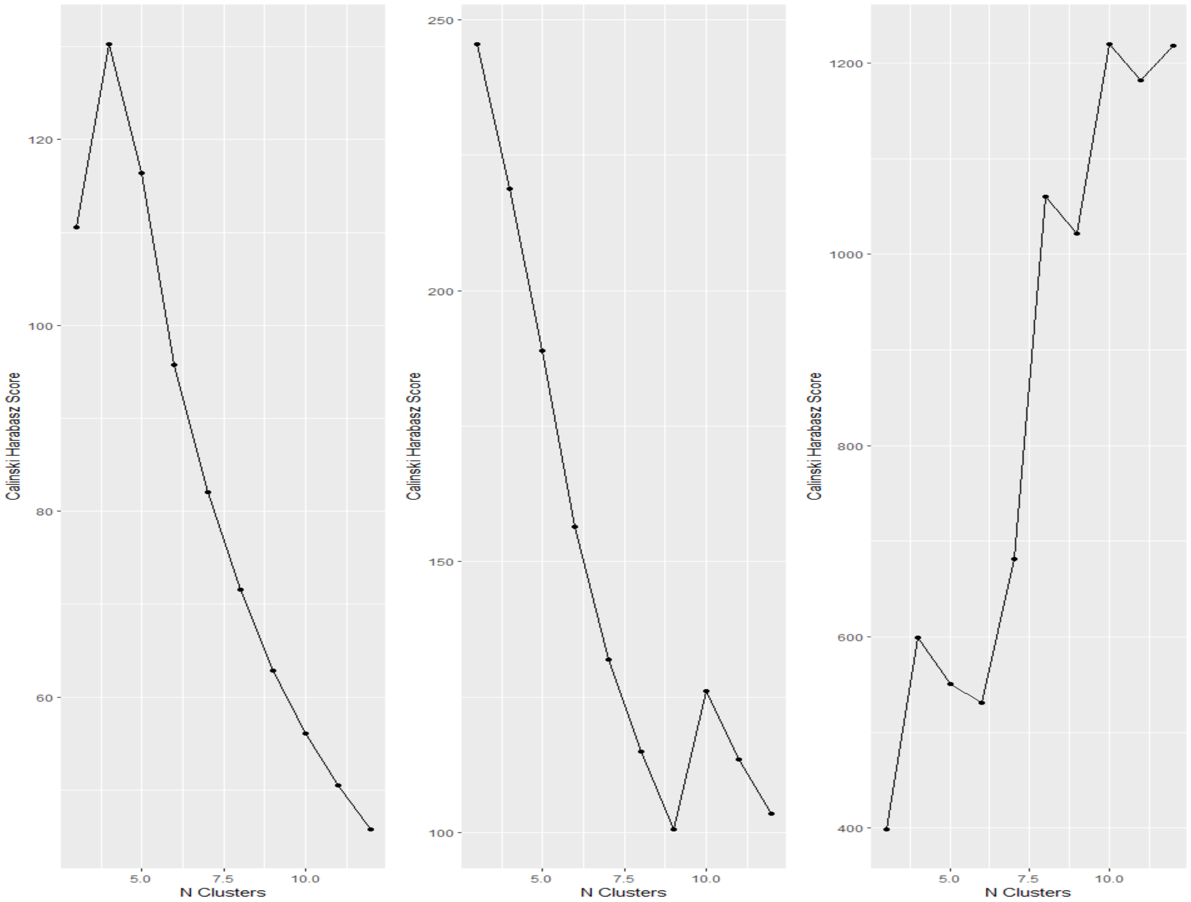


Figure A.19: Quality Criteria for APRI, FIB-4, and NFS Generalized Anchored K-Medoids Clustering (Testing Set)

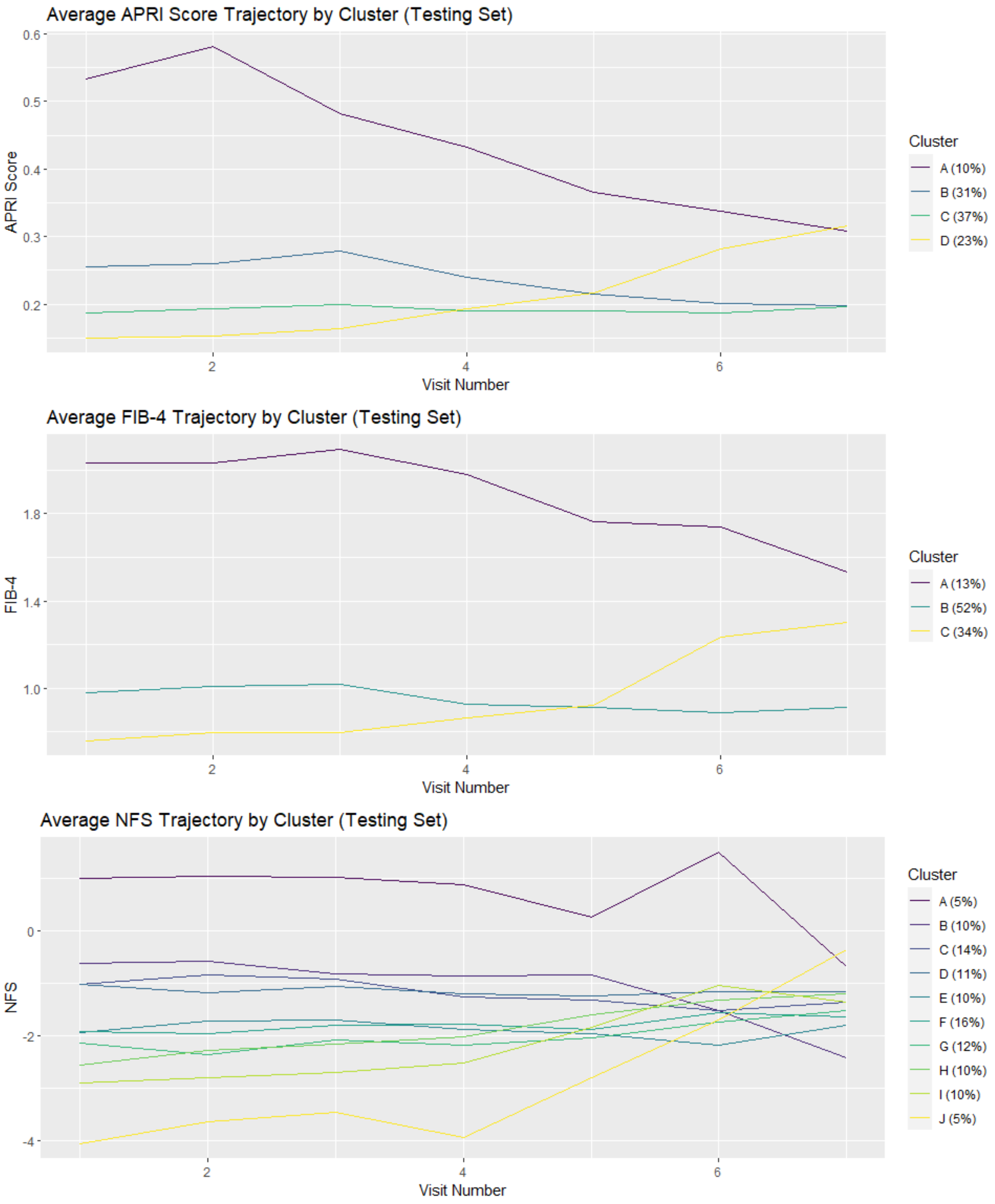


Figure A.20: Average APRI, FIB-4, and NFS Trajectories by Cluster over Visits (Testing Set)

Based on the plotted quality criteria in Figure A.19 and the corresponding elbow method, an optimal number of clusters chosen for APRI, FIB-4, and NFS were four, three, and ten respectively. Participants APRI trajectories were partitioned into clusters containing 10, 31, 37, and 23% of participants. FIB-4 trajectories were partitioned into clusters containing 13, 52, and 34% of participants. NFS trajectories were partitioned into clusters containing 5, 10, 14, 11, 10, 16, 12, 10, 10, and 5% of participants. None of the clusters for the three liver indicators contained < 5% of participants (Figure A.20).

Cluster A for APRI shows a slight increase in the average APRI and then gradually decreases over the course of the visits. Clusters C and D slightly increase at around the 3rd visit and then return to around their starting value afterwards while Cluster D begins to increase at around the 5th visit. While Clusters A and B for FIB-4 show a slight decrease over time, Cluster C also begins to increase at around the 5th visit. The ten clusters for NFS show considerable overlap over the course of the seven visits, with cluster J showing the sharpest increase and cluster B the largest decrease (Figure A.20).

Discordance between the resulting clusters of the training set and testing set can be observed in Figures A.18 and A.20. An example of this is the nine-cluster solution proposed for APRI in the training set vs. the four in the testing set. The opposite can be seen for NFS, where the training set produced a cluster solution of four while the testing set produced one of ten. The behavior of the average trajectory of the clusters for the three liver indicators also appears to be inconsistent between the two sets for all three liver indicators.