

SISCER 2023: Exploratory Analysis, MACS Data

Katie Wilson, Anna Plantinga (Instructors)

Yiqun Chen (former TA)

Contents

About This Document	1
Data Overview	1
R packages	2
Scientific Question	2
Exploratory Analyses	2
Loading and Processing the Data	3
Baseline Viral Load	3
Trajectory of CD4+ cells	4
Distribution of CD4+ cells in years 1, 2, 3, and 4 following seroconversion	6
Characterizing the correlation among CD4+ cell measurements	8
Missing Responses	10
References	11

About This Document

This file suggests key questions one might try to answer using an exploratory analysis of longitudinal data collected from the Multicenter Aids Cohort Study (MACS), a study that aimed to characterize the time course of CD4 cell depletion (Kaslow et al. 1987; see also Fitzmaurice et al. 2018). For sample code to address these questions, please see the “full” version of this document.

Data Overview

This data set contains repeated measures on $n=307$ subjects that were living with HIV, the virus that causes AIDS. The primary variable of interest is the number of CD4 cells, which is an immunologic marker of the impact of HIV that was taken at approximately semi-annual follow-up visits. The count of CD8 cells, another type of immune cell, is also included. The dataset contains measurements obtained in the first 4 years after acquisition of HIV infection and excludes subjects that died within the first 4 years.

There are several variables that give time information. First, the variable MONTHS measures the number of months since acquisition of HIV infection. The variable VTIME is the number of calendar months since January 1984. ATIME is the calendar time (months since 1/1984) at which a subject is diagnosed with AIDS, the disease caused by HIV. Note that although at MONTHS=0 a subject is diagnosed with HIV, there may be a long delay until the serious consequences of disease manifest – this latter time being the time of AIDS diagnosis. The variable ATIME is missing if a subject was not observed to be diagnosed with AIDS. Finally, DTIME is the calendar time of death (months since 1/1984) or follow-up time depending on the value of IDEATH.

Some subjects miss scheduled visits or choose to discontinue study participation and thus have fewer than the expected number of measurements.

Variables: (columns of the data file)

- ID = subject ID
- MONTHS = months since seroconversion (detection of HIV)
- AGE = age of subject
- CD4-COUNT = # of CD4 positive cells (helper cells) per mm^3
- CD8-COUNT = # of CD8 positive cells (suppressor cells) per mm^3
- VLOAD0 = viral load at baseline (copies per ml)
- AIDSCASE = 1 if no AIDS observed; 2 if AIDS observed; 3 if died prior to AIDS
- VTIME = calendar time of study visit in months since January 1984
- SCTIME = calendar time of seroconversion (detection of HIV) in months since 1/1984
- ATIME = calendar time of AIDS diagnosis in months since 1/1984
- DTIME = calendar time of death in months since 1/1984, or follow-up time
- IDEATH = indicator of death at DTIME (1=death, 0=censored)

Note 1: ATIME is missing (NA) if the time was not observed during study follow-up (i.e., subjects remained AIDS free and alive).

Note 2: There is a lower limit of detection for viral load and thus measurements at 300 reflect this detection limit.

Note 3: The ability to measure viral loads actually became available many years after the study was started, and for many subjects this measurement needed to be obtained from stored samples. Thus, not all subjects have a viral load at baseline (perhaps due to limited blood samples).

For the rest of the analysis, we will consider the baseline viral load categorized into the following groups (see Chapter 18 of van Belle et al.):

- Low viral load: baseline value less than 15×10^3
- Medium viral load: baseline value between 15×10^3 and 46×10^3
- High viral load: baseline value greater than 46×10^3

R packages

We first load the packages we will need for this activity.

```
library(tidyverse) # used for data manipulation
library(ggplot2)   # used for plotting, included in tidyverse
library(ggcorrplot) # used for visualizing cor matrix
```

If you need to install any of the listed packages (e.g., if you do not have the **geepack** package installed, you might get an error message “Error in library(geepack) : there is no package called **geepack**”. In this case, you can install the missing package by

```
install.packages('geepack')
```

Scientific Question

Is baseline viral load associated with rate of decline in CD4+ cells among men who are HIV+?

Exploratory Analyses

First, we will look at baseline viral load. What is the distribution of baseline viral load — how many individuals are in each category (low, medium, high)?

Next, we will look at CD4 counts, specifically focusing on how CD4 counts change over time. Subject-specific trajectories are helpful for seeing individual-level patterns in change. We could also look at the distribution of CD4 counts by year (we have 4 years of data) and at the correlation between them.

Finally, we will consider CD4 both over time and by baseline viral load. We could look at the distribution of CD4 counts by year by baseline viral load category (e.g., boxplots, table of mean and standard deviation).

Another important consideration is missing data. While the details of properly accounting for missing data are beyond the scope of this module, we will explore some aspects of it in exploratory data analysis. First, it is important to note how much is missing, when is it missing, specific patterns of missingness (e.g., dropout, where once patients have a missing outcome, all future outcomes are missing) and then investigate relationships with other variables. For example, does it appear that subjects with missing CD4 measurements tend to have higher baseline viral loads? Are they younger? Are the CD4 measurements that they do have lower than the CD4 measurements of the individuals who do complete the study?

Loading and Processing the Data

First we load in the data. Note that it is in the “long format,” where each line corresponds to one measurement on an individual, so there are multiple rows per individual.

```
macs <- read.csv("./macs.csv", row.names = 1)
head(macs, 5)

##      id months age cd4 cd8 vload0 aidsstage vtime sctime atime dtime ideath
## 1 1022      6  27 391 300 70737      3    18    12   NA    66     1
## 2 1022     12  27 361 596 70737      3    24    12   NA    66     1
## 3 1022     16  28 288 845 70737      3    28    12   NA    66     1
## 4 1022     27  29 378 774 70737      3    39    12   NA    66     1
## 5 1022     33  29 197 868 70737      3    45    12   NA    66     1

## check the number of unique participants
length(unique(macs$id))

## [1] 307
```

Baseline Viral Load

For the purpose of this analysis, we will create a new variable for baseline viral load category. Note that because some participants are missing baseline viral load, we have to be careful with defining the categories.

```
# Create categorical baseline viral load variable
macs <- macs %>%
  mutate(vload0_cat = as.character(cut(vload0, breaks = c(0, 15000, 46000, Inf),
    labels = c("Low", "Medium", "High"), right=FALSE))) %>%
  mutate(vload0_cat = ifelse(is.na(vload0_cat), "Missing", vload0_cat)) %>% # Make "Missing" a level
  mutate(vload0_cat = factor(vload0_cat, levels = c("Low", "Medium", "High", "Missing")))

# Check counts
unique_vload0 <- unique(macs[, c("id", "vload0_cat")])
table(unique_vload0$vload0_cat)

##
##      Low  Medium   High Missing
##      75     76     77      79

prop.table(table(unique_vload0$vload0_cat))

##
```

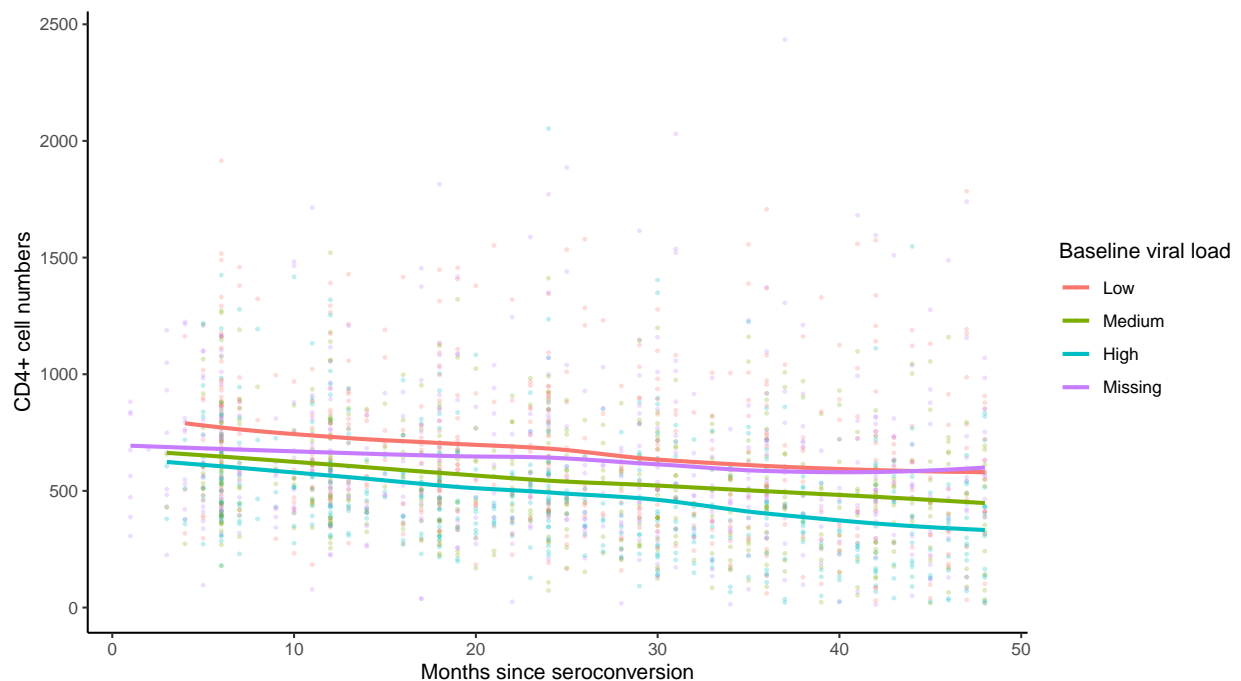
```
##           Low      Medium      High      Missing
## 0.2442997 0.2475570 0.2508143 0.2573290
```

From the table above, we see that 75 people (24.4%) are categorized as having low, 76 people (24.8%) having medium, and 77 people (25.1%) having high baseline viral load. The remaining 79 people (25.7%) are missing their baseline viral load.

Trajectory of CD4+ cells

We could first consider a simple time plot with a scatterplot smoother added by baseline viral load category:

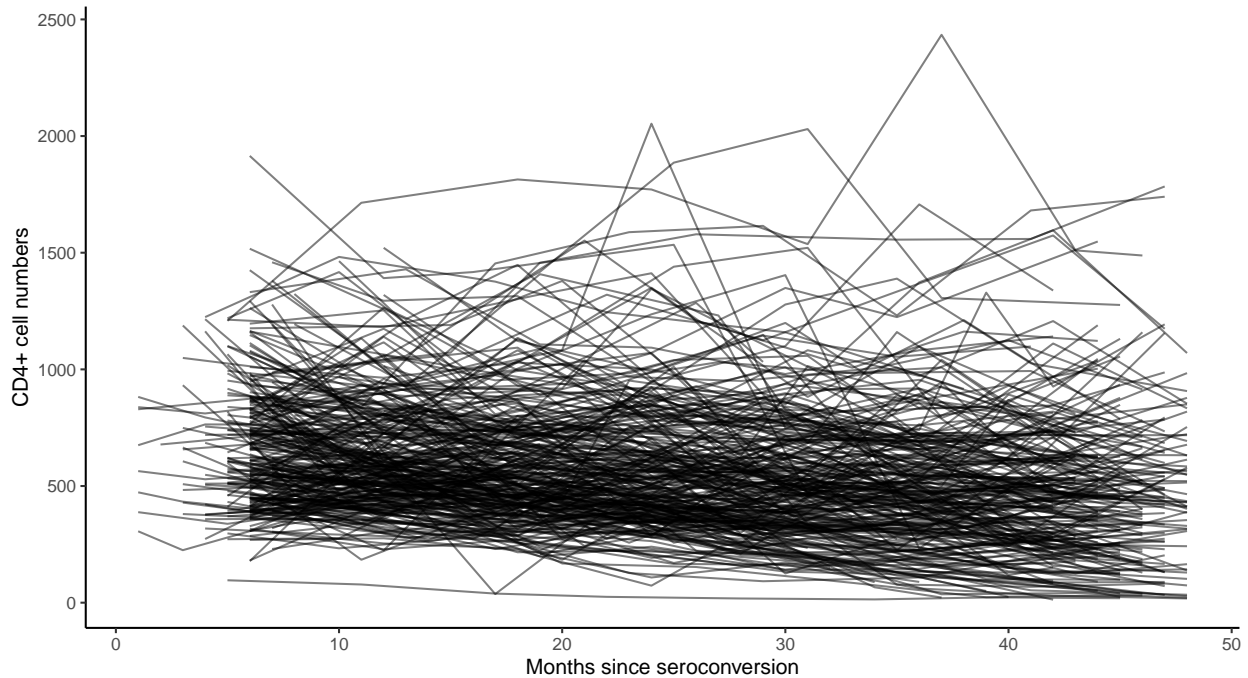
```
ggplot(macs, aes(x=months, y=cd4, group = vload0_cat, color = vload0_cat)) +
  geom_point(size=0.5, alpha = 0.25) +
  geom_smooth(se=F) +
  theme_classic() +
  ylab("CD4+ cell numbers") +
  xlab("Months since seroconversion")+
  scale_color_discrete(name = "Baseline viral load")
```



It appears there is a decrease in CD4+ cells over time (and it appears fairly linear). In addition, subjects with higher baseline viral loads tend to have a lower count of CD4+ cells on average. However, since the points for the same individual are not connected, we cannot determine right away whether the decreasing pattern holds for most individuals.

We can use the following “spaghetti plot” to visualize the individual trajectories.

```
ggplot(macs, aes(x = months, y=cd4, group=id)) +
  geom_line(alpha=0.5) +
  xlab("Months since seroconversion") +
  ylab("CD4+ cell numbers") +
  theme_classic()
```

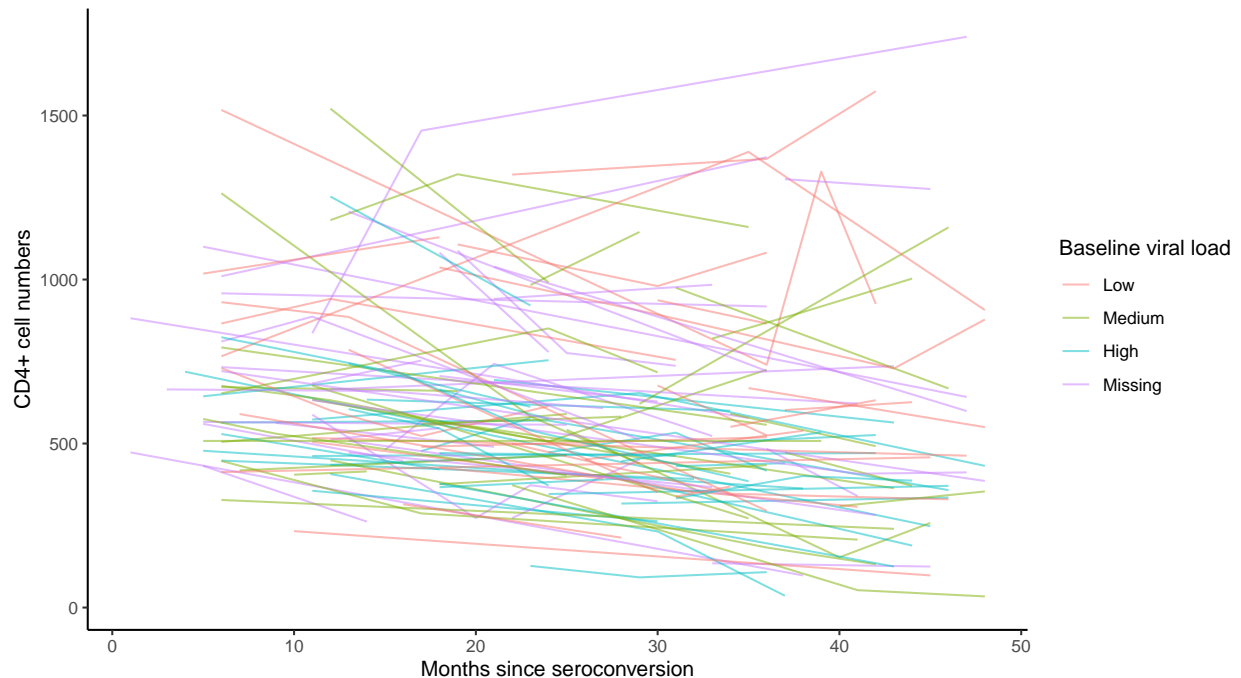


Given the large number of individuals, this plot is very busy and unhelpful. Instead, we will only highlight a few individuals. Which ones should we choose? We could select a random group or highlight ones with specific features. Here, for simplicity, we will randomly sample 20% of the participants from each of the baseline viral group.

```
set.seed(2023)

macs_subsample <- macs %>%
  group_by(vload0_cat) %>%
  sample_frac(0.2) %>%
  ungroup()

ggplot(macs_subsample, aes(x = months, y=cd4,
                           group=id, colour=vload0_cat)) +
  geom_line(alpha=0.5) +
  xlab("Months since seroconversion") +
  ylab("CD4+ cell numbers") +
  theme_classic()+
  scale_color_discrete(name = "Baseline viral load")
```



In general, we see “tracking” of the individual curves for these participants (the idea of “between-person variability” where participants tend to maintain their relative level of CD4 cell numbers, e.g., participants with higher levels tend to stay high); however, there does appear to be a bit of variability within an individual.

Distribution of CD4+ cells in years 1, 2, 3, and 4 following seroconversion

First, we create a variable `year` that indicates which year the measurement came from and obtain some summary statistics:

```
macs$year <- cut(macs$months, c(0, 12, 24, 36, 48))
table(macs$year, useNA = "always") # number of measurements in each year
```

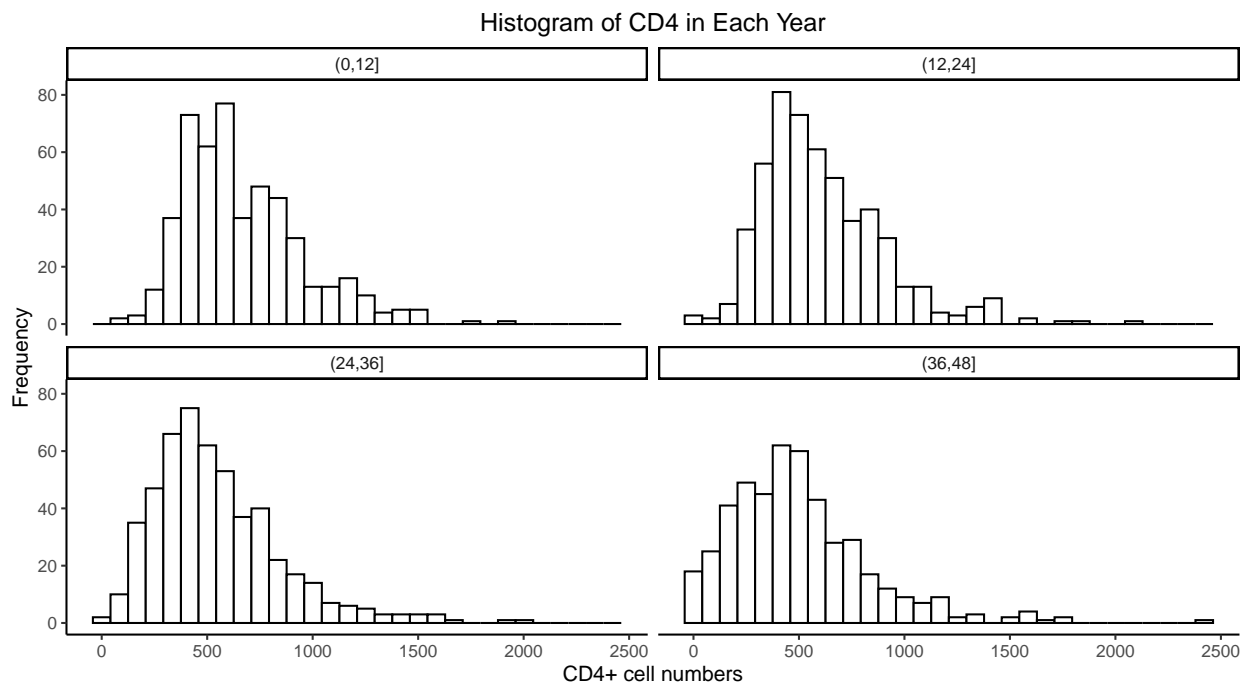
```
##
## (0,12] (12,24] (24,36] (36,48] <NA>
## 506 529 515 471 0
```

```
macs %>%
  group_by(year) %>%
  summarise(nobs = sum(!is.na(cd4)),
            nmiss = sum(is.na(cd4)),
            cd4_mean = mean(cd4, na.rm=T),
            cd4_sd = sd(cd4, na.rm=T))
```

```
## # A tibble: 4 x 5
##   year      nobs nmiss cd4_mean cd4_sd
##   <fct>   <int> <int>   <dbl> <dbl>
## 1 (0,12]    493    13    668.  282.
## 2 (12,24]   526     3    613.  289.
## 3 (24,36]   513     2    546.  307.
## 4 (36,48]   469     2    499.  331.
```

We can also use histograms to visualize the distribution of CD4 in each year:

```
ggplot(macs,aes(x=cd4)) +
  geom_histogram(color="black", fill="white")+
  xlab("CD4+ cell numbers") +
  ylab("Frequency") +
  ggtitle("Histogram of CD4 in Each Year")+
  theme_classic()+
  facet_wrap(vars(year)) + # creates one sub-graph for each year
  theme(plot.title = element_text(hjust = 0.5)) # center title
```



Next, we can obtain summary statistics (mean, sd) of CD4+ cell counts by baseline viral load category:

```
macs %>% group_by(id, year, vload0_cat) %>%
  summarise(cd4_year = mean(cd4,na.rm=T)) %>%
  group_by(year, vload0_cat) %>%
  summarise(nobs = sum(!is.na(cd4_year)),
            nmiss = sum(is.na(cd4_year)),
            cd4_mean = mean(cd4_year, na.rm=T),
            cd4_sd = sd(cd4_year, na.rm=T))
```

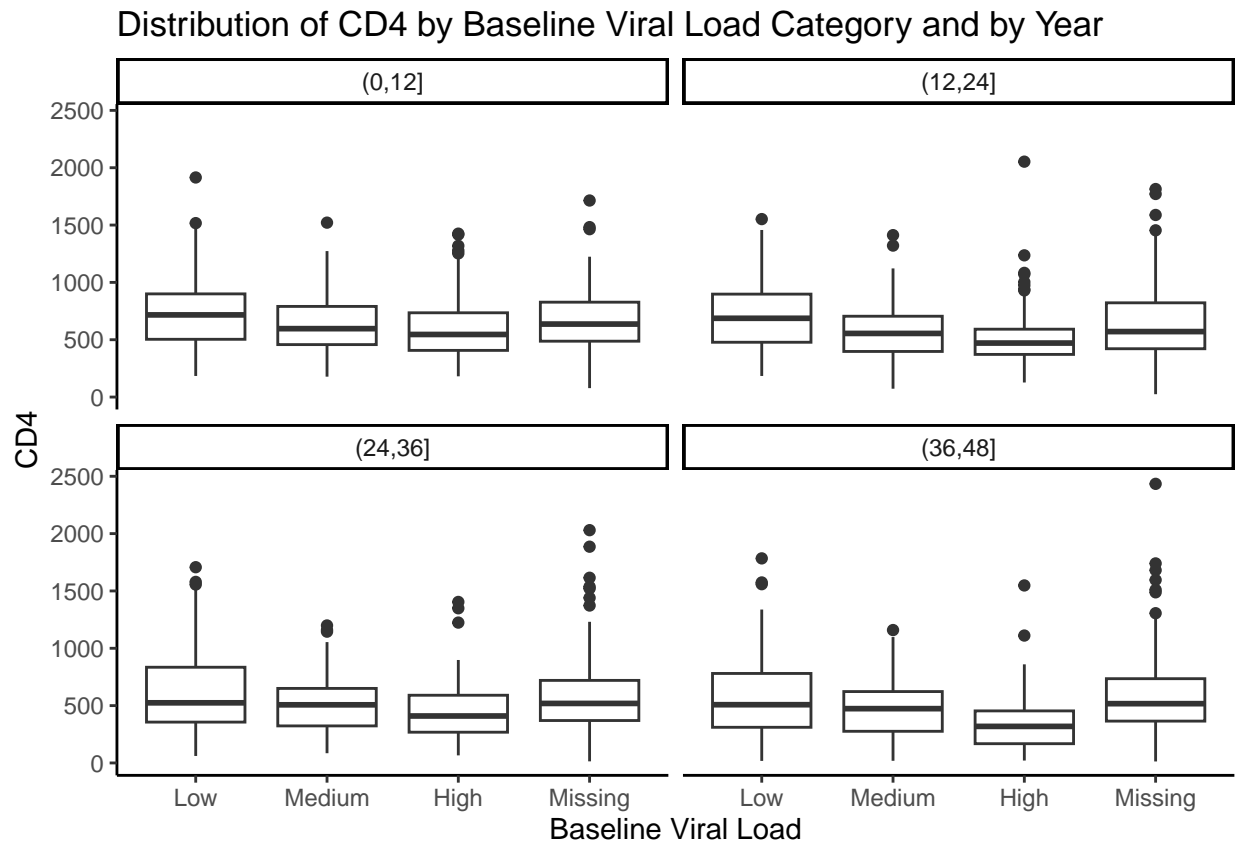
```
## # A tibble: 16 x 6
## # Groups:   year [4]
##   year   vload0_cat  nobs nmiss cd4_mean cd4_sd
##   <fct>   <fct>      <int> <int>   <dbl> <dbl>
## 1 (0,12] Low         73     2    761.  305.
## 2 (0,12] Medium     76     0    663.  256.
## 3 (0,12] High      76     1    612.  261.
## 4 (0,12] Missing   78     1    687.  265.
## 5 (12,24] Low      70     0    703.  287.
## 6 (12,24] Medium   72     0    574.  226.
## 7 (12,24] High     74     1    510.  203.
## 8 (12,24] Missing  71     0    664.  311.
## 9 (24,36] Low      65     1    634.  318.
```

```
## 10 (24,36] Medium      69    0    510.   219.
## 11 (24,36] High       71    0    446.   222.
## 12 (24,36] Missing    73    0    600.   342.
## 13 (36,48] Low       61    0    575.   351.
## 14 (36,48] Medium    64    0    471.   246.
## 15 (36,48] High      65    0    349.   250.
## 16 (36,48] Missing   71    0    566.   339.
```

From this we can see that mean CD4+ cell count tends to be lower at later time points. Within a given year, those in the lower viral load category have the highest mean CD4+ cell counts, followed by medium, with those in the high category having the lowest mean CD4+ cell counts. Those with missing baseline viral load have a mean CD4+ cell count similar to the low and medium groups. Spread (as measured by standard deviation) tends to be greater with larger mean values.

Below, we visualize the distribution of CD4 counts by baseline viral category and year.

```
ggplot(macs) +
  geom_boxplot(aes(x=vload0_cat, y=cd4)) +
  facet_wrap(vars(year)) +
  xlab("Baseline Viral Load") +
  ylab("CD4") +
  ggtitle("Distribution of CD4 by Baseline Viral Load Category and by Year") +
  theme_classic()
```



Characterizing the correlation among CD4+ cell measurements

We will compute the pairwise correlations across CD4 measurements by year. Recall that some individuals have multiple measurements in a given year; for those individuals, we will first obtain the mean CD4 per

person per year. Ultimately we want a dataset in the “wide format” where each row corresponds to an individual and there are columns for each year containing the mean CD4+ cell number.

```
mac_mean_cd4 <- macs %>%
  group_by(id, vload0_cat, year) %>%
  summarise(cd4_mean = round(mean(cd4, na.rm=T))) %>%
  pivot_wider(names_from=year, values_from=cd4_mean)
head(mac_mean_cd4)
```

```
## # A tibble: 6 x 6
## # Groups:   id, vload0_cat [6]
##       id vload0_cat `(0,12]` `(12,24]` `(24,36]` `(36,48]`
##   <int> <fct>      <dbl>      <dbl>      <dbl>      <dbl>
## 1  1022 High         376        288        288         39
## 2  1049 High         722        430        326        204
## 3  1120 Low          1144       1305       1108       1142
## 4  1164 Missing        714        477         NA         NA
## 5  1214 Medium         847        766        717        661
## 6  1235 High          384        220        158         28
```

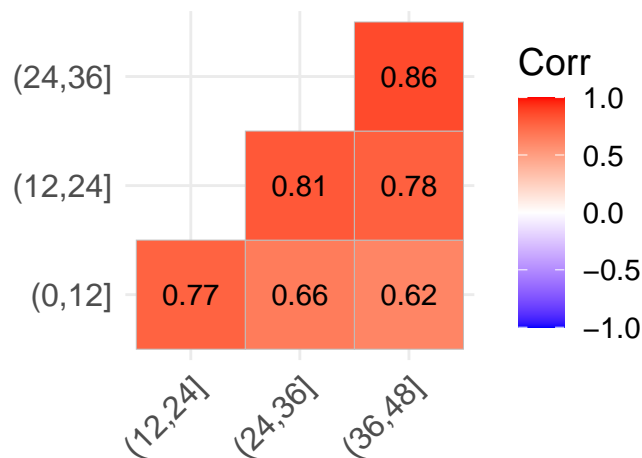
Next, we compute the correlation matrix using the `cor` function in R.

```
cor_mat <- cor(mac_mean_cd4[,c("(0,12]", "(12,24]", "(24,36]", "(36,48]")],
               use="pairwise.complete.obs")
cor_mat
```

```
##           (0,12]  (12,24]  (24,36]  (36,48]
## (0,12]  1.0000000 0.7664469 0.6577924 0.6221988
## (12,24] 0.7664469 1.0000000 0.8080691 0.7821300
## (24,36] 0.6577924 0.8080691 1.0000000 0.8581926
## (36,48] 0.6221988 0.7821300 0.8581926 1.0000000
```

Here is some code for creating a color-coded correlation plot:

```
ggcorrplot(cor_mat, lab=TRUE, type="lower") +
  theme(text=element_text(size=14))
```



Overall, there appears to be fairly strong positive correlation among the CD4 counts for the same individual, with some evidence that it decreases with increasing time lag.

Missing Responses

Note that aside from those missing baseline viral load (a covariate), we have not investigated those missing CD4 (response). According to the data description, measurements of CD4 cells were taken approximately semi-annually for the first 4 years following HIV acquisition. In other words, a subject with “complete” observations should have roughly 8 CD4 counts. Additionally, the description notes that some subjects missed scheduled visits or choose to discontinue study participation and thus have fewer than the expected number of measurements. There are many things to consider when studying missing CD4 measurements. The easiest and first one would be to look at a simple count of the number of measurements each person has:

```
visit_counts <- macs %>% group_by(id) %>% summarise(n=n(), nobs=sum(!is.na(cd4)))
visit_counts
```

```
## # A tibble: 307 x 3
##       id     n  nobs
##   <int> <int> <int>
## 1  1022     6     6
## 2  1049     8     8
## 3  1120     8     8
## 4  1164     4     4
## 5  1214     8     8
## 6  1235     8     8
## 7  1259     6     6
## 8  1290     8     8
## 9  1308     7     6
## 10 1350     7     7
## # i 297 more rows
```

```
# number of visits with cd4 measurement available
table(visit_counts$nobs)
```

```
##
##  0   1   2   3   4   5   6   7   8   9
##  2   4  12   9   9  29  46  82 111   3
```

From this, we note that there are 114 individuals that have 8 or 9 CD4 measurements.

Bonus question: what could be a next step for examining the missing responses using exploratory analyses?

References

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2018), Applied Longitudinal Analysis, Wiley & Sons, Limited, John.

Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F., and Rinaldo, C. R., Jr (1987), “The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants,” American Journal of Epidemiology, 126, 310–318. <https://doi.org/10.1093/aje/126.2.310>.

Van Belle, G., Fisher, L. D., Heagerty, P. J., & Lumley, T. (2004), Biostatistics: a methodology for the health sciences, John Wiley & Sons.