

# SISCER 2023: Exploratory Analysis, MACS Data

Katie Wilson, Anna Plantinga (Instructors)

Yiqun Chen (former TA)

## Contents

<b>About This Document</b>	<b>1</b>
<b>Data Overview</b>	<b>1</b>
R packages . . . . .	2
<b>Scientific Question</b>	<b>2</b>
<b>Loading and Processing the Data</b>	<b>2</b>
<b>Exploratory Analyses</b>	<b>3</b>
<b>References</b>	<b>3</b>

## About This Document

This file provides questions one might want 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). Code (but minimal discussion) is provided in the “partial” version of this document, and full code and discussion is provided in the “full” version.

## 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<sup>3</sup>
- CD8-COUNT = # of CD8 positive cells (suppressor cells) per mm<sup>3</sup>
- 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  $15 \times 10^3$
- Medium viral load: baseline value between  $15 \times 10^3$  and  $46 \times 10^3$
- High viral load: baseline value greater than  $46 \times 10^3$

## 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+?

## Loading and Processing the Data

```
# Load data
macs <- read.csv("./macs.csv", row.names = 1)

# 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")))

# View dataset
head(macs, 5)
```

##	id	months	age	cd4	cd8	vload0	aidscale	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

```
## vload0_cat
## 1 High
## 2 High
## 3 High
## 4 High
## 5 High
```

## Exploratory Analyses

You may want to consider how to explore the following questions:

1. What is the distribution of baseline viral load – how many individuals are in each category (low, medium, high)?
2. How do average CD4 counts change over time for groups defined by baseline viral load?
3. How do CD4 counts change over time for **particular individuals** with different baseline viral loads?
4. What is the distribution of CD4 counts by year? What is the correlation in CD4 counts across years?
5. How could one explore CD4 counts both over time *and* by baseline viral load?
6. How much missing data is present? How is it distributed – any patterns by time point, covariate values, outcome values, etc.?

## 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.