SAS® Macro to Determine Rates of Change in Markers of HIV-1 Disease Progression

Baibai Chen, Johns Hopkins University, Baltimore, MD
Cynthia A. Kleeberger, Johns Hopkins University, Baltimore, MD

ABSTRACT

Two markers to measure the degree of progression of human immuno-deficiency virus type 1 (HIV-1), namely, the level of plasma viremia and the number of $CD4^+$ lymphocytes, have been developed. Both the level and the slope of these markers play an important role in characterizing disease progression and in identifying study participants for future research. In addition, changes in markers (or "viral load") have been key in establishing the efficacy of highly active antiretroviral therapies (HAART). This article presents a simple SAS macro to compute the slopes using a longitudinal data set. The code consists of a DATA step, SAS functions, PROC FREQ and/or REG, and macro programming. The macro can be used to compute a slope from a single patient, multiple patients or an entire cohort from an epidemiological study. The macro code allows the investigator to specify the following factors for the computation of the slope: (1) the initial time point, (2) the time interval of interest, and (3) the number of time points.

INTRODUCTION

The Multicenter AIDS Cohort Study (MACS) is an ongoing prospective study of the natural history of HIV-1 infection conducted since 1984 across four centers located in Baltimore, Chicago, Pittsburgh, and Los Angeles. Between March 1984 and April 1985, the MACS enrolled a cohort of 4,954 men. Recruitment was open again from 1987 - 1991 during which an additional 668 men were enrolled. Details about the recruitment and characteristics of the MACS cohort and study protocol have been reported previously[1]. Participants were extensively evaluated and lymphocyte data and other laboratory parameters were collected and

incorporated into the central database at 6-month intervals. Data from each semi-annual period existed as a separate file. Computing the change (slope) in lymphocyte level over the course of 13.5 years, 27 files and 66,154 CD4 measurements was difficult, time consuming and prone to errors. In this paper, we investigate the use of a SAS macro to lessen the burden of analysis within a large longitudinal data set, and provide one slope per person-marker as an indication of HIV disease progression.

METHODS

A SAS macro was used to compute the slope using the following coding scheme. The macro calls macro SLOPE and returns a data set named SLOPE, which includes the patient identification number and the computed slope by year. The macro is introduced step by step using the slope of $CD4^+$ number as an example. However, the same procedures can be used to compute slopes of virus burden in the plasma.

The parameter list has the following form:

```
%SLOPE(IDLIST,DATAFILE,
INTLTIME,ENDTIME,NTEST);
```

That is, the first parameter is the patient's identification number (ID) or the name of the data set which contains the patient identification only (it could be either a SAS data file or ASCII data file, which also requires two names with a '.' in between, e.g. patient.id). The second parameter is the name of the data set which includes the longitudinal data (patient ID number and the number of CD4+ lymphocytes) at each time point and the corresponding test date. The third and fourth parameters allow you to decide the initial time point and the interval of interest. Both could be either a specific date (e.g. MDY(06,15,97)) or the name of a variable in your longitudinal data set (e.g. LASTDATE). The final parameter is the number of time points needed in the analyses. In order to compute a slope, it is necessary to have at least two time points. Depending on the analysis, more points may be advised.

Step I: Identify patients of interest

```
%MACRO SLOPE(IDLIST,DATAFILE,
INTLTIME,ENDTIME,NTEST);
```

```
DATA IDS;
 %IF %INDEX(&IDLIST,.) = 0 %THEN
    %DO;
  IF ID IN &IDLIST;
 %END;

 %ELSE %DO;
  %IF %INDEX(&IDLIST,SSD) ^= 0
    %THEN %DO;
   SET &IDLIST;
   KEEP ID;
  %END;
  %ELSE %DO;
   INFILE "&IDLIST" MISSOVER;
   INPUT ID;
  %END;
 %END;
PROC SORT; BY ID;
```

It is possible to compute a single or multiple patients' slope by (1) including the ID number(s) in parentheses, separating the IDs by a comma if there are more than one; or (2) typing the name of the data file either in SAS format or in ASCII. The first IF statement reads in the ID(s) when you type ID number(s). The ELSE IF statement is used to read in the ID if you specify a data file name.

Step II: Merge the ID with the longitudinal data set and determine the time interval

```
DATA ONE;
MERGE IDS(IN=IN1)
        SSD.&DATAFILE;
                BY ID;
%IF %LENGTH(&INTLTIME) NE 0
%THEN %DO;
 IF TESTDATE < &INTLTIME THEN
DELETE;
 %END;

%IF %LENGTH(&ENDTIME) NE 0
%THEN %DO;
 IF TESTDATE > &ENDTIME;
                %END;
PROC SORT; BY ID;
```

If you want to use all the data points in your data set, just leave space instead of a value for the variables INTLTIME and ENDTIME.

Step III: Use PROC FREQ to calculate the number of tests for each patient after the restrictions specified on the start and the end time points.

```
PROC FREQ DATA = ONE NOPRINT;
        TABLES ID / OUT = NTEST
                (KEEP=ID COUNT);
```

Step IV: Rename the date of the first test as the start date.

```
DATA START(KEEP=ID DTESTART);
 SET ONE;
        BY ID;
 IF FIRST.ID;
 RENAME TESTDATE = DTESTART;
```

Step V: Define the time from the start date of each test date in years and then run a regression on the test value by the time. Output a data set, SLOPE, which is the final data set with patients' identification number and the value of slopes

```
DATA;
  MERGE ONE START(IN=INS)
                NTEST;
  BY ID;
  IF INS;
  SLOPEYR=(TESTDATE-DTESTART)
              /365.25;
 IF COUNT >= &NTEST;

PROC REG OUTEST=      SLOPE
(KEEP=ID SLOPEYR) NOPRINT;
   MODEL CD4N = SLOPEYR /
        NOPRINT;
   BY ID;
%MEND SLOPE;
```

CONCLUSION

The SAS macro described here provides an example of using this programming technique to simplify the analysis of complicated longitudinal laboratory data. In HIV-1 infection, $CD4^+$ lymphocyte numbers and plasma viremia levels have proven to be the most powerful predictors of clinical outcome[2] and markers of therapy efficacy, and therefore have been used as our examples. Resulting slopes (trajectories) may be used to characterize distinct groups with comparable initial $CD4^+$ cell decline (i.e. stable or positive slopes, moderate decline and rapid decline)[3]. Identification of these sub-groups within a cohort study permit the investigation of factors related to pathogenic mechanisms of HIV-1 infection including studies on long-term survival. Although these are the markers selected here to demonstrate the usefulness of the macro, other laboratory parameters may be substituted within the same coding scheme (i.e. cholesterol, vitamin A).

REFERENCES

1 Kaslow RA, DG Ostrow, R. Detels, et al. The Multicenter AIDS Cohort Study: rational, organization and selected characteristics of the participants. Am J. Epidemiology 1987; 126:310-318

2 Mellors JW, A. Muñoz, J Giorgi et al. Plasma viral load and $CD4^+$ lymphocytes as prognostic markers of HIV-1 infection. Ann Intern Med. 1997; 126:946-954

3 Muñoz A, AJ Kirby, Y He et al. Long-term Survivors with HIV-1 infection: incubation period and longitudinal patterns of $CD4^+$ lymphocytes. JAIDS 1995; 8:496-505