A SAS program for calculating cumulative incidence of events (with confidence limits) and number at risk at specified time intervals with partially censored data

Alan D. Penmana, W.D. Johnsonb

a Department of Medicine, University of Mississippi Medical Center, Clinical Research Program (Building LK), 876A Lakeland Drive, Jackson, MS, United States
b Department of Preventive Medicine, University of Mississippi Medical Center, Jackson, MS, United States

Article info
Article history:
Received 10 February 2007
Received in revised form 17 July 2007
Accepted 8 October 2007

Keywords:
Survival data
Cumulative incidence
Number at risk

Abstract
Correct analysis and interpretation of longitudinal (cohort) studies with partially censored time-to-event data requires that the cumulative count of events and censored observations as well as the number at risk be calculated at appropriate time points (for example, every year), by baseline group or stratum. We present here a simple SAS program, for use in situations in which competing risks do not need to be accounted for, that calculates, by baseline group or stratum, the cumulative event count, cumulative event probability (with upper and lower 95% confidence limits), and number at risk at selected time points that can be chosen by the user. We demonstrate the use of the program in the analysis of longitudinal time-to-event data from a prospective study, the Atherosclerosis Risk In Communities (ARIC) Study, for four groups and a 10-year follow-up. The SAS code presented here is easy to follow and modify and can be incorporated quickly by the user for immediate use. It provides an especially valuable tool for less experienced SAS users.

© 2007 Elsevier Ireland Ltd. All rights reserved.

1. Introduction
In longitudinal (cohort) studies with time-to-event data a main outcome measure of interest is an estimate of the cumulative incidence probability (where the incidence probability is 1—survival probability). This may be presented in crude (unadjusted) form or, if comparisons are to be made with an external standard or among baseline groups (or strata), adjusted for baseline group differences in age, sex, race, or other covariates. The Kaplan–Meier product-limit method is used to estimate the unadjusted cumulative incidence probability, [1] and Cox proportional hazards regression is used to estimate the adjusted cumulative incidence probability by various methods [2–4]. The results of both methods are usually presented graphically with an appropriate time scale on the X-axis. Correct interpretation of these graphs, and, in particular, of group differences in the incidence curves, requires that the number at risk in each baseline group be displayed at appropriate time points, for example, every year [5]. We present here a simple SAS program that, in situations in which competing risks do not need to be accounted for, calculates the cumulative event count, cumulative event probability (with upper and lower 95% confidence limits), and number at risk at selected time points that can be chosen by the user.
Table 1 – Endpoint frequencies by follow-up time interval—first quartile

<table>
<thead>
<tr>
<th>Follow-up time interval (years)</th>
<th>Endpoint frequencies</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of events</td>
<td>Number of censored observations</td>
</tr>
<tr>
<td>0–1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>1–2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2–3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>3–4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>4–5</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>5–6</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>6–7</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>7–8</td>
<td>4</td>
<td>49</td>
</tr>
<tr>
<td>8–9</td>
<td>1</td>
<td>164</td>
</tr>
<tr>
<td>9–10</td>
<td>0</td>
<td>136</td>
</tr>
<tr>
<td>&gt;10 years</td>
<td>0</td>
<td>62</td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>434</td>
</tr>
</tbody>
</table>

2. The SAS program

The following program, written in SAS Version 9.1,[6] has three sections. The first section produces numbers of events and censored observations by time interval. The second section produces cumulative frequencies and number at risk at each time point. The third section produces cumulative incidence probabilities, with upper and lower 95% confidence limits, again by time interval. We have written the program for four baseline groups and a 10-year follow-up, but it can be easily modified by straightforward addition or deletion of code to analyze data with any number of baseline groups and any length of follow-up.

The program makes minimum use of SAS macro coding, plus (in Section 2) some SAS output delivery system (ODS) code. The following macro-variables should be set by the user:

1. In Section 1, the dataset name, the name of the variable for follow-up time, the name of the censoring variable, and the name of the grouping (or stratum) variable, thus: %macro section1 (name, time, event, group);
2. In Section 2, the number of groups (strata), the dataset name, the name of the grouping (or stratum) variable, and the number of subjects in each group (stratum), thus: %macro section2 (numgrps, name, group, grpnum1, grpnum2, grpnum3, grpnum4);
3. In Section 3: the dataset name and the name of the grouping (or stratum) variable, thus: %macro section3 (name, group).

These are the only pieces of information required for the user. Of course, if a permanent dataset is being used the macros must be preceded by a LIBNAME statement to indicate path/folder. The complete SAS program is given in Appendix A.

3. Example

We demonstrate the use of the program in the analysis of longitudinal time-to-event data from the Atherosclerosis Risk in Communities (ARIC) Study, a prospective study, begun in 1987, designed to investigate the etiology and natural history of cardiovascular disease (CVD) in four U.S. communities. The study design and procedures have been reported previously [7]. At the Jackson field center (Mississippi) only African-Americans were recruited. During the third ARIC examination (1993–1995), echocardiographic examination was performed at the Jackson field center on 2445 middle-aged and elderly African-American men and women, 51–70 years of age. Of these 2445 participants, 1792 had an adequate M-mode echocardiogram for assessing left ventricular mass and were free from prevalent stroke, left ventricular aneurysm, aortic stenosis, and mitral stenosis. The aim of the analysis was to determine the association between echocardiographically defined indexed left ventricular mass (LVMI), in quartiles, and the incidence of first stroke [8].

This dataset (dataset1) includes variables for follow-up time (ISCstrfuyrs), censoring status (IN03ISC), and group (LVMIQUARTILE) with four levels (n = 448 persons in each). The macros are run with these user-defined parameters:

- %section1 (dataset1, ISCstrfuyrs, IN03ISC, LVMIQUARTILE);
- %section2 (4, dataset1, LVMIQUARTILE, 448, 448, 448, 448);
- %section3 (dataset1, LVMIQUARTILE).

Table 2 – Cumulative frequency of endpoints and number at risk in each time interval—first quartile

<table>
<thead>
<tr>
<th>Follow-up time interval (years)</th>
<th>Cumulative frequency of endpoints</th>
<th>Number at risk in each time interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–1</td>
<td>3</td>
<td>445</td>
</tr>
<tr>
<td>1–2</td>
<td>7</td>
<td>441</td>
</tr>
<tr>
<td>2–3</td>
<td>12</td>
<td>436</td>
</tr>
<tr>
<td>3–4</td>
<td>15</td>
<td>433</td>
</tr>
<tr>
<td>4–5</td>
<td>21</td>
<td>427</td>
</tr>
<tr>
<td>5–6</td>
<td>25</td>
<td>423</td>
</tr>
<tr>
<td>6–7</td>
<td>32</td>
<td>416</td>
</tr>
<tr>
<td>7–8</td>
<td>85</td>
<td>363</td>
</tr>
<tr>
<td>8–9</td>
<td>250</td>
<td>198</td>
</tr>
<tr>
<td>9–10</td>
<td>386</td>
<td>62</td>
</tr>
<tr>
<td>&gt;10</td>
<td>448</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 3 – Cumulative incidence probabilities and 95% confidence limits—first quartile

<table>
<thead>
<tr>
<th>Duration of follow-up (years)</th>
<th>Cumulative incidence probability (%)</th>
<th>Lower 95% confidence limit</th>
<th>Upper 95% confidence limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
<td>0.1</td>
<td>0.8</td>
</tr>
<tr>
<td>3</td>
<td>0.7</td>
<td>0.2</td>
<td>2.1</td>
</tr>
<tr>
<td>4</td>
<td>0.9</td>
<td>0.3</td>
<td>2.4</td>
</tr>
<tr>
<td>5</td>
<td>1.4</td>
<td>0.6</td>
<td>3.0</td>
</tr>
<tr>
<td>7</td>
<td>2.1</td>
<td>1.1</td>
<td>3.9</td>
</tr>
<tr>
<td>8</td>
<td>3.0</td>
<td>1.8</td>
<td>5.2</td>
</tr>
<tr>
<td>9</td>
<td>3.4</td>
<td>2.0</td>
<td>5.6</td>
</tr>
</tbody>
</table>

The complete SAS output consists of 12 tables:

1. Four tables (one per quartile) of numbers of events and censored observation by year;
2. Four tables (one per quartile) of cumulative frequency and number at risk year;
3. Four tables (one per quartile) of cumulative incidence probabilities, with upper and lower 95% confidence limits, by year.

To save space, only the three tables for the first quartile (modified slightly for clarity) are shown here. For each group (quartile), the number of events (new strokes) and censored observations by time period (year) are shown in Table 1 and the numbers at risk by time period (year) are shown in Table 2. The cumulative incidence probabilities by time period (year) are shown in Table 3. The cumulative incidence graph (Fig. 1) shows the number at risk in each group (quartile) in the footnote lines; as can be seen, the numbers for the first quartile are taken from Table 2.

4. Discussion

A number of SAS macros for calculating cumulative incidence can be found on the Internet for public use (see, for example, [2–4] and [9]). They are usually more complicated and harder to follow, which may limit their adoption. The SAS code presented here is easy to follow and modify and can be incorporated quickly by the user for immediate use. It provides an especially valuable tool for less experienced SAS users.

A particularly useful output from this program, in our view, is the table of number at risk in each group, by time interval. This is not produced by any of the survival analysis procedures in SAS (PROC LIFETEST, PROC LIFEREG); it can be calculated by hand using the routine output but this can be tedious to do. Correct interpretation of survival plots requires this information, [5] and more and more journals now ask for it.

One slight limitation of the program is that no table row is created for time points without events or censored observations; for example, in Table 3, which shows the cumulative incidence probabilities for the first quartile, the row entry for...
time 6 years is missing (as well as row entries for follow-up times greater than 9 years) because there were no events in that time period. This should be kept in mind when the data are being copied into a figure.

Our program has been written for four baseline groups and a 10-year follow-up, but it can be easily modified by straightforward addition or deletion of code to analyze data with any number of baseline groups and any length of follow-up. We have also assumed that follow-up time has been recorded in years (which is then converted in Section 1 into months). However, if in the original dataset follow-up time has been recorded in months, this conversion step can easily be skipped.

It should be remembered that in situations in which competing risks need to be accounted for conventional Kaplan–Meier procedures overestimate the cause-specific cumulative incidence probabilities [9,10]. Use of this SAS macro is best reserved for the analysis of overall cumulative incidence probabilities (such as overall mortality/survival or overall event probabilities).

Appendix A. SAS CODE

************************************************************ SECTION 1************************************************************

%macro section1 (name, time, event, group);

title1 'Events & Censored Observations' ;
title2 ' ' ;
data &name; set &name;

fumonths = &time* 12; *convert years to months;

if &event = 0 then do; d = 0; endpoint = 'b Censor ' ; end;
if &event = 1 then do; d = 1; endpoint = 'a Event ' ; end;

if fumonths le 12 then grpt = ' a. 1 yr ' ;
if fumonths gt 12 and fumonths le 24 then grpt = 'b. 2 yrs ' ;
if fumonths gt 24 and fumonths le 36 then grpt = 'c. 3 yrs ' ;
if fumonths gt 36 and fumonths le 48 then grpt = 'd. 4 yrs ' ;
if fumonths gt 48 and fumonths le 60 then grpt = 'e. 5 yrs ' ;
if fumonths gt 60 and fumonths le 72 then grpt = 'f. 6 yrs ' ;
if fumonths gt 72 and fumonths le 84 then grpt = 'g. 7 yrs ' ;
if fumonths gt 84 and fumonths le 96 then grpt = 'h. 8 yrs ' ;
if fumonths gt 96 and fumonths le 108 then grpt = 'i. 9 yrs ' ;
if fumonths gt 108 and fumonths le 120 then grpt = 'j. >10 yrs ' ;

run;

proc freq data=&name;
tables &group * &grpt* endpoint / nopct norow nocol;
title2 ' By Group ' ;
run;

%mend section1;

********************************************************** SECTION 2**********************************************************

%macro section2 (numgrps, name, group, grpnum1, grpnum2, grpnum3, grpnum4);

title1 ' Number at Risk ' ;
title2 ' ' ;
%do i = 1 %to &numgrps;
proc sort data=&name; by &group; run;
ods listing close;
ods output Freq.ByGroup&i . . Table1.OneWayFreqs = Table&i
(keep = grpt Cum Frequency);
proc freq data=&name;
tables grpt / nopct list;
by & group;
run;
ods output close;
ods listing;

data Table & i; set Table & i; Number At Risk=(&&grpnum&i - CumFrequency);
proc print data=Table & i noobs; title 2 "Group &i" ; run;
%end;
%mend section2;

****************************************SECTION 3**********************************************
%macro section3 (name, group);
title 1 'Cumulative Incidence Probabilities' ;
title 2 ' ';
proc lifetest data=& name method=km noprint outsurv = out1;
time fumonths* d (0);
strata & group;
run;

data out1; set out1;
if fumonths le 12 then do; dum= 1; duration = ' 1 yr ' ; end;
if fumonths gt 12 and fumonths le 24 then do; dum= 2; duration = ' 2 yrs ' ; end;
if fumonths gt 24 and fumonths le 36 then do; dum= 3; duration = ' 3 yrs ' ; end;
if fumonths gt 36 and fumonths le 48 then do; dum= 4; duration = ' 4 yrs ' ; end;
if fumonths gt 48 and fumonths le 60 then do; dum= 5; duration = ' 5 yrs ' ; end;
if fumonths gt 60 and fumonths le 72 then do; dum= 6; duration = ' 6 yrs ' ; end;
if fumonths gt 72 and fumonths le 84 then do; dum= 7; duration = ' 7 yrs ' ; end;
if fumonths gt 84 and fumonths le 96 then do; dum= 8; duration = ' 8 yrs ' ; end;
if fumonths gt 96 and fumonths le 108 then do; dum= 9; duration = ' 9 yrs ' ; end;
if fumonths gt 108 and fumonths le 120 then do; dum= 10; duration = ' 10 yrs ' ; end;
if fumonths gt 120 then do; dum= 11; duration = '>10 yrs' ; end;
if sdf_lcl ., then delete;
cumrisk = round (100*(1 - survival), .1);
lcl = round (100*(1 - sdf_ucl), .1);
uc1 = round (100*(1 - sdf_lcl), .1);
run;
proc sort data=out1; by & group dum fumonths; run;

data out2; set out1;
by & group dum;
if last.dum;
run;
proc sort data=out2; by & group; run;
proc print data=out2 noobs;
var duration fumonths cumrisk lcl ucl;
by & group;
title2 'By Group' ;
run;
%mend section3;
REFERENCES


