Methods of Validating Cumulative Incidence Rate Plots

Lin Yan, Merck & Co., Inc., Rahway, NJ
Aiming Yang, Merck & Co., Inc., Rahway, NJ

ABSTRACT

Validating plots generated by internal SAS macro calls is an important job responsibility for SAS programmers. For validating survival plots, the accuracy of the number of patients at risk is one of the major focuses. However, this is not straightforward. In this paper, the authors will first introduce a simple method using data steps only to calculate the number of patients at risk for a cumulative incidence rate plot. Furthermore, the authors will present a more complex approach that uses the output from the product-limit table from PROC LIFETEST to verify the number of patients at risk and the cumulative incidence rates as well.

INTRODUCTION

Many plots are generated through internal SAS macros. In many cases, a graphics macro generates a plot and a corresponding table at the same time. For example, a line plot for change from baseline and a box plot, each has a corresponding table provided that contains the data. Checking the plot against the corresponding table can serve as the preliminary validation. But for some graphics plots, no table is given. In some other cases, the output table is too large or does not have all the needed information, and, therefore, it is too time-consuming or impossible to check the plots. For example, with a cumulative incidence plot macro, the estimated values are usually not provided in the output table. In this case, some extra steps are needed to validate the plot.

A cumulative incidence plot is usually used to visualize the estimated probability of an event prior to a specified time. For example, Figure 1 is a cumulative incidence plot. In this plot, the y-axis is the cumulative incidence rate and x-axis is the duration of the study in months. The numbers at the bottom of the plot are the numbers of patients at risk at specified time points, 0, 1, 2, 3, 4, 5, and 6 months for two different treatments (treatments 1 and 2).

The validation of a cumulative incidence rate plot usually includes validating the numbers of patients at risk and/or incidence rates. In Section 2.1, the authors introduce a simple method involving data steps only for validating the numbers of patients at risk. In Section 2.2, the authors introduce a more complex method for validating the incidence rate and the numbers of patients at risk. The second method uses the product-time limit table from PROC LIFETEST. The second method does more than the first one, but it is not very easy to understand. An example will be given to show the effectiveness and difference of the two methods.
# Patients at Risk

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>472</td>
<td>398</td>
<td>351</td>
<td>312</td>
<td>275</td>
<td>243</td>
<td>208</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>405</td>
<td>349</td>
<td>304</td>
<td>267</td>
<td>237</td>
<td>205</td>
<td>169</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 1 Cumulative Incidence Rate**

**A Method Involving Data Steps Only for Calculating Number of Patients at Risk**

The number of patients at risk at a time point is the number of patients who stays in the study at that time point, i.e., how many patients have not had events and have not been censored before the time point.

These numbers can be obtained by using simple data steps only. The algorithm may be as follows:

1. Order the data by treatment and duration.
2. Find the rank of duration within each treatment group
3. Find the number of patients in each treatment
4. Patients at risk at a timepoint (or duration D) = total number of patients in a treatment – the number of patients whose duration < D.

Consider a trial having two treatments 1 and 2. Assume that in the SAS dataset EVENT, from the time a patient enters a study to the event occurrence or the end of study is recorded in the value of DURATION. The CENSOR variable indicates whether a patient had an event (=1) or not (=0) and the treatment is represented by variable TREAT.

The sample code as follows calculates the number of patients at risk.

```plaintext
data event;
  set mydata;
  keep  treat censor duration ;
run;

proc sort data=EVENT out=ATRISK;
  by treat duration ;
run;

**get total number of patients in each treatment group;**
data ATRISK1;
  set ATRISK;
```
by treat duration ;
retain intrtDurationOrd ;
if first.treat then intrtDurationOrd   = 1;
else intrtDurationOrd =intrtDurationOrd +1;
if last.treat then call symput('trttot'||cat(treat),
intrtDurationOrd);
run;

**get number of patients at risk in each treatment;
proc sort data=atrisk1;
   by treat duration ;
run;
data atrisk2;
set atrisk1;
   by treat duration ;
   if treat= 1 then atrisk= &trttot1 -intrtDurationOrd +1;
   if treat= 2 then atrisk= &trttot2 -intrtDurationOrd +1;
   if first.duration=1 then output;
run;

Table 1 Code for Method 1

The output at desired time points from the code above is as follows:

<table>
<thead>
<tr>
<th>treat</th>
<th>duration1</th>
<th>atrisk</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>472</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>398</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>351</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>312</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>275</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>243</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>208</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>405</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>349</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>304</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>267</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>237</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>205</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>169</td>
</tr>
</tbody>
</table>

Table 2. Output from method 1

The numbers obtained by this method match the numbers at the bottom of Figure 1.

A Method Using Product Limit Table for Calculating Numbers of Patients at Risk and Cumulative Incidence Rates.

In addition to validating the number of patients at risk, sometimes, we also need to validate the curve. In that case, we can make use of the SAS procedure PROC LIFETEST to get the incidence rate from product-limit table. The sample code is given in Table 3.

Again, assume that in the dataset EVENT, the elapsed time, or the time duration from the time when a patient entered a study to the time when an event occurred, or to the times when the study ended is captured in the DURATION field. The CENSOR variable indicates whether a patient had an event (=1) or not (=0). The treatment is represented by variable TREAT.

ods output productlimitestimates=kmdata;
The output dataset KMDATA is shown in Table 4-1 and Table 4-2. However, the dataset is usually large. In order to save space, the tables only display a portion of the dataset.

In the output dataset KMDATA, which was generated from the product-limit estimate, the values in y-axis (cumulative incidence rate) are from the 4th column – FAILURE (see Tables 4-1 and 4-2). The number of patients at risk is the last column LEFT (see Table 4-2). The time point is in the second column – DURATION (see Tables 4-1 and 4-2).

There might be no exact time points of interest. For example, durations of months 1.000, 2.000, and 3.000 are not included in the tables. In those cases, we can pick the last available one before the time point. From Table 4-1, it can be seen that the incidence rate for treatment 1 is 0.00928 (or 0.928%) at month 1, and 0.0197 (or 1.97%) at month 2. The cumulative incidence rates for treatment 2 at month 1 and month 2 are 0.0153 (or 1.53%) and 0.0212 (2.12%), respectively. These numbers match the values in the plot.

But if there is more than one event or censored value at a time point, the output table only reports the failure rate for the last occurrence of the duration. Therefore, all failure rates will be set to missing at all occurrences of that duration except the last occurrence. For example, the failure rate in Table 4-1 has three occurrences of duration 0.3472, the value of FAILURE is only missing for the first two occurrences and the cumulative value 0.0100, or 1% is reported at the third occurrence of that duration (0.3472).

```
proc lifetest data=event;
   time duration * censor(0) ;
   strata treat;
run;
```

**Table 3 Code for Method 2**

The output dataset KMDATA is shown in Table 4-1 and Table 4-2. However, the dataset is usually large. In order to save space, the tables only display a portion of the dataset.

In the output dataset KMDATA, which was generated from the product-limit estimate, the values in y-axis (cumulative incidence rate) are from the 4th column – FAILURE (see Tables 4-1 and 4-2). The number of patients at risk is the last column LEFT (see Table 4-2). The time point is in the second column – DURATION (see Tables 4-1 and 4-2).

There might be no exact time points of interest. For example, durations of months 1.000, 2.000, and 3.000 are not included in the tables. In those cases, we can pick the last available one before the time point. From Table 4-1, it can be seen that the incidence rate for treatment 1 is 0.00928 (or 0.928%) at month 1, and 0.0197 (or 1.97%) at month 2. The cumulative incidence rates for treatment 2 at month 1 and month 2 are 0.0153 (or 1.53%) and 0.0212 (2.12%), respectively. These numbers match the values in the plot.

But if there is more than one event or censored value at a time point, the output table only reports the failure rate for the last occurrence of the duration. Therefore, all failure rates will be set to missing at all occurrences of that duration except the last occurrence. For example, the failure rate in Table 4-1 has three occurrences of duration 0.3472, the value of FAILURE is only missing for the first two occurrences and the cumulative value 0.0100, or 1% is reported at the third occurrence of that duration (0.3472).
The numbers of patients at risk are the values in column \textit{LEFT} in the product limit table as shown below in Table 4-2. As the value of \textit{LEFT} in a row is the number of patients remaining in the study at the time point whose value is (\textit{DURATION}) in the same row, the number of patients at risk at a time point is the value of \textit{LEFT} in the previous row. For example, for treatment 1, the number of patients at risk at month 1 is 398 rather than 397, and at month 2, the number of patients at risk is 351 rather than 350. Similarly, the number of patients at risk for treatment 2 at months 1 and 2 are 349 and 304, respectively.

\begin{verbatim}
<table>
<thead>
<tr>
<th>treat</th>
<th>duration</th>
<th>Censor</th>
<th>Failure</th>
<th>Left</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.000000</td>
<td>0.0000</td>
<td>0</td>
<td>0</td>
<td>472</td>
</tr>
<tr>
<td>1.000000</td>
<td>0.1069</td>
<td>1</td>
<td>.</td>
<td>471</td>
</tr>
<tr>
<td>1.000000</td>
<td>0.9377</td>
<td>1</td>
<td>.</td>
<td>400</td>
</tr>
<tr>
<td>1.000000</td>
<td>0.9454</td>
<td>1</td>
<td>.</td>
<td>399</td>
</tr>
<tr>
<td>1.000000</td>
<td>0.9680</td>
<td>1</td>
<td>.</td>
<td>398</td>
</tr>
<tr>
<td>1.000000</td>
<td>1.0273</td>
<td>1</td>
<td>.</td>
<td>397</td>
</tr>
<tr>
<td>1.000000</td>
<td>1.9444</td>
<td>1</td>
<td>.</td>
<td>351</td>
</tr>
<tr>
<td>1.000000</td>
<td>2.0084</td>
<td>1</td>
<td>.</td>
<td>350</td>
</tr>
<tr>
<td>2.000000</td>
<td>0.0000</td>
<td>0</td>
<td>0</td>
<td>405</td>
</tr>
<tr>
<td>2.000000</td>
<td>0.1223</td>
<td>1</td>
<td>.</td>
<td>404</td>
</tr>
<tr>
<td>2.000000</td>
<td>0.9767</td>
<td>1</td>
<td>.</td>
<td>349</td>
</tr>
<tr>
<td>2.000000</td>
<td>1.0042</td>
<td>1</td>
<td>.</td>
<td>348</td>
</tr>
<tr>
<td>2.000000</td>
<td>1.0266</td>
<td>1</td>
<td>.</td>
<td>347</td>
</tr>
<tr>
<td>2.000000</td>
<td>1.9121</td>
<td>1</td>
<td>.</td>
<td>305</td>
</tr>
<tr>
<td>2.000000</td>
<td>1.9806</td>
<td>1</td>
<td>.</td>
<td>304</td>
</tr>
<tr>
<td>2.000000</td>
<td>2.0139</td>
<td>1</td>
<td>.</td>
<td>303</td>
</tr>
<tr>
<td>2.000000</td>
<td>2.0832</td>
<td>1</td>
<td>.</td>
<td>302</td>
</tr>
</tbody>
</table>
\end{verbatim}

\textbf{CONCLUSIONS}

Two methods for validating cumulative incidence rate plots were introduced. In addition, an algorithm for the first method and sample code for both methods were given. The first method that involves data steps only is straightforward. It can be used to check the numbers of patients at risk. The second method uses the product-limit table from PROC LIFETEST. It is more complex and can be used to validate the number of patients at risk and the incidence rate as well.

\textbf{ACKNOWLEDGEMENT}

Authors thank John Bowen and Hong Qi for reviewing the paper and their helpful comments and thank the support from the Department of Scientific Programming, Merck Co., & Inc.

\textbf{REFERENCES}

CONTACT INFORMATION

Your comments and questions are valuable and appreciated. Authors can be reached at

Lin Yan
Merck & Co., Inc.
Rahway, NJ 07065 U.S.A.
Email: lin_yan2@merck.com

Aiming Yang
Merck & Co., Inc.
Rahway, NJ 07065 U.S.A.
Email: aiming_yang@merck.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.