_hitlist Macro -

A Tool to Obtain Consistent Results in an Integrated Database

Diane E. Foose, GlaxoSmithKline, Collegeville, Pa
David C. Izard, GlaxoSmithKline, Collegeville, Pa

copyright © GSK, 2001

ABSTRACT

When using an integrated database to support a submission to the FDA, it is critical for all programs to consistently identify and utilize the same data for analysis of the same patient subgroups. Often many programmers are assigned to perform the programming tasks of this massive effort. The utilization of the variety of programming styles and techniques available within SAS® may introduce variations in the patient populations identified, thereby creating the possibility of erroneous results.

This paper will present the concept of using a SAS macro as an easy-to-use solution to this problem. Considerations for how to design and use this "develop once, utilize many" strategy for patient population identification and reporting within your own environment will be discussed. An implementation of this concept, the macro %_hitlist, will also be presented.

INTRODUCTION

An NDA (New Drug Application) submission usually requires an integrated database of data from all protocols to be analyzed. Typically the analyses are performed on several distinct groupings of patients. A common method for creating the reporting database is to combine this data together into common data sets with common variables in one directory. It is optional to have a master data set that has all the common demographic variables. Having a master data set with the common variables and other variables that will help distinguish which patient is in which analysis group can be a great advantage to processing the data more accurately. Making use of this data set, a macro can then be created to pull in the correct patient data for each report.

DATA EXTRACTION

As in the flowchart below, in most reporting efforts the standard procedure is to use a stacking or appending program to integrate the data first. Then, each report program would extract and subset the data it needs. Each program would have to determine the correct patient population. Occasionally different programmers would not extract the data the same way and cause discrepancies across reports.

To prevent these discrepancies from occurring, a consistent approach is needed for data extraction. With a master data set and a macro to utilize this data set a dependable approach can be developed for reliable data extraction as demonstrated below.

MASTER DATA SET

A master data set is essential to the process of creating the various patient populations. It should contain at the very least the patient identifier (PID), treatment group (TRX), basic demographics (i.e. sex, age, race, etc.), and any other classification variables (i.e. INDICAT: disease indication, STDYTYPE: type of study, STDYDAYS: duration on study, etc.). The first step for using this macro is to determine what the needs are for this submission. How will the data be subsetted? What are the patient populations? Then create a variable for determining each of these groupings in the master data set. This data set should also have a variable for the main ‘by’ group
analysis (i.e. response, completers, protocol violators, etc.).

<table>
<thead>
<tr>
<th>Example of some variables from Master data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>pid</td>
</tr>
<tr>
<td>001.001.00301</td>
</tr>
<tr>
<td>001.002.00401</td>
</tr>
<tr>
<td>002.003.00501</td>
</tr>
<tr>
<td>020.001.00301</td>
</tr>
<tr>
<td>020.002.00401</td>
</tr>
<tr>
<td>020.003.00510</td>
</tr>
<tr>
<td>030.050.01002</td>
</tr>
</tbody>
</table>

THREE MAIN SECTIONS OF THE MACRO

1. Creating working copy of master data set – The first step in the macro should create a working copy of the master data set. It should be subsetted for the patient population of interest. Having an option for additional subsetting is beneficial, for example, if a report only needs all males, or elderly patients, within a particular subgrouping. This step should have one record for each patient/treatment combination (if a patient can be on more than one treatment).

2. Merge master with each data set desired – The next step is to merge the work.master data set with the data sets needed for the analysis. The macro should be able to pass in which variables to keep and any additional subsetting.

3. Special process handling (optional) - If any data sets need to be further processed before they can be used, then the last section of the macro can have a placeholder for these cases. An example could be if the lab data set has to be specially processed to determine the correct baseline record for extension patients; or if relative day needs to be recalculated for the combined feeder and extension study patients.

WHAT ARE WE LEFT WITH?

The output from this macro will be the specific data sets required for a program including only the patients in the analysis group desired. It is further enhanced with only the variables needed. A work copy of the master data set will also be available that can be used for patient counts (n=xxx). At this point the analysis of the data can be performed with the correct patient population.

EXAMPLE USING _HITLIST MACRO

The remainder of this paper will describe the actual creation and implementation of the _hitlist macro. This macro was created to extract up to 10 data sets from the integrated database for the patient population of interest. It also has the ability to specify for each data set only those variables needed and if there are any special subsets required for each table. For the submission effort in which this code was used, there were seven possible groupings of patients, where a patient could be in more than one group. List of the Groups:

A = Controlled studies
B = Uncontrolled studies
C = All studies for COPD
D = Controlled studies of 24-weeks duration
E = Controlled dose-ranging studies
F = Patients in COPD studies with >= 6 months exposure
G = All Studies for Asthma

An added complication to this integrated database is that three protocols were extension studies with patients continuing with the same or a different treatment and with a new Patient Identifier (PID) number. The requirement for all reports was to use their feeder PID number on all listings.

The macro was programmed to use a master (mart_master) data set that contains all the information needed to determine which populations a patient belongs to. It holds information like what type of study the patient is enrolled in (stdytype); if it is an extension study (extstdy); what indication it is for (indicat); plus other basic demographics such as age, race, sex, treatment, etc.

The first step of the macro has specific coding to classify each patient into one of the seven distinct populations. This will enable the program to extract the correct patients and create the final work_master data set that contains one record per feeder PID and treatment. For example, if a patient switched from ‘Placebo’ in the feeder study to ‘Study Drug’ in the extension study then this patient would have two records in the work_master data set, otherwise if they stayed on the same treatment there would only be one record for this patient. In the example below of an abbreviated work_master data set, PID 001.001.00301 has only one record since this patient had the same treatment in both the feeder and the extension study. PID 002.003.00501 has two records, one for each treatment.

| Example of work_master data set
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pid</td>
</tr>
<tr>
<td>001.001.00301</td>
</tr>
<tr>
<td>001.002.00401</td>
</tr>
<tr>
<td>002.003.00501</td>
</tr>
<tr>
<td>020.001.00301</td>
</tr>
<tr>
<td>020.002.00401</td>
</tr>
<tr>
<td>020.003.00510</td>
</tr>
<tr>
<td>030.050.01002</td>
</tr>
</tbody>
</table>

Due to the complexity of having extension patients with different patient numbers, a temporary work_master data set is created with records from each study preserving the feeder and extension PID numbers. This data set is used to merge with each of the requested data set(s) to pull the correct patients data.

| Example of work_master data set
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pid</td>
</tr>
<tr>
<td>001.001.00301</td>
</tr>
<tr>
<td>001.002.00401</td>
</tr>
<tr>
<td>002.003.00501</td>
</tr>
<tr>
<td>020.001.00301</td>
</tr>
<tr>
<td>020.002.00401</td>
</tr>
<tr>
<td>020.003.00510</td>
</tr>
<tr>
<td>030.050.01002</td>
</tr>
</tbody>
</table>
After creation of the requested data set(s), special processing would conditionally take place if the group requested an extension protocol. The feeder PID and protocol number would be assigned for all records for a patient so it would correspond with the \texttt{work._master} data set.

\texttt{/* Correct pid \\& prot to use acute pid number */}
\texttt{%if \&extensn \then \&do;}
\texttt{data \&\&data\&i;}
\texttt{set \&\&data\&i;}
\texttt{if \&apid \^= \'' \then \&do;}
\texttt{pid = \&apid;}
\texttt{prot = \texttt{str}ubstr(pid,1,3);}
\texttt{end; \&end;}
\texttt{run;}

For several data sets other special processing was necessary, consequently separate code was created to handle these specific situations. For these data sets there would be a program using the naming convention of \texttt{h_\&data\&i} to make adjustments to the working data set as needed. For example, for the lab data set a program was created called \texttt{h_lab.sas}. This would be "%included" into the \_hitlist macro for processing. It needs to be "%included" since the resulting program name could be greater than 8 characters and therefore could not be a simple macro call. The \texttt{h_lab.sas} program recalculated the relative days from the start of the feeder study for the group of interest is available for patient counts or as use for calculating percentages.

\[ %\texttt{macro \_hitlist(} \]
\texttt{\texttt{data1=}\&ae, \texttt{var1=}\&ae\_pref, \texttt{ae\_body, \texttt{subset1=}\&ae\_pop='ON');} \]

\textbf{HOW TO CALL THE MACRO}

The intent is to call this macro at the start of each reporting program. To obtain the data sets needed in each program the names of the data sets, variables and any additional subsetting for each data set are passed into the macro. This is achieved by using keyword parameters of \texttt{data\&i=data10, var\&i=var10, subset\&i=subset10, base\&i-base10}. The \texttt{data} macro variables are for the data sets, the \texttt{var} variables are for the names of the variables to be kept, the \texttt{subset} variables are used for if any additional subsetting is needed, and the \texttt{base} variables are for optionally keeping the baseline record for the given data set.

\begin{tabular}{|l|l|l|l|l|}
\hline
\textbf{AE data set before hitlist macro call} & & & & \\
\hline
\textbf{Pid} & \textbf{ae\_pref} & \textbf{ae\_body} & \textbf{ae\_pop} & \textbf{rel} \\
\hline
001.001.00301 & Headache & CNS System & \texttt{ON} & 1 \\
001.002.00401 & Nausea & GI System & \texttt{ON} & 2 \\
002.003.00501 & Dyspnea & Respiratory & \texttt{ON} & 3 \\
020.001.00301 & Headache & CNS System & \texttt{ON} & 1 \\
020.002.00401 & Dizziness & CNS System & \texttt{POST} & 1 \\
020.003.00501 & Headache & CNS System & \texttt{ON} & 1 \\
020.004.00510 & Tooth Ache & GI System & \texttt{ON} & 1 \\
030.050.01002 & Gingivitis & GI System & \texttt{ON} & 2 \\
\hline
\end{tabular}

\begin{tabular}{|l|l|l|l|l|}
\hline
\textbf{AE data set after hitlist macro call} & & & & \\
\hline
\textbf{Pid} & \textbf{ae\_pref} & \textbf{ae\_body} & \textbf{ae\_pop} & \textbf{trx} \\
\hline
001.002.00401 & Nausea & GI System & \texttt{ON} & 1 \\
002.003.00501 & Dyspnea & Respiratory & \texttt{ON} & 1 \\
001.003.00101 & Headache & CNS System & \texttt{ON} & 1 \\
002.003.00510 & Headache & CNS System & 0 & \\
002.004.00510 & Tooth Ache & GI System & \texttt{ON} & 1 \\
\hline
\end{tabular}

This call to macro \_hitlist will create a \texttt{work.AE} data set with all patients in group A and only contain records where \texttt{ae\_pop} = \texttt{ON}. The variables on the data set will be a few standard variables such as \texttt{PROT, PID, TRX} and the requested \texttt{AE\_PREF} and \texttt{AE\_BODY} variables. The PID number is now the feeder PID for all patients.

The advantage of using this code is that every time a program calls for "Group A" patients, the \texttt{work._master} data set will be exactly the same. The total counts for that group can be derived from this data set. Additionally, each time a request for a data set for a specific patient population is used, it will consistently return the correct patients and treatment assignments.
CONCLUSION

Each submission has different needs and requirements. These variations can be programmed into your version of the macro. An organized approach to data extraction is necessary to help ensure accurate results. This approach is not limited to submission work, it can also be implemented in every day protocol work. The macro _hitlist is an example of one way to produce consistent output.

ACKNOWLEDGEMENTS

SAS and all other SAS Institute Inc. product or service names are registered trademarks of SAS Institute Inc. in the USA and other Countries.® indicates USA registration.

CONTACT INFORMATION

Diane E. Foose
GlaxoSmithKline
1250 S. Collegeville Road, OK1105
Collegeville, Pa 19426-0989
(610)917-7168 (W)
e-mail: diane_foose-1@gsk.com

David C. Izard
GlaxoSmithKline
1250 S. Collegeville Road, UP4335
Collegeville, Pa 19426-0989
(610)917-6911(W)
e-mail: david_c_izard@gsk.com