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.

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, treatment group, basic demographics (i.e. sex, age, race, etc.), and any other classification variables (i.e. ITT, disease indication, type of study, duration on study, etc.). You first need to determine what your needs are for your submission. How will you be subsetting your data? What patient populations do you have? Then create a variable for determining each of these groupings. This data set should also have a variable for your main ‘by’ group analysis (i.e. response, completers, protocol violators, etc.).

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. Additional subsetting might be called for, so this should be another option. For instance, if you just need 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 subsetted master data set with the data sets you need for your analysis. You should be able to pass it which variables to keep and any additional subsetting.

3. Special process handling (optional) - If certain data sets have to be further processed before it can be used, then the last section of the macro can have a placeholder for these cases. An example could be if your lab data set has to be 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 your program for 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 your 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 _hitlist macro was created to fill the need to generate consistent output for all subpopulations across the many reports needed for a NDA submission. It 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 those variables you want to keep and if there are any special subsettings for each table. In the circumstances 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 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 the seven distinct populations. This will enable you 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 the work._master data set, PID 001.001.00301

<table>
<thead>
<tr>
<th>pid</th>
<th>indicat</th>
<th>stdy type</th>
<th>trx</th>
<th>ext stdy</th>
<th>apid</th>
</tr>
</thead>
<tbody>
<tr>
<td>001.001.00301</td>
<td>COPD</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>001.002.00401</td>
<td>COPD</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>002.003.00501</td>
<td>COPD</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>020.001.00301</td>
<td>COPD</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>001.001.00301</td>
</tr>
<tr>
<td>020.002.00401</td>
<td>COPD</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>001.002.00401</td>
</tr>
<tr>
<td>020.003.00510</td>
<td>COPD</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>002.003.00510</td>
</tr>
<tr>
<td>030.050.01002</td>
<td>ASTHMA</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

...
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.

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.

After creation of the requested data set(s), special processing for the extension patients would then conditionally take place if the group requested had an extension protocol. The feeder PID and protocol number would be assigned for all records for a patient so it would correspond with the work._master data set.

The last step in this macro is an option to keep only the baseline (first) record for each patient if that was needed.

Once this step was completed you had your working data sets for use in your individual program with the correct patient population. Each patient is represented by their feeder PID number, but may have different treatments for each protocol they came from. You would also have the _master work data set with one record per PID and treatment for the group of interest.

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 your program you pass in the names of the data sets, variables and any additional subsetting for each data set. This is achieved by using keyword parameters of data1-data10, var1-var10, subset1-subset10, base1-base10. The data macro variables are for your data sets, the var variables are for the names of the variables you wish to keep, the subset variables are used for if any additional subsetting is needed, and the base variables are if you need only the baseline record for the given data set.

```sas
%_hitlist(
  data1=, var1=, subset1=, base1=N,
  data2=, var2=, subset2=, base2=N,
  data3=, var3=, subset3=, base3=N,
  data4=, var4=, subset4=, base4=N,
  data5=, var5=, subset5=, base5=N,
  data6=, var6=, subset6=, base6=N,
  data7=, var7=, subset7=, base7=N,
  data8=, var8=, subset8=, base8=N,
  data9=, var9=, subset9=, base9=N,
  data10=, var10=, subset10=, base10=N);
```
Example of use in a program:

<table>
<thead>
<tr>
<th>pid</th>
<th>ae_pref</th>
<th>ae_body</th>
<th>ae_pop</th>
<th>rel</th>
</tr>
</thead>
<tbody>
<tr>
<td>001.001.00301</td>
<td>Headache</td>
<td>CNS System</td>
<td>PRE</td>
<td>1</td>
</tr>
<tr>
<td>001.002.00401</td>
<td>Nausea</td>
<td>GI System</td>
<td>ON</td>
<td>2</td>
</tr>
<tr>
<td>002.003.00501</td>
<td>Dyspnea</td>
<td>Respiratory</td>
<td>ON</td>
<td>3</td>
</tr>
<tr>
<td>020.001.00301</td>
<td>Headache</td>
<td>CNS System</td>
<td>ON</td>
<td>1</td>
</tr>
<tr>
<td>020.002.00401</td>
<td>Dizziness</td>
<td>CNS System</td>
<td>POST</td>
<td>1</td>
</tr>
<tr>
<td>020.003.00501</td>
<td>Headache</td>
<td>CNS System</td>
<td>ON</td>
<td>1</td>
</tr>
<tr>
<td>020.004.00510</td>
<td>Tooth Ache</td>
<td>GI System</td>
<td>ON</td>
<td>1</td>
</tr>
<tr>
<td>030.050.01002</td>
<td>Gingivitis</td>
<td>GI System</td>
<td>ON</td>
<td>2</td>
</tr>
</tbody>
</table>

%let group = A;  
%_hitlist(data1=ae,var1=ae_pref ae_body, subset1=ae_pop='ON');  

<table>
<thead>
<tr>
<th>pid</th>
<th>ae_pref</th>
<th>ae_body</th>
<th>ae_pop</th>
<th>trx</th>
</tr>
</thead>
<tbody>
<tr>
<td>001.002.00401</td>
<td>Nausea</td>
<td>GI System</td>
<td>ON</td>
<td>1</td>
</tr>
<tr>
<td>002.003.00501</td>
<td>Dyspnea</td>
<td>Respiratory</td>
<td>ON</td>
<td>1</td>
</tr>
<tr>
<td>001.001.00301</td>
<td>Headache</td>
<td>CNS System</td>
<td>ON</td>
<td>1</td>
</tr>
<tr>
<td>002.003.00510</td>
<td>Headache</td>
<td>CNS System</td>
<td>ON</td>
<td>0</td>
</tr>
<tr>
<td>002.004.00510</td>
<td>Tooth Ache</td>
<td>GI System</td>
<td>ON</td>
<td>1</td>
</tr>
</tbody>
</table>

This call to macro _hitlist will create a work.AE data set with all patients in group A and only contain records where ae_pop = 'ON'. The variables on the data set will be a few standard variables such as PROT, PID, TRX and the requested AE_PREF and 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 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 you request a data set for a specific patient population 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 your 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  
450 Cresson Blvd, OK1105  
Oaks, Pa 19456  
(610)917-7168 (W)  
e-mail: diane_foose-1@sbphrd.com

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