

Challenges Involved with the Integration of Clinical Study Data

Ryan DeCosta, Satellite, London, UK

ABSTRACT

It is in a pharmaceutical company's best interest and also an FDA requirement to review the safety of drugs in clinical development on an ongoing basis. In order to perform meta-analyses on multiple studies and across different clinical phases there are a number of challenges involved with pooling the data. Variations in study indications, study durations, data standards and medical dictionaries all contribute to the resource intensive nature of producing integrated summaries. The author reflects on his recent experience of creating such summaries and how some of these challenges were overcome.

INTRODUCTION

An Integrated Safety Output (ISO) is a display such as a Table, Listing or Graph that is produced after pooling or combining data across several clinical studies to summarise safety aspects of the drug. The FDA requires annual IND (Investigational New Drug) updates which involve integrating all completed studies into the existing data in a cumulative manner. There are some important considerations to address when approaching this work.

CONSIDERATIONS FOR DATA INTEGRATION

DATA STANDARDS

The time from discovery to approval of a new drug can be around 15 years. During this time systems change and standards are refined. To maximise the expandability of the integrated data it makes sense to map legacy data to the most current data standards. When doing so, decisions that are made need to be thoroughly documented. An example of where this might be encountered is when looking at values of RACE. New FDA standards suggest a 2-tier categorisation of Race, those being Race and Ethnicity. However, in older standards there may only be a single level of categorisation.

Old Standard	New Standard	ETHNICITY
RACE	RACE	
Asian	South Asian	Hispanic or Latino
Oriental	East and South-East Asian	Not Hispanic or Latino
White	White – Arabic/North African Heritage	
	White – White/Caucasian/European Heritage	

In the above case, one might map 'Asian' to 'South Asian' and 'Oriental' to 'East and South-East Asian'. As Ethnicity cannot be guessed for the older studies it might be omitted from ISO summary tables. Alternatively a single, higher level of categorisation might be used with both 'Asian' and 'Oriental' being classed in a single 'Asian' category and 'White – Arabic/North African Heritage' and 'White – White/Caucasian/European Heritage' being classed as 'White'. It is imperative to the success of future integration efforts that documentation is in place to describe how the mapping has been applied, stating the reasoning behind the decision which might be influenced by study specific requirements.

When mapping data between standards, it promotes clarity and facilitates the future maintenance of programs if mapping is done within a separate program at the individual study level rather than within the program that integrates the datasets.

CODE AUTOMATION

In order to reduce future programming effort, code was written, where possible, in a style to automate the integration of future completed studies. In addition the code had to be written as efficiently as possible anticipating large datasets such as laboratory data which exceeded 600mb. Some of these efficiencies included dropping variables not required when reading in the dataset, performing any sub-setting before commencing data manipulation, using (precompiled) functions where possible and where large datasets needed sorting by a number of variables, creating a concatenated variable of those variables required for sorting.

PhUSE 2006

E.g. Writing SAS code to automate the integration of new studies.

```
%let studystr= study1 study2 study3 study4 study5;

%macro integrate_pop();

data popiso;
  set
  %let count=1;
  %let study=%scan(&studystr,&count);
  %do %while(%quote(&study)~=);
    &study..pop
    %let count=%eval(&count+1);
    %let study=%scan(&studystr,&count);
  %end; ;
run;

%mend integrate_pop;
%integrate_pop;
```

In the above code the librefs to the study data were named study1 to studyn. The macro variable studystr is then broken into words and for each word, which equates to the libref, the pop dataset is read in and appended onto the existing dataset. If macro variable studystr is defined as a global macro variable in the autoexec.sas then new studies (i.e. study6 and so on) can be added to the string in one central location and their data automatically integrated by the programs that use this method.

RE-CALCULATIONS

In order that the pooled data were reported consistently and sensibly, a decision was made to recalculate visit windows. The studies to integrate were a combination of Acute (single dose) and Chronic (multiple doses over a longer period) durations. Visit windows were required to be constructed in such a way that they captured as much data as possible for the ISO reporting, whilst maintaining any subtle variations between phases.

The integrated visit windowing had a knock-on effect and subsequently there was a requirement to recalculate baseline, corresponding change from baseline variables, evaluability for tabulation flags (for instances where more than one assessment fell within a visit window) and other flags such as change from baseline indicators. A drawback of recalculating this is that it made it very difficult to match the numbers back to individual study outputs. In addition there is a risk of losing intentional study specific logic. However, as a benefit of doing this, there was confidence that the data was being reported consistently across all studies which ultimately was the factor that determined this decision.

REDUCING THE VOLUME OF OUTPUT

Where a typical output for a study would have around 2 to 4 treatment groups, the ISO integration had 15 treatment groups. This induced wrapping onto numerous pages for what normally would be a 1 page output. In addition data had to be reported by study indication, increasing the amount of output significantly. A key requirement in order to make it manageable was to reduce the amount of output yet not cram too much onto a page. So how was this done? :

1. "A picture paints a thousand words". Where possible data was presented graphically as this is often more efficient and easier to interpret. The data was such that we could produce individual plots of each treatment group vs placebo to avoid too many lines appearing on a single graph.
2. A decision was made not to use completetypes options within Proc Summary and only report the data that was there. This removed many pages of what would be zero filled rows.
3. At some point of ISO integration it might be worth considering combining treatment groups if feasible (into dose ranges). This is ultimately a Lead Statistician's decision but worth investigating it if there are simply too many treatment groups to manage individually.
4. Focus on displays that are required for the regulatory reporting process and that assist with signal identification.

CONCLUSION

Pooling data from numerous studies is not as easy as it seems. There are a number of points to consider when doing so. Some of the key considerations are described in this paper but company and study specific needs will inevitably give rise to other considerations.

PhUSE 2006

Thorough and precise documentation is essential for mapping data between standards and to assist future development. It is important to remember that ISO reporting is not a one-off effort but is ongoing and repeatable.

There is an emphasis on writing dynamic code, where possible, to automate the integration of recently completed studies. However, it is necessary to weigh up the benefits of writing dynamic code against the ease of future maintenance and legibility. One can detract from the other.

Due to the repeat nature of ISO reporting, it makes sense to keep programs modular, to facilitate manageability. This is preferable to having e.g. one program that attempts to map old datasets as well as integrate datasets across studies.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Ryan DeCosta
Satellite
Artemis House
Odyssey Business Park
South Ruislip HA4 6QF

ryan.m.decosta@gsk.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.