Paper DH09

May Your Data Pool Become Super Cool

Daniil Shliakhov, Intego Group, Kharkiv, Ukraine
Anastasiia Tiurdo, Intego Group, Kharkiv, Ukraine

ABSTRACT

Regulatory authorities require both ISS and ISE submissions for new drug applications. The idea of this analysis is quite simple: pool data from a number of studies and provide the statistical analysis - which is most likely to have already been done for each of the studies - in order to receive more accurate statistical results. At first glance, it does not sound like a difficult task, so the timelines for the ISS and ISE are usually tough. However, the process of pooled data harmonization needs a thorough and careful approach. This paper will describe some steps that may help to understand the process of aggregated dataset creation, as well as to show some tips and tricks to deal with possible difficulties. It contains general technical aspects of the pooled data handling to make it ready for analysis. In addition, the paper will describe a macro which combines SDTM datasets for ISS and ISE together.

INTRODUCTION

ISS and ISE mean Integrated Summary of Safety and Integrated Summary of Effectiveness, respectively. Both ISS and ISE are integrated documents that are created in order to describe the results of more than one clinical studies. The data from all studies (from Phase 1 to Phase 3), performed on the investigated product, are generally combined into one database usually called pooling, so the results are statistically analyzed as a whole.

Usually sponsors are required to summarize the safety information from all clinical studies for any new drug submission. The main idea of ISS analysis is to detect safety signals that may not be detected in individual studies. Minimum requirements for ISS include: summary of the clinical studies, treatment exposure information, demographics and background characteristics, adverse events summary and clinical laboratory assessments details.

There can also be an ISE analysis requested and, obviously, it is more complex than ISS. It makes sense to carry it out since pooling database is much larger than the individual studies and has better chance of detecting statistically significant difference between the treatment groups. Minimum ISE content is primary endpoints summary, which is completely study-specific.

COMBINE RAW, SDTM OR ADAM DATASETS? PROC AND CONS

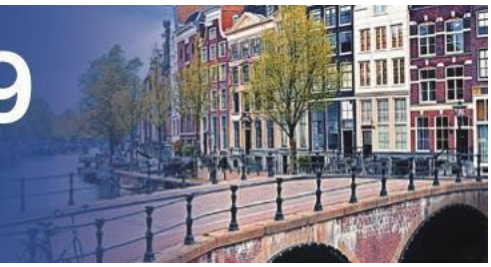
Before creating any summary outputs, the way of combining data from different clinical studies should be chosen. There are 3 possible ways to proceed: either to combine the raw data, or SDTM datasets, or even to aggregate data on the ADaM level from each clinical study. There is no obviously right or wrong way, since lots of points should be considered to make a decision, like the goal of creating the pooling, whether it is just one-time submission or ad-hoc requests are expected. The development of a strategy of data mapping from the study to a pooled level is quite a complex task, especially if the pool includes a large number of studies that are using a variety of data standards. Additionally, legacy data may be integrated as well. However, it needs to be converted to the latest data standard first.

Another important point that can help to choose the way of data pooling is that there may be existing programs available to help in expediting this process. Generally speaking, all the ways have their pros and cons, so let us look at each way and try to find the best solution.

POOLING RAW DATASETS

The most complex and the most inconvenient way of creating the integrated database for ISS and ISE is to pool the raw datasets from each individual study. Since the raw data could differ from study to study and furthermore, the structure of the raw data is likely to be study-dependent, combining the raw data for the ISS/ISE may take a lot of time. If it has been decided to combine raw datasets for the ISS/ISE, the following main points should be considered:

1. It is necessary to remember that the structure of the raw datasets could vary from study to study, so the names of variables and even the names of datasets could be different.



2. Some data could be collected as numeric (for example, gender may have values 1 and 2 for male and female respectively) in one study and similar information could be saved as character variable in another study (for example, sex would have values “MALE” and “FEMALE”).
3. There is a big chance that all collected data in raw datasets will not be straight-forward enough to produce SDTM compliant datasets easily based on it, so it would take lots of extra time to create pooled SDTM datasets based on pooled raw datasets.

However, pooling from the raw data may be a suitable option in cases where there is significant variability in the data structures, standards or endpoint derivations across studies. Although this method assumes lots of time and programming resources, one distinct advantage of this method is that if endpoints, baseline definitions etc. are recreated, this will maximize the consistency across studies.

The scheme of preparing the ISS/ISE using the raw datasets could be found below:

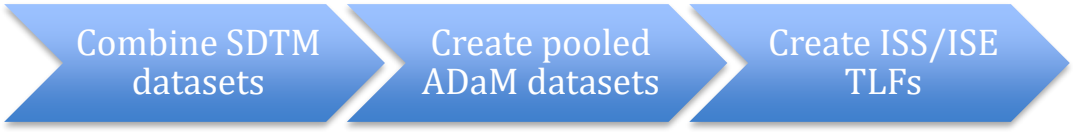
POOLING SDTM DATASETS

ISS and ISE can be built by pooling the individual study SDTMs too. Then integrated SDTMs could be used to create integrated analysis datasets (ADaM). This is the way we recommend to use for pooling datasets for the ISS/ISE analysis. The main advantages of this method are:

1. It’s easy to combine data on the SDTM level rather than on any other level, since the structure of each SDTM dataset is strictly defined by the CDISC Implementation Guide, so at the end there is no or almost no difference in the structure of each particular dataset across all individual studies.
2. Usually SDTM datasets are validated and all work around the data issues has already been done for each individual study.

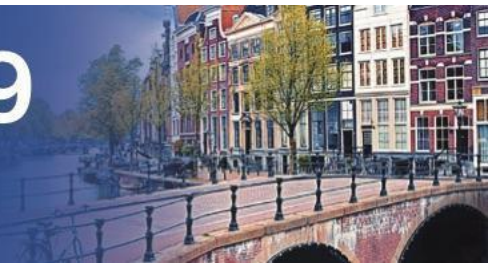
Some guidelines on how to combine the SDTM datasets for the ISS and ISE are provided below.

There are 2 possible solutions on how to receive pooled ADAM datasets. The first one is to combine the SDTMs from each individual study right in the ADAM program, so no pooled SDTM package is assumed to be prepared in such a scenario. And the second one, which is more appropriate, from our point of view, is to create pooled SDTMs first and only after that to prepare pooled ADaM datasets using SDTMs combined in advance. The reason of using such a solution is very simple: statistical programmers may forget to add some studies by mistake, especially if ISS/ISE is based on a big number of clinical studies. Another important point is that some SDTM datasets can have different length of variables which may cause unpredictable results and warnings. According to the FDA requirements the length of each variable should be equal to the maximum length of its value in each particular SDTM dataset. So, before combining SDTMs, a programmer should identify the maximum length and re-set the length of each character variable in a final combined dataset. A quite simple but very useful macro to combine the SDTMs for the pooling could be found in the appendix of this paper.



POOLING ADAM DATASETS

Another approach to create integrated summary is by pooling ADaM datasets from individual studies. From the first glance this approach may take less time than pooling SDTM or RAW datasets. But before going with this, every programmer should be aware of the next important points:



1. All ADaM datasets from an individual study may contain variables with the same name but with the different content. For instance, analysis flag ANL01FL in the ADVS dataset could define the last non-missing observation within the visit window in one study but may flag the worst value per visit in another study. Since the name and label of ANL01FL variable does not contain any meaningful information, programmers should check its derivation in ADAM specification for each individual study, then define the single algorithm to derive the flag in the pooled database. This would finally mean that some ADAMs from individual studies are required to be updated respectively before pooling.
2. Special care should be taken to ensure that all derivations and algorithms are consistent across individual studies. At least definition of populations, visit windowing and the key study parameters should be checked.
3. Sometimes programmers include only tests required for analysis to the final ADaM dataset. Let us look at one simple example. Initially ALT and AST laboratory results are collected but not analyzed in Study A. So, statistical programmers made a decision not to add ALT and AST test to the final ADLB dataset, but kept them in the SDTM LB dataset only. At the same time Study B assumes ALT and AST parameters to be analyzed, so ADLB for this study has both above mentioned tests included. If we are going to combine the analysis datasets only, all ALT and AST results will not be presented in the combined ADLB dataset without additional manipulation. In such situation they should be added manually from the individual SDTM LB dataset and, moreover, all analysis flags, grades, etc. should be re-derived for those two tests again.

Nevertheless, the advantage of this method is in particular relevance for pooling of efficacy data which often includes a large number of derived endpoints, so no additional derivations are necessary in such case.

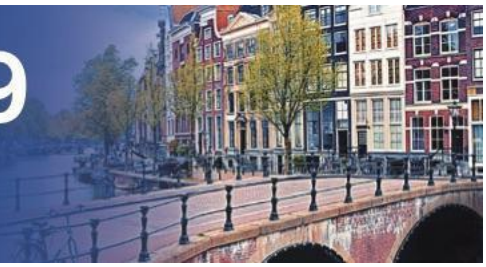
MEDDRA, CTCAE AND WHO

Sometimes dictionaries like MedDRA, CTCAE and WHODrug may cause some issues during the data pooling for the ISS.

According to the FDA requirements “to avoid potential confusion or incorrect results, the preparation of the adverse event dataset for the ISS should include MedDRA terms from the most current version of MedDRA at the time that data across studies are pooled”. Since MedDRA dictionary is updated twice a year there is a good chance that MedDRA version is not the same for all individual studies included in the pooling. For example, if a clinical study that started in 2017 and coded with MedDRA version 20.1 is going to be integrated to the ISS analysis that is planned to be submitted at the beginning of 2020, then ISS should be coded using MedDRA version 22.1.

The reason for ISS to be based on the latest version of MedDRA is that the reviewers often want to analyze AEs across studies looking at the Adverse Events of Special Interest which are based on SMQs (Standardized MedDRA Queries). The reviewers are likely to use the SMQs related to the latest MedDRA version, so to avoid any confusion and to make their job easier programmers are recommended to follow the last available coding version.
Reconciling the MedDRA version should be carried out by coding specialists before data is pooled for ISS. If the version used for an individual study is different from the version planned to be used for the ISS dataset, a document describing any changes to coded variables should be provided to the regulatory authorities. It should help the reviewers to understand any potential differences identified during the review of the pooled data against the individual study data.
There is only one exception not to update MedDRA version: if the results from legacy studies are included in the ISS submission, there is no need to update the version of such individual studies.

WHODrug (World Health Organization Drug Global) is a dictionary typically used to code concomitant medications data. WHODrug contains unique product codes for identifying drug names and listing medicinal product information including active ingredients and therapeutic uses. Similar to the Adverse Events coding process, concomitant medications should be coded to the single version of WHODrug.



The National Cancer Institute (NCI) Common Terminology Criteria for Adverse Events (CTCAE) are a set of criteria for recognition and grading toxicity of Adverse Events. While for adverse events CTCAE toxicity grading is directly collected from the sites on the CRF, for laboratory results toxicity grading is assigned by statistical programmers in the later stage, namely, during the ADLB development. The example of the Hemoglobin grading is presented below:

Blood and lymphatic system disorders					
CTCAE Term	**Grade 1**	**Grade 2**	**Grade 3**	**Grade 4**	**Grade 5**
Anemia	Hemoglobin (Hgb) <LLN - 10.0 g/dL; <LLN - 6.2 mmol/L; <LLN - 100 g/L	Hgb <10.0 - 8.0 g/dL; <6.2 - 4.9 mmol/L; <100 - 80g/L	Hgb <8.0 g/dL; <4.9 mmol/L; <80 g/L; transfusion indicated	Life-threatening consequences; urgent intervention indicated	Death

Usually the calculation of CTCAE grading for laboratory results requires an individual derivation for each lab test.

If it was decided to combine ADaM datasets from single studies without SDTMs or RAW data pool prepared, it is very important to be careful with the ADLB variables containing toxicity grades, as different studies could use different versions of CTCAE. ISS should use the latest single CTCAE version, so all grades should be recalculated in the final pooled dataset.

LEGACY STUDIES

Sometimes legacy studies should be a part of ISS. What are the legacy studies? These are prior studies conducted for a drug currently being studied. These studies can be in a separate indication or conducted by another sponsor before the drug is acquired by current sponsor who is going to submit the ISS. Per FDA guidance on format and content of ISS the sponsor should “integrate safety information from all sources, including pertinent animal data, clinical pharmacology studies, controlled and uncontrolled clinical trials (including controlled trials for indications not claimed in the application) and foreign marketing experience or epidemiologic studies related to any use of drug”.

If legacy studies are part of the ISS the following points should be considered:

1) Studies may have been conducted many years ago and/or may span across several years. This means that there is a big chance that data will not be standardized using CDISC.
2) Studies may have been developed by multiple companies and as a result there may be a lack of consistency across studies, the different method of data collection may be used, etc.
3) Such studies may not reflect current regulatory requirements, because these studies are pre-ICH.

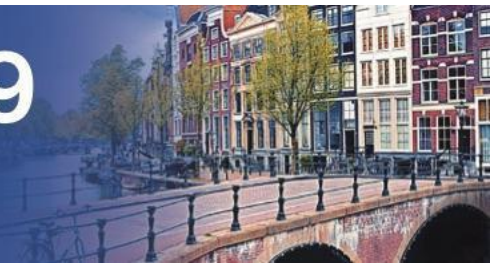
So, considering all these points, including legacy studies in ISS may cause big problems for the statistical programmers, since they should create SDTM datasets for such studies to standardize the data. There is a big chance that for legacy studies the ADaM datasets will not be presented at all as well as some study documents and derivation rules could me missing.

DATA HARMONIZATION

Probably the most annoying and time consuming process related to ISS/ISE analysis is data harmonization starting from the attributes unification and up to the use of the same derivation algorithms for all studies in the pooling. This section provides some recommendations regarding the challenging points that should be deeply checked in order to receive analysis ready integrated data.

Talking about the attributes, if all the individual studies are prepared according to the CDISC standards and the data is pooled on the SDTM level, as it was suggested above, then only minimal check of attributes is necessary. The main thing we recommend is to set length to 200 characters (as maximum possible) for all variables that are going to be pooled, otherwise some cut values may appear. As for the other attributes, there is no special manipulations to be done, since SDTM structure is quite strict. All the variables that are allowed to be included in each dataset are described in the Implementation Guide in advance, so just Pinnacle 21 CDISC compliance check should be sufficient on the ISS/ISE level.
Regarding harmonization of the variables values, it makes sense to apply the steps suggested below.

First of all, it is a good practice to put all the character data to uppercase (considering the control terminology, surely) to avoid any mixed case issues, especially for the ‘free’ text value. In general, if there is no terminology assumed to be applied to the variable values but the variable is available in more than one individual study, we would advise to create the list of all possible values right after all the data is combined and look it through. It should help to identify the



same values in terms of meaning, but they are not identical, in fact, due to typos, different word order or just rephrasing.
For instance, all the following reasons of death can be remapped to the unique value:

Death due to Adverse Event
Adverse Event Leading to Death
Death due to AE

→ Adverse Event

Considering Laboratory, Vital Signs, ECG data in the integrated analysis, special attention should be paid to the fact that most likely a number of lab vendors have been used across the pooled studies. Moreover, the sensitivity and calibration of the equipment may vary. This would definitely influence the normal ranges and may lead to the limited interpretability of the final statistical result. If the analysis based on the normal ranges is going to be provided, then it might be worthwhile to standardize ranges and the way of standardizing is to be described in the SAP.

Another key point related to findings data is to ensure that all values are converted to and reported in the standard units. The problem is that the standard units may vary depending on the region, so, theoretically, it can be a combination of SI and US units for different studies. That is why, when pooling the laboratory data, confirmation of the units used is necessary and additional adjustment might be required.

As for the test names, they are less likely to become a problem in the integrated analysis, since in most cases parameters should be defined based on the control terminology. But what really can cause the issue is parameters from Supplemental domains, since there are no requirements as for them, so cross-check of all the QNAM/QVAL values available in each particular domain is strongly recommended.

Additionally, the important thing to take into account is the way of how the partial dates were imputed, since for each particular dataset the method may be study-dependent. We particularly advise to check Adverse Events start/end dates imputation. Usually it is not really straight-forward and with a good chance is based on the worst-case scenario, meaning that the imputation will be done with the idea that AE occurred right after the start of the study treatment. This, in particular, would affect treatment emergent flag, and as a result the number of events included in the safety summary – one of the key ISS outputs. In terms of treatment emergent flag, its algorithm should be checked across all the studies, since for some individual studies it may consider toxicity grade worsening, while for the others initial toxicity would not be available at all. The same thing needs to be checked for the concomitant medications data to ensure prior/concomitant categorization, which depends on the date imputation as well, is done properly.

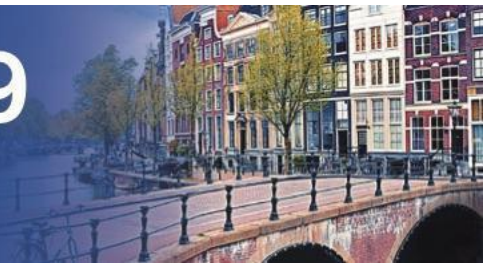
Another tricky question is related to the sites and patients in general. At the study level it is relatively simple to identify unique sites, but once data is pooled this may not be such a trivial task. It is unlikely that Site 001 in Study A is the same as Site 001 in Study B. A unique site number is needed to identify the same site in various studies if site effects are to be analyzed with the pooled data. The same idea is for the situation when the same patient participated in more than one of the pooled studies. The unique ID should be assigned to them on the pool level and the decision should be made as for treatment start definition, treatment group to be used and so on.

To sum up, how is the integrated structure determined? For variables that are common across the pooled studies:

1) Determine the most common definition of the variable.
2) If any imputation is assumed, define the single way of how to proceed.
3) Make variables in all studies consistent with the “majority” (for example, Height is in meters in 6 out of 8 studies. Therefore, convert height from centimeters to meters in the other two studies).
4) Map code lists of categorical or ordinal variables into a common standard if necessary.

VISITS AND ANALYSIS WINDOWING RELATED ISSUES

Special attention should be paid to visits and analysis windowing development during ISS/ISE creation. In case of big number of studies included in the pool this topic may be quite problematic. Talking about SDTM it should make sense to review all the findings datasets in order to have the name of visits and visit numbering consistent between the studies. For example, a spelling of the follow-up visit may vary as ‘Follow-up’, ‘Follow-Up’, ‘Follow-up Visit 10’, ‘Follow-up, Day 28’ and so while, in fact, they all mean the same visit. Logically it should be updated to one unique value across all the above cases.



In addition, even if the names of visits are identical across all the individual studies, the following situation may happen to visit numbering. Let us assume that in one of the pooled studies patients are scheduled to visit sites and have all the assessments once per 3 weeks, while in the second study the scheduled visits are planned once per 4 weeks. Most likely visiting details would look like the ones presented below as well as the suggestion on how it should be organized in the pool.

Visit name	Visit Number		
	Study A	Study B	Pooling
Screening	-1	-1	-1
Day 1	1	1	1
Week 3	2	NA	2
Week 4	NA	2	3
Week 6	3	NA	4
Week 8	NA	3	5
Week 9	4	NA	6
Week 12	5	4	7

To resolve this issue, our suggestions are as follows:

1) Select all the unique values of study ID/ visit name/ visit number in the pooling.
2) Look through all the individual study protocols in terms of schedule of assessments. It might be the case that visit names are not informative enough, so it is not really clear what the visit consequence should be without checking the study documentation.
3) Unify the names of the visits. It makes sense to put all the values to uppercase first and then use information found in the protocol.
4) As soon as all the visits details are clearly defined, create a separate dataset or excel spreadsheet (whatever is more convenient in your case), with the whole list of visits in the pool database. Define a logical order for all the visits as it is shown in the table above.
5) To distinguish between visit numbers, you can use non-integer values. For example, if “Week 3” corresponds to the visit number 3, then the “Week 3 Day 2” can be assigned to 3.02.
6) If unscheduled visits are included in the analysis (it is not necessary that this is the case for the pool), they should be numbered as more logical as possible. For instance, the first Unscheduled visit that occurred after Week 8 (VISITNUM = 3) should ideally be numbered as 3.1, the second – 3.2 and so on. Otherwise, if it would take too much time, just assign all the unscheduled visits to the one visit number to avoid any confusion after remapping, especially if unscheduled visits are not going to be analyzed, it is not necessary to waste time on that.
7) If study period/phase is included in the name of the visit, they should not be ignored. For example, ‘Double Blind Period, Week 2’ and ‘Open Label Extension, Week 2’ are not the same visits. Possible solution is to use numbers as 1000.2 and 2000.2, respectively.

The same idea of consistency between the individual studies should be applied to the time points values. The thing that needs to be carefully defined is Pre-dose time-point, since different protocols may define pre-dose assessments in different ways: 1 hour / 30 minutes / 10 minutes before dosing or just without any details. In such a situation the decision should be made if this is meaningful to present all pre-dose assessments under one single pre-dose time point. Otherwise, if there is something in the study design which means that a pre-dose value at a given visit is influenced by the previous dose, these values may not be directly comparable, so should be named differently.

While working with analysis windows on the ADaM stage, programmers need to consider all the individual study details, especially for the visits related to key study points. For example, it may happen that two individual studies have the key visit “Week 12”, that is defined as

	Week 12 Analysis window	Week 12 Target Day
Study A (visiting schedule is once per week)	[78 ; 84]	81
Study B (visiting schedule is once per 3 weeks)	[74 ; 95]	84

Theoretically, they should correspond to the same visit, but in practice they could represent absolutely different results. That is why the most optimal and unique analysis windowing definition across all the pooled studies should be found and described in the SAP.
Furthermore, even providing that analysis windows are defined exactly in the same way, protocols may address different algorithm of finding analysis value per window. Of course, it should be within the defined limits and as closer to the target day as possible, but in case of 2 results available on the same distance from the target day, one protocol may request to use the earliest one, while the second protocol requires the latest one.

It is necessary to note that especially if a big number of studies is pooled, it may not be necessary to work on all the visits data. For instance, if there are some visits/time points that are available for only one or two of twenty individual studies included in the pooling, it would be logical to exclude them from analysis at all since the results will be meaningless compared to other visits anyway. It does not make sense to spend time on the harmonization of such visits, but it should be agreed with the study team in advance, as well as whether unscheduled visits results are going to be included in the integrated analysis, since their exclusion may save lots of time too.

Talking about the visits that are reasonable to be reported, first of all, these are the visits that intersect across all the studies. Namely, if in some studies scheduled assessments are available once per 3 weeks up to week 24 and in the rest of the studies – once per 4 weeks, it is obvious to include visits Baseline, Week 12 and Week 24 to the final summary only. But the situation may not be so clear and then there might be a necessity to derive the endpoints that utilize all data e.g. minimum or maximum post baseline value or change from baseline, or just the latest one in the treatment period. This is particularly the case for laboratory data analysis – that is always a part of ISS. To avoid this routine work around analysis visits harmonization, shift table of laboratory abnormalities (baseline vs the worst on-treatment value) is requested rather than by visit summary.

The last but not the least point to discuss is Baseline definition. It is possible that baseline flags in individual studies are not derived consistently, especially when complicated study designs are involved to ISS/ISE. Another tricky case is when one subject participates in multiple studies. The most obvious things to check are:

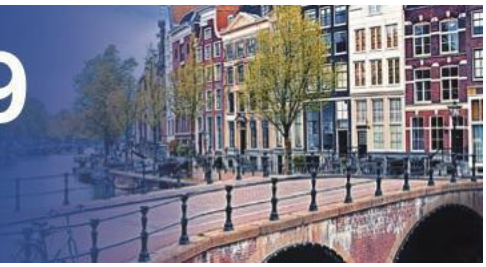
- Whether baseline value should be found strictly before the first treatment dose or before/on treatment start.
- If the study day 1 results are to be considered as baseline in case time is not available.
- the type of baseline may vary from the last to worst or average pre-treatment value – to be unified as well.

TREATMENT GROUPS ASSIGNMENT

One of the tricky topics related to ISS/ISE creation is a treatment group assignment. If you need to combine the data from studies that are of similar design i.e. dose, duration, control and population, you are a lucky one. In fact, it is often the case for pooled data that treatments are presented differently in individual studies, so it is not really obvious how to combine the patients based on them. The matter can be in regimen, dose frequency and value, or exploratory studies may be grouped into an “other active treatment” type group. The potential questions to be raised whether patients that dosed BID should be combined with the patients that dosed QD, if their total daily dose was the same? Or if the patients that were allowed to titrate their dose can be combined with patients that followed a specific dosing regimen?

Of course, all the answers should be well discussed in advance with the whole ISS/ISE team. Here we would like to summarize some points that should be taken into account during the treatment groups assignment:

- Special attention should be paid for studies with multiple periods of treatment as well as for the cross-over designed studies. It can be the case that only some of the periods will make sense to be analyzed in the pooling, so all the data should be cut-off accordingly. Otherwise, patient may be assigned to each dose group to which they were exposed to and this solution is far from being perfect, since the same patient would be counted multiple times in the denominators of the ISS treatment groups, as a result, the treatment groups are not independent.



- One more challenging case is the studies with up/down titrations – it should be also agreed on how to proceed. The good thing is that this is a typical situation for Phase I studies and in the most cases, short-term studies in healthy volunteers are not combined with Phase II and Phase III studies in patients, so such a scenario is rare to take place.
- Add-on therapies should be taken into account as well, since they easily can influence the treatment effect.
- In case there are really lots of different treatment types to be combined, then the easiest possible grouping would be: placebo-controlled studies vs. uncontrolled or active-controlled studies. In general, it is a common practice to have an "all dose groups" category combining various doses of the investigational drug in the pooling summary, so theoretically only such kinds of data can be sufficient at the end.

Finally, the most important thing is that the required treatment groups mapping is well documented, since most likely it will not be clear just from the treatment group names or output headers as received. A good idea would be to include respective footnotes to all the summaries explaining the way treatment groups were formed.

POOLING STRATEGY FROM THE PROGRAMMING SIDE

"Forewarned is forearmed" – the phrase that plays a major role in the planning of any integrated analysis. To be able to generate ISS/ISE analysis with the minimum time/programming resources, the pooling strategy should be developed as early as possible and should clearly explain the purpose, objectives and requirements of the pooled analysis. The plan is to be prepared to define which studies and which data should be pooled to add value to the integrated analysis for any potential deliverables from both efficacy and safety perspectives.
If at the time of the individual study analysis the project team is already aware that in the future this study is planned to be included to the integrated analysis, it is a good chance to organize all the individual studies in the similar manner, especially in terms of key algorithms. Anyway, even if there is no information available as for the future plans, one of the crucial things that should make any pooling process as easy as possible is that data standards should be used within projects for each study to enable more efficient data combining if it would be requested later.

A good practice also would be to make all the codes for individual studies universal, optimal and well organized, so they can be easily used on the pooling level. Especially this is important for the outputs programs and the minimum point is that they should not depend on the number of patients / studies / treatment groups and so on.

The last thing to discuss is that any ISS/ISE analysis assumes work with the huge number of data, for instance, pooled laboratory dataset can easily contain tens of millions of records. In order to make your codes run faster on such data it is necessary to prepare the programs for this process. If you are using SAS® environment to prepare the integrated analyses our advices are the following:

- Avoid sorting datasets or re-setting if this is not needed.
- Make use of as few data steps as possible, for instance, use OUTPUT statement instead of several data steps if you need to create several datasets from the same input dataset based on different conditions.
- Make use of DROP or KEEP statement to remove un-needed variables even in temporary datasets.
- Make use of WHERE statement while setting instead of making record selection in the data steps.
- Drop temporary datasets when they are not needed anymore.
- Carefully reduce LENGTH of the variables.
- Use indexes or hash-objects where possible.

What is also good to remember is that there is a big chance that individual study and ISS/ISE programming teams will not be the same, so other colleagues may take over your code at the end. That is why do not make it too complex, it has to be commented and understandable - this aspect can potentially save even more time than any above mentioned point.

The next no-less important part of ISS/ISE programming is validation process. It can be done in the standard way via double programming, especially if some QC programs have already been prepared on individual study level. Additionally, once data is pooled it is important to ensure that the pooled data represent the data from individual studies consistently where expected. For instance, obviously, pooling from raw data will result in reprogramming of the same endpoints produced for individual studies before. In a such case, checks on consistency are strongly recommended to ensure that re-derived results are the same as the ones received on each individual study level. If there are any discrepancies due to known differences in algorithms, they have to be documented if further questions regarding inconsistencies between the integrated analysis report and the individual summaries are raised.

CONCLUSION

To conclude, it is, of course, ideal to plan ISS/ISE analysis when designing the individual studies, but, in any case, the planning process should be started as soon as the decision to submit integrated data is reached. Pooled datasets can be created from the existing raw or derived datasets and the decision to use one or other will depend on the availability of standard and derived data, common endpoints, etc. When preparing pooling strategy, it is important to understand the rationale of pooling data and the purpose it is created for. The default should not be to pool everything but to do so when it adds value. Even with a well-defined data standard in place a certain amount of variability is to be expected. Detailed documentation is important to be prepared as well as the validation that may be based on reports or by study checks.

CONTACT INFORMATION

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

Daniil Shliakhov and Anastasiia Tiurdo
Intego Group
Gagarin avenue, 43/2
Kharkiv, Ukraine / 61001
daniil.shlyakhov@intego-group.com, anastasiia.tiurdo@intego-group.com

APPENDIX 1. MACRO TO COMBAINE THE SDTM DATASETS

```
%macro change_v_length(
                        lbin=
                       ,dsin=
                       ,lbout=
                       ,dsout=
                       );

proc sql noprint;
       select name into :names separated by " "
       from dictionary.columns
       where libname="&lbin" and memname="&dsin" and type='char';
quit;

proc sql noprint;
       select name into :names_retain separated by " "
       from dictionary.columns
       where libname="&lbin" and memname="&dsin" ;
quit;

data &lbout..&dsout;
       retain &names_retain;
       length &names $200.;
       set &lbin..&dsin;
run;
%mend change_v_length;

%macro pool_sdtm(
                  list_inlibs=
                 ,list_inds=
                 ,list_outlibs=
                 ,out_lib=
                 );

data _null_;
       numoflibs = countw("&list_inlibs");
       call symputx("numoflibs", numoflibs);

       numofds = countw("&list_inds");
       call symputx("numofds", numofds);
run;

%do lib=1 %to &numoflibs;
       %do ds=1 %to &numofds;
              %change_v_length(lbin=%scan(&list_inlibs,&lib,' '),
                               dsin=%scan(&list_inds,&ds,' '),
                               lbout=WORK,
                               dsout=%scan(&list_inds,&ds,' ')&ds);
       %end;
%end;
%mend pool_sdtm;
```