Paper SI09

The Integration Dilemma

Angelo Tinazzi, Cytel Inc., Geneva, Switzerland

ABSTRACT

As of today, our Industry has not defined any approach, nor does an official regulatory agency preference / recommendation exists on how to integrate data of different studies to support either ISS/ISE.

In 2020 PHUSE released a white paper, "Integration Strategies in Support of ISS/ISE Submissions" [2], where three integration approaches are proposed:

1. SDTM(s) → iADaM
2. ADaM(s) → iADaM
3. SDTM(s) → iSDTM → iADaM

Over the last five years at Cytel we did experience all three options for several submissions. All three options have pros/cons and their applicability strictly depend on whether and how standards were applied in individual studies.

The purpose of this presentation is to share experience we gathered by applying these three options, such as challenges and some technical tips we did apply to streamline for example the integration of several legacy studies into an iSDTM.

INTRODUCTION

PURPOSE OF ISS/ISE

The Integrated Summary of Safety is required by the U.S. Food and Drug Administration (FDA) as a critical component in any New Drug Application (NDA) or similar market approval requests. Integrated Summary Efficacy (ISE) might be also required under certain circumstances, although most of the time efficacy results from individual pivotal study, or studies, might be sufficient. This is not simply a summary but rather a detailed integrated analysis of all relevant data from the clinical study reports with the aim of providing a more transparent understanding of responses across different populations (demographics, disease related, etc.) and dosing regimens.

Both ISS and ISE allow reviewers to easily compare individual outcomes, tracking subject's results across the entire clinical development, facilitating broad views of the investigational product's overall efficacy and safety profile. With ISS and ISE, a "single database" is formed by pooling the results of all concerned clinical trials.

PLANNING FOR THE ISS/ISE

With the draft label created, the intended key messages documented, and the pivotal studies designed and recruiting, the integration/pooling of the data requires careful consideration. Typically, you may have multiple studies to consider for supporting your safety and/or efficacy claims.

At this stage, it is invaluable to engage with a well-seasoned statistician who will collate the legacy study designs and assist with your initial pooling plan, considering the following topics:

- patient populations and cohort e.g., dosing groups
- dosing appropriateness
- regimens and trial durations
- evaluation of safety and efficacy in various subgroups
- evaluation of secondary efficacy endpoints, which were underpowered in individual studies
- impact of concomitant medications safety and efficacy
- in general, getting a more robust assessment of safety in subgroups if sample size is sufficiently large. Safety assessment might include assessment of laboratory data, ECG, and vital signs
- long term effect of the product or chronic conditions
- devise the pools and related justifications for those

It is then time to share those plans and meet with the authorities, obtain guidance, and fine tune the pooling plan; from these meetings you can get an agreement from the FDA about the proposed pooled strategy.

Focus then turns to the data and it should be ensured that the ongoing pivotal studies are reported and delivered in compliance with data standards requirements i.e., CDISC standards (SDTM and ADaM) as well as ICH E3 compliant study reports (CSRs). This will save time later when the data integrations to support the pooled data analysis are required.

Whilst the pivotal studies are running, it is an ideal time to finalize the ISE and ISS statistical analysis plans. There is also time to assess legacy studies datasets (non in standard format) and CDISC packages from closed studies to be

incorporated in the submission and plan the necessary conversion to applicable standards. It is important to ensure that the electronic Common Technical Document (eCTD) required submission documents are inclusive, conclusive, robust and the analyses results production has full traceability from CRF data through to results.
It may be necessary to retrospectively create CDISC SDTM and ADaM entities and validate any utilized endpoints within them against the existing CSRs to ensure consistency.
Each legacy study will have its own issues and it is imperative that a

gap analysis plan is devised and deployed against each study to ensure that there is absolute data traceability and accuracy.

There are several submission requirements that surround the data elements, it is important that your chosen 'Data experts' are aware of these and avoid the simplest of pitfalls. For example, but not limited to:

- Coding of the pooled data (unique terms) with the latest dictionary version
- Study level domains that are coded should retain the reported coding and version
- Reviewers guides and defines.xml must be prepared and complete

With all your supporting studies polished and their respective dossier contributions prepared, the Pivotal studies will be ending. Your ISS, and ISE, SAPs will be finalised (or close to!) and statistical programming should be progressing using available data from closed studies, either in standard or legacy format, and pivotal test or draft data as required. The clinical overview should be well advanced and drafted template ISE/ISS reports will be set out and have the key message sections ready to receive the pooled outcomes.
With careful planning and front-loaded preparations, the final pivotal study and final pooled analysis can be generated and reported.

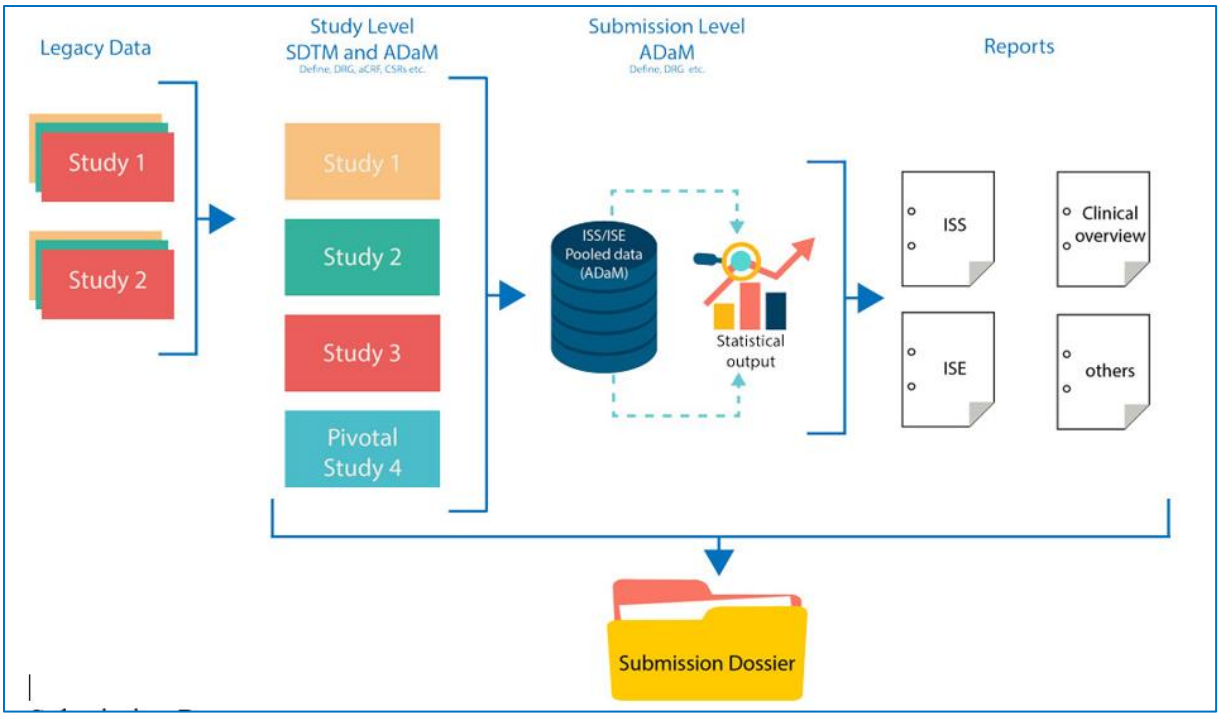


Figure 1: Overall Plan for ISS/ISE

DATA INTEGRATION OPTIONS

As of today, our Industry has not defined any approach, nor does an official regulatory agency preference / recommendation exists, on how to integrate data of different studies to support either an Integrated Summary of Safety (ISS) or an Integrated Summary of Efficacy (ISE). When it comes to data integration, neither the SDTM IG nor the ADaM IG do not give very much guidance, since both currently cover single study data structure.
An "attempt" was made by the CDISC ADaM Team in 2018 [3] and in 2019 [4] with a CDISC webinar, with the release of a draft version of "ADaM Structure for Integration"; unfortunately, that new set of ADaM standards was never confirmed and released. In any case, that guidance was only covering one side of the problem, that is in the integration of the data in ADaM for analysis purpose (that is somehow a requirement), while there is much discussion about the need for Integrated SDTM (iSDTM) and how to create these versus not needing iSDTM.
It was October 2020, PHUSE released a white paper, "Integration Strategies in Support of ISS/ISE Submissions", where three integration approaches are proposed with pros and cons discussed:

1. SDTM(s) → iADaM: ADAM Integration only Using Individual Studies SDTM datasets
2. ADaM(s) → iADaM: ADAM Integration only Using Individual Studies ADAM datasets
3. SDTM(s) → iSDTM → iADaM: SDTM and ADAM Integration

The whitepaper was followed by a webinar [5], with panelists including two FDA representatives, where the three integration approaches were discussed.

All three integration options require the creation of an iADaM from which integrated analyses are generated. While in option 1, iADaM is generated from individual study SDTM(s) and in option 2 individual study ADaM(s) are used, option 3 requires an intermediate set of SDTM datasets (iSDTM) to be created from individual study SDTM(s), from which iADaM is then generated. In all three scenarios, it is also accepted to integrate either legacy raw datasets or legacy analysis datasets directly into either iSDTM or iADaM for studies started before or on December 17, 2016. This reduces the effort and budget needed when dealing with submissions with several old studies yet to be integrated.

WHICH DATA INTEGRATION OPTION SHOULD YOU CHOOSE?

During the PHUSE webinar the participants were asked to reveal which approach they commonly used for Integrating their studies, surprisingly (to me), their top preference was option 2 (53%), followed by option 3 (32%); option 2 also appeared to be the preferred choice of the two FDA representatives although it was recognized that option 3 too is a valid option under certain circumstances, provided that traceability is guaranteed.

In all cases, rationale for the approach used should be mentioned in the reviewer guide, either in the Integrated Clinical Study Data Reviewer Guide (iSDRG) or in the Integrated Analysis Data Reviewer Guide (iADRG). Most importantly, it should be discussed when meeting the relevant agency prior to pre-NDA/BLA meeting, for example, during a type-C meeting and planned integration strategy can be anticipated in the Study Data Standardization Plan – SDSP [6]. Figure 2 and 3, show two examples of SDSP "Pooled Studies" section where the integration strategy is shared with the FDA, adopting respectively option 2 and 3.

4.3 Pooled Studies

Both Efficacy (ISE) and Safety (ISS) will include data collected in the listed studies in the two following tables. Integrated safety ADaM dataset (iADaM) will be created from individual study ADaM datasets and included in the submission. Data from individual SDTM study packages might be integrated in the iADaM when an information is not available in the individual study ADaM datasets.

The following approach will be used:

- Differences in versions used for CDISC Standard Controlled Terminology and Medical Dictionaries, such as the MedDRA, will be aligned in the iADaM, meaning that all Adverse Events, for example, will be coded using a single version of MedDRA (original medical coding will be also kept in the iADaM); any other mapping discrepancies between studies will be fixed in the iADaM. Consequently, the original individual study ADaM datasets will be not modified.
- Derivations used in individual studies will be checked and eventually modified in the iADaM
- define-xml and the Analysis Data Reviewer Guide (ADRG) will be provided together with the iADaM datasets package; the define-xml and the ADRG will provide details of any major applied harmonization

More details about standards <Sponsor> intends to use i.e., ADaM Ig, Controlled Terminology and MedDRA versions, are detailed in the table on the next page.

Data Pool Identifier	Data Pool (List of Studies)	Pool Status	Pool Description	Exchange Standards	Terminology Standards
ISS	Study-1 Study-2 Study-3 Study-4 Study-5	PLANNED	Integrated Summary of Safety	ADaM v2.1/ ADaM IG 1.1 ADaM define.xml 2.0	CDISC ADaM Terminology TBD MedDRA (Adverse Events/ Medical History): 25.1 WHO-DD (Medications) B3 March 2022
ISE	Study-1 Study-2 Study-4	PLANNED	Integrated Summary of Efficacy	ADaM v2.1/ ADaM IG 1.1 ADaM define.xml 2.0	CDISC ADaM Terminology TBD

..

Figure 2: Sharing Integration Strategy with the FDA through the SDSP (ADaM(s)→iADaM)

4.3 Pooled Studies

In addition to study data packages for XXXX Phase II/III studies (see section 4.1), <Sponsor> will create pooled Integrated SDTM (iSDTM) datasets for the domains that will be used for the ISS; the iSDTM will be used to create the Integrated ADaM (iADaM), from which ISS summaries will be created e.g., tables. Both iSDTM and iADaM datasets will be submitted together with define-xml and cSDRG and ADRG respectively for iSDTM and iADaM.

The iSDTM will be created from the studies described in the table in section 4.2. In this iSDTM the following approach will be followed:

- Only domains needed for the ISS will be integrated e.g., SDTM domains specific to one study only and for which no integrated analyses are planned will be not integrated in the iSDTM
- Studies for which only Legacy Datasets are available will be converted to SDTM directly into the iSDTM; for these legacy studies only data-domains needed for the ISS analysis will be converted
- Screen Failures subjects' data will not be mapped to the iSDTM; screen failures details can be found in the individual CSR and individual study data packages
- Differences in versions used for CDISC Standard Controlled Terminology and Medical Dictionaries, such as the MedDRA, will be aligned in the iSDTM, meaning that all Adverse Events, for example, will be coded using a single version of MedDRA; any other mapping discrepancies between studies will be fixed in the iSDTM. Consequently, the original individual study datasets will not be modified
- define-xml and the Clinical Study Data Reviewer Guide (cSDRG) will be provided together with the iSDTM datasets package; the define-xml and the cSDRG will provide details of any major applied harmonization
- Based on the current versions of the ISS SAP, at least the SDTM domains listed in the following table will be integrated from all studies contributing to the ISS

<Follow details of pooled SDTM domains that will be provided>

ts.xpt for the iSDTM will be not provided giving the fact pooled datasets will be submitted to eCTD section 5.3.5.3 "Reports of Analyses of Data from More than One Study and ts.xpt is not required for eCTD section 5.3.5.3".

No iSDTM Trial Design Datasets will be provided as well as special purpose domain SV (Subject Visits) and SE (Subject Elements). "Visit Windowing" will be applied in iADaM wherever applicable and needed the ISS planned analysis.

Subjects participating in more than one study will be identified using the same USUBJID, that is the USUBJID assigned in the study the subject was first enrolled in. The following table lists studies concerned by the multiple across studies enrolments:
<Follow list of studies>
Throughout the pooled iSDTM datasets, the STUDYID and the subject / enrolment ID (SUBJID) assigned in each study will be used to distinguish data "pertaining" to each study participation e.g., Double Blind Study data vs Open Label Study Data. For example, in DM, if a subject participated in the 001 Double-blind Study, then enrolled into the 002 Open Label Study, there will be two records with same USUBJD, but different SUBJID, being the original USUBJID assigned in the original study datasets for the study 002, with STUDYID indicating study participation:

STUDYID	USUBJID	SUBJID	RACE	BRTHDTC	AGE	RFIDTC
001	001-001-001	001-001	WHITE	1975-01-01	45	2020-01-01
002	001-001-001	001-015	WHITE	1975-01-01	46	2021-01-01

In all other datasets, the assigned USUBJID in DM will be used, with again STUDYID and SUBJID indicating the original study ID. Information. For example, in the following AE table, subject 001-001-001 had two occurrences of the same event, NAUSEA, one in the double-blind, randomized study, study 001 (record nr. 1), and one in the open-label study, study 002 (record nr. 3).

STUDYID	USUBJID	SUBJID	AETERM	AESTDTC	AEENDTC
001	001-001-001	001-001	NAUSEA	2020-01-02	2020-01-05
001	001-001-001	001-001	ANEMIA	2020-01-12	2020-01-13
002	001-001-001	001-015	NAUSEA	2021-01-04	2021-01-07

More details about standards <Sponsor> intends to use i.e., SDTM Ig, Controlled Terminology and MedDRA versions, are detailed in the table on the next page.
<Follow a table with proposed ISS Pool>

Figure 3: Sharing Integration Strategy with the FDA through the SDSP (SDTMs→iSDTM→iADaM)

As mentioned above, option 2 seems to be the preferred choice of the Pharma Industry Community and the FDA. However, in deciding what option to go,

the decision should be based on data standardization status of the studies we "intend" to use in our ISS/ISE.

If for example, like in the use case 3 later discussed and object of the SDSP in figure 3, we have to pool for ISS purpose several "historical" studies, not only with various SDTM IG and controlled terminology versions, but also with different legacy formats, option 3 then become an obligated choice as it allows "in one shot" the full data integration of all studies already in SDTM or not.

If instead a sponsor was able to full apply the standardization effort early in the life-cycle of its drug development, and as such having all studies mapped and analyzed consistently, thus having also individual study ADaM packages with not only similar terminology but also with consistent derivations and the same can be replicated in the in the ADaM for the ISS with minimal changes, obviously a sponsor will want to save time during the data integration effort and start by pooling together all ADaMs from all concerned studies in the ISS to build the iADaM (see use case 2).

Lastly, option 1 can be considered a valid option when we have a limited number of studies to integrate, all in SDTM format with same or very closed versions, and the ISS/ISE requiring different or slightly different analysis approaches, requiring slightly different way of deriving endpoints, for example considering different periods when analyzing the whole subject exposure when a subject was rolled over into an Open Label study after participating to a double blind study; in such a situation, the sponsor could still opt with option 3 before creating the iADaM if one wants to provide, for example, a sole source for the iADaM and a sole source for the reviewer, after integrating and harmonizing data from all concerned studies (harmonized controlled terminology, one single version of medical dictionary, etc.).

PROS AND CONS OF OPTIONS

The PhUSE with paper, and many other previously discussed sponsor experience (see "Recommend reading" section), discuss the pros and cons of the three options and activities you should consider when choosing one of three possible options.

For example, if you choose option 2, you need, as previously mentioned,

sufficient pre-planning and good scrutiny of the work done where individual studies analyses

especially if conducted by the different CROs as you need to make sure ADaM datasets for individual studies are created consistently. This option in theory guarantees the maximum level of traceability because, not only most of the derivations/transformations were done at the study level but also, we will apply the same derivations, including conventions and eventually any imputation rules, which mean when doing for example an ISS we will get consistent results with what was presented in the individual study CSRs. You however assume here traceability was already done from individual SDTMs to individual ADaMs.

On the other hand, option 3 could be a good choice to go, and it is still my preferred approach, because despite this option will require the reprogramming of endpoints, variables and derivations that were possibly already programmed at a study level, it provides a sole source for the iADaM and for the reviewer. This option however requires some backward re-validation to make sure, once we have harmonized and integrated individual study SDTM into the iSDTM, we do not get different "numbers" from those obtained during the individual CSRs, although this anyway becomes a necessity if the endpoints in the final integration differ from those defined for one or more studies (and may require more time for validation if the derivations are complex) and as such it is an obliged step with whatever option we chose.

Pros / Cons	**Option 1 SDTM(s) →iADaM**	**Option 2 ADaM(s) →iADaM**	**Option 3 SDTM(s)→iSDTM→iADAM**
Alignment of algorithms, controlled terminology, and data domain			
Sole source for the iADaM			
Sole source data for the reviewer			
Cost: additional datasets, define-xml and cSDRG for iSDTM			
Clear Traceability			
Requires consistent analysis approach and terminology			
Can keep original dictionary versions	Standard variables in OCCDS	Standard variables in OCCDS	In SUPP datasets (?)

Figure 4: Pros and Cons of Data Integration Option

THINGS TO TAKE CARE OF WHEN INTEGRATING DATA FROM DIFFERENT STUDIES

Because studies are conducted in different periods (years), potentially by different vendors, when preparing ISS/ISE, regardless of the integration option you choose, the following aspects should also be taken into consideration:

- Subjects participating to more than one study e.g., roll-over. Care should be taken

in ensuring that the SDTM variable of USUBJID is depicting a single unique subject in the integrated datasets

  Subjects may have taken part in more than one study within the integrated datasets, and it is important that these subjects have the same USUBJID across all studies in the integrated datasets. A typical example is an Opel Label Study enrolling subjects from an earlier Blinded Study (see SDSP example in figure 2)

Medical dictionaries should be up versioned

  so that all Adverse Events (AEs), Concomitant Medications (CMs), etc. are coded to a single version. Changes that have occurred during the up-versioning process should be properly documented, for example, in the reviewer guide or with ad-hoc documentation (i.e., "Bridge" Document), so that the impact of the up-versioning on the results presented in the original study Clinical Study Report (CSR) is clear
- For CMs, consider "The PHUSE Recommendations for Pooled Submissions with WHODrug B3 Format Data White Paper" [7] for additional useful and relevant inputs and recommendations on integrating prior and concomitant medications across different formats of the WHODrug classification
- Requirements for CDISC-CT are similar. If you choose option 3, these "alignments" can occur in the pooled SDTM together with other structure harmonization to remove or align study differences, for example:
    - Standards unit conversion i.e., Laboratory Data
    - Visit naming and numbering conventions, when applicable i.e., "Study Day 1" in study 1 and "DAY 1" study 2 are highly likely to have the same meaning.
    - Other terminology such as supplemental qualifiers

If you chose option 3, the icSDRG should describe the changes you applied to individual studies when integrated to the iSDTM.

Furthermore, also consider the following recommendations:

- Trial Design Domains i.e., TS, do not have to be pooled. This was confirmed by the FDA through edata@fda.hhs.gov

The pooled datasets should be submitted to eCTD section 5.3.5.3 Reports of Analyses of Data from More than One Study. Ts.xpt is not required for eCTD section 5.3.5.3

  and as such rejection criteria will not apply e.g., missing for TS
- The same is applicable to datasets like SV, unless really needed in the analysis
- Integrate in the iSDTM only those domains that will be used in the ISS/ISE analysis. For example, if the integrated analysis plan does not mention summarization of prior and concomitant medications, there is no need to integrate the SDTM CM domains (nor the ADaM ADCM dataset) and of course it will not be necessary to up-version concomitant medical dictionary.

Conformance Issues

When integrating datasets from multiple studies, and conformance of the created datasets is checked using validation tools such as Pinnacle21, several CDISC conformance rules can be broken regardless of the option used; this is expected, and the issues should be documented in the relevant reviewer guide. More details on the type of conformance issues you may encounter are discussed in the PHUSE whitepaper and later discussed in the Use Case number 3.

Many of the topics mentioned here have already been discussed and experiences shared by different companies, either sponsors or CROs, in public events, as such I recommend the reading listed in the "recommended Reading" section.

PLANNING THE ISS/ISE INTEGRATION

The integration should be planned right at the start when the strategy for submission and the ISS/ISE Statistical Analysis Plan is drafted by the Biostatisticians. At Cytel we have set up a standard process to aid sponsors in planning data integration for ISS/ISE, starting with a Gap Analysis, where closed and ongoing studies "candidate" for the ISS/ISE are assessed. The Gap Analysis is made of two "dimensions" 1) Biostats Gap-analysis 2) Data Gap-Analysis. The data gap-analysis starts with the study data elements inventory, which includes determination of available and their format, either legacy or standard; this will be the basis for drafting the SDSP. See Appendices 1-3 for more details on what should be tracked and inventoried.

USE CASES

Looking back at the last 10 major FDA submissions we did at Cytel in the last three years, of which one is ongoing, we did apply the following data integration options:

Data Integration Option	N
Option 1 / Individual Study SDTMs to iADAM	4
Option 2 / Individual Study ADaMs to iADAM	2
Option 3 / Individual Study Datasets to iSDTM to iADaM	4

Table 1: Data Integrations options used in the latest 10 Cytel Submissions

Our numbers, even if based on a limited sample size, demonstrate there is not one unique option to adopt, but the option to adopt depends on the status and conformance, and variability, of individual study datasets and in some cases on sponsor preference (or I would say "prejudice").

In the next sections I'm going to analyze some additional key learnings and challenges from all three pooling scenarios I have discussed in previous sections based on some of our recent experiences.

A common denominator of all the experiences we had, in particular with sponsor of small dimensions with lack of internal resources, was probably the client themselves, along with the timelines and updates to the ongoing individual study packages, given the fact Cytel was not appointed for all ongoing studies and in some instances, we were only appointed for the ISS (and ISE) part. With most of the sponsors, it was often the case they would change their minds, and sometimes opt for very general and vague ISS SAP definitions. However, I believe this is possibly a "normal" scenario due to the nature of the projects where clinical drug development is still ongoing and new "inputs" might come from ongoing studies / research. As such, personal opinion, sometimes ISS/ISE are extremely exploratory meaning that things and approach, and required analyses, could change from time to time.

USE CASE 1 USING OPTION 1: SDTM(S) TO iADAM

This submission involved 2 studies in a Gynecologic indication conducted between 2017 and 2020. Both studies were conducted by Cytel, so a continuous harmonization effort was maintained throughout the two studies lifecycles by the same Cytel team.

The chosen pooling approach for this ISS/ISE was option 1. The decision to not re-use individual study ADaMs, and as such go with option 2, was taken after carefully assessing the analyses performed in the two individual studies as opposite to what was required by the planned analyses in the ISS/ISE SAPs, where several additional slightly different endpoints were used, also requiring aggregation and definition of different periods.

We could have opted to first pool at the SDTM level, so creating an iSDTM from which iADaM could have been derived but giving the similarity between the two study SDTM packages and the limited number of studies (2), we did decide that creating the iADaM package by directly using the two study SDTM packages was best option and it was not worth to create an additional SDTM package requiring also the creation of an additional define.xml and cSDRG.

USE CASE 2 USING OPTION 2: ADAM(S) TO iADAM

This submission involved six studies in an Endocrinology indication conducted between 2017 and 2021. The pooling approach for the planned ISS/ISE was already decided by the sponsor and agreed with the FDA prior to involving Cytel. The sponsor, given the fact all studies were applying CDISC standards, both SDTM and ADaM, decided to go with option 2 hoping to be able to re-use a lot of the work already performed for the individual CSRs; the hope was that the safety analysis done at the study level was same or very similar to the analysis planned for the ISS, and also the hope was that the structure of ADaM datasets "designed" by the individual CROs were consistent. This is "on paper" the ideal scenario where we can re-use what was done in the individual studies and so, theoretically, skipping a lot of data derivations and transformations.

After all, this ISS/ISE posed several challenges,

in particular with regards to timelines, when for example your final ISS/ISE submission is close to the end of the pivotal study(s), and you need to integrate ADaM datasets from more than one ongoing study.

However, a study can be integrated in the iADaM only when the individual CROs provide the individual ADaM, which sometimes is long after the availability of the SDTM.

This also meant in parallel that we received many different versions of the ongoing individual study ADaMs that would often change in structure (as Sponsor indecision would affect the CRO programming the individual study ADaMs).

Initially, the sponsor suggested we rederive a lot of variables in the pooled ADaMs that already existed in the individual study ADaMs, such as BASE, CHG, SHIFTy etc. and they would meticulously check these against the individual study outputs (which we did not have). Eventually, we would just pick up as much as we could from the individual study ADaMs and only rederive if missing/necessary to minimize any inconsistencies with the individual study outputs/ADaMs. We suggested the latter approach is used for pooled ADaMs and discussed early:

- Sometimes it could be difficult when handling subjects who participated in multiple studies within the pooled ADaMs without the additional granular detail from the SDTMs (e.g., specific exposure dose levels, disposition, source data for derived parameters in the ADLB). However, it wasn't a major problem, and in most cases was entirely possible to use the variables in the individual study ADaMs
- You can't validate the pooled ADaMs in Pinnacle 21 against SDTMs so you cannot avoid the three errors: "Traceability rules not executed due to missing XX dataset" for DM, EX, and AE. But I believe it is a minor issue and it was documented in the iADRG
- Derivation criteria often varied between individual studies. For example, for classifying specific disease conditions, the originally derived variables were differing due to differences in age groups/sites and there were numerous discussions on how to handle this as well. Initially the sponsor had given us a set of criteria that we should apply in the ISS ADaMs, but they had various concerns about this. In the end we used the criteria as defined in the individual studies and explained this in the SAP text/iADRG, but inconsistent approaches could make this slightly difficult (for example use AVALCATy vs CRITy vs PCSFL etc.). It just meant extra time spent combing through the datasets, specifications, define.xmls, as there's not always one "correct" method when ADaM programming compared to the more rigid structure of SDTMs
- Inconsistent derivations/definitions of analysis flags between the individual studies when compared with the iADaMs often meant it was very difficult to have 100% consistency with individual study outputs. Again, we think this is fairly standard for an ISS/ISE, but was something we had to explain several times to the Sponsor and spent considerable time checking/investigating
- Applying visit windowing could potentially require some extra thought, because we may not always have the original CRF captured visits or VISIT/VISITNUM from the SDTMs kept in the individual study ADaMs, and only AVISIT(N), ATPT(N) present, which may have already been remapped
- Origins and terminology of the source datasets could quickly become confusing. For example, the source data for the pooled ADSL could likely be a combination of 5 ADSLs, 5 ADEXs and 3 ADLBs (if a baseline characteristic is required in the ISS/ISE) etc. Explaining and emphasizing that there are no circular dependencies in the iADaMs is crucial in the specifications and/or ADRG.

Handling Subjects participating in more than one study

In this ISS/ISE we had an open label study where subjects participating in previous double-blind studies (core study) were offered to continue the experimental treatment in an open-label study (extension), or to start the experimental treatment if previously randomized in the Placebo arm (roll-over subjects). To satisfy the CDISC and FDA requirements for uniquely identifying subjects in the submitted data packages, the same USUBJD was assigned and a variable, ASTUDYID (Analysis Study), was added in all datasets containing the study id where a record originates.

For roll-over subjects, the value of STUDYID in the core studies was populated in the corresponding pooled ADaM STUDYID. For example, a roll-over subject who participated in the 001, 002 and 003 studies could have STUDYID equal to "001" for every record, with ASTUDYID populated as "001" or "002" or "003", depending on the timing of the given observation i.e., during the core study or the extension study. For example, if a rollover subject had two adverse events in both the core and extension study, the ADAE dataset would contain two records with consistent USUBJID and STUDYID, but a unique ASTUDYID for each period (core vs extension); of course, appropriate APERIOD variable was also derived. The variable SUBJID was also retained in all ADaM datasets containing the original subject id. assigned in each individual study.

Furthermore, to satisfy the ADSL constraint requiring to have one record per subject/USUBJID, the variables TRT01P, TRT01A, etc. for period 1 were used to identify boundaries and characteristics of the treatment given in the blinded study, while TRT02P, TRT02A, etc. for period 2 were used to identify boundaries and characteristics of the treatment given in the open-label study (when applicable). Similar approach was used to handle other variables such as age at enrolment or baseline values of each specific study participation.

Medical Coding up-versioning

For the ISS, medical dictionary coding such as MedDRA was up versioned from three different versions used in the individual studies to one unique version of MedDRA (the current version available at the time of the up versioning). Giving the fact we did pool in iADaM directly from individual study ADaMs, we did apply the up versioning in the iADaM, for example in ADAE, and for traceability we kept original medical coding assigned in the individual studies in the iADaM dataset, and so in the individual CSR, using the standard OCCDS variables e.g., DECDORGx / PT in Original Dictionary x, BDSYORG1 / SOC in Original Dictionary x, etc.

Multiple Baseline

The baseline definition used in individual studies did not differ from the one used in the ISS/ISE. However, for roll-over subjects the baseline value occurred in the core study might be needed versus the one in the extension study, depending on the circumstances. The variable BASETYPE was used to handle the different baseline scenarios and records duplicated accordingly as suggested in the ADaM IG.

Re-deriving parameters from individual study ADaMs

To identify parameters copied from individual ADaMs from those re-derived in the iADaM because of different derivation approach in the ISS/ISE, in addition to describe the scenario in the iADRG and the corresponding define.xml, the ADaM variable PARAMTYP was added to distinguish "original" vs "new" ISS/ISE parameters.

USE CASE 3 USING OPTION 3: SDTM/LEGACY(S) TO iSDTM TO iADAM

This submission involved twenty-five studies in a Neuro-Degenerative indication conducted between 2004 and 2022. Cytel was appointed to support the pivotal Phase III study and the ISS. Following a data gap-analysis where data from all closed and ongoing studies were assessed, simultaneously with discussion with Cytel biostatistician and the sponsor on the ISS strategy, it was decided to go with data integration option number 3 (see figure 3 for SDSP describing the approach used).

The decision to go with option 3, and as such to first pool into an iSDTM, was driven by the high heterogeneity of data structure used in the individual studies, including use of legacy format, and unavailability of proper analysis datasets for many of the older studies. Table 2 summarizes the different SDTM versions or legacy data structure used.

Version of SDTM / Legacy	Number of Studies
SDTM-like or Legacy	10
3.1.2 Amended or below	10
3.2	5

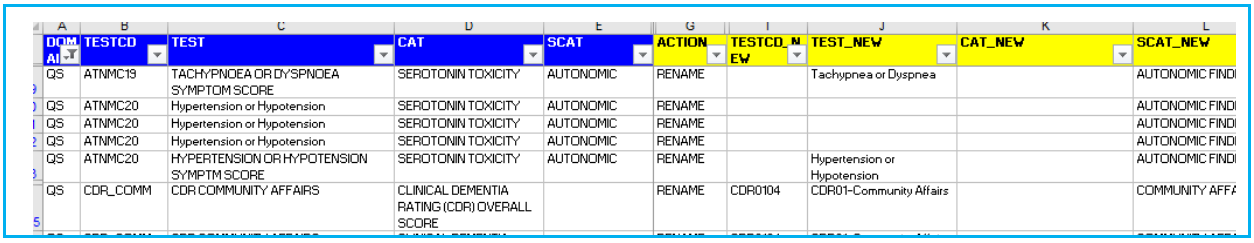
Table 2: Data Structure used in individual studies contributing to the ISS

Aligning CDISC Terminology and handling of harmonization

Given the number of studies, and the different "era" where studies were conducted, the main effort one could expect from this data integration "exercise" is with data harmonization process across studies, and that's what happened with this submission. The process for data harmonization was tedious and it required most of the resources assigned to the iSDTM part. The PHUSE white paper already contains some recommendations on what should be done when integrating data into an iSDTM and those recommendations match what we have done in this ISS, such as:

- We limit records in the pooled iSDTM to information needed in the ISS analysis, as such for example not all lab parameters found in the individual studies, more than 100, were mapped into the iSDTM
- all --TESTCD / --TEST were harmonized whenever possible. We initially preferred for example to not fully harmonize ECG parameters as we were unsure of specific mapping decisions taken in some studies at the time of original study mapping and there was no clear documentation from central labs providing ECG results that a specific method was used e.g., aggregate vs average formula, until the sponsor was able to confirm, but we preferred to have the harmonization done during the analysis so in ADaM
- harmonization of standard units e.g., laboratory data. All applied conversions were documented in both icSDRG and in the ISS SAP as per standard Cytel process
- For VISIT/VISITNUM we simply tried to harmonize terminology when obvious e.g., "VISIT 1" to "Screening" when from the protocol we could see visit 1 was a Screening visit. Any other harmonization / aggregation requiring the application of some windowing was done in the iADaM
- For EPOCH, similarly to what we did for VISIT/VISITNUM, we harmonized wording or similar concepts. We also created EPOCH for studies where EPOCH was not defined, while, apart the obvious wording harmonization, we didn't try to re-derive the variable, as such different methods might have applied in individual studies, so methods were described in the icSDRG

All changes in the individual studies applied in the iSDTM were documented in the icSDRG. To streamline the harmonization effort, its documentation and its application, and the continuous medical review, we created some SAS macros to automatically generate SAS code out of the excel file that we had used to track changes. See example in figure 5.



A	B	C	D	E	G	I	J	K	L
DOM AI	TESTCD	TEST	CAT	SCAT	ACTION	TESTCD_NEW	TEST_NEW	CAT_NEW	SCAT_NEW
QS	ATNMC19	TACHYPNOEA OR DYSPNOEA SYMPTOM SCORE	SEROTONIN TOXICITY	AUTONOMIC	RENAME		Tachypnea or Dyspnea		AUTONOMIC FINDI
QS	ATNMC20	Hypertension or Hypotension	SEROTONIN TOXICITY	AUTONOMIC	RENAME				AUTONOMIC FINDI
QS	ATNMC20	Hypertension or Hypotension	SEROTONIN TOXICITY	AUTONOMIC	RENAME				AUTONOMIC FINDI
QS	ATNMC20	Hypertension or Hypotension	SEROTONIN TOXICITY	AUTONOMIC	RENAME				AUTONOMIC FINDI
QS	ATNMC20	HYPERTENSION OR HYPOTENSION SYMPTM SCORE	SEROTONIN TOXICITY	AUTONOMIC	RENAME		Hypertension or Hypotension		AUTONOMIC FINDI
QS	CDR_COMM	CDR COMMUNITY AFFAIRS	CLINICAL DEMENTIA RATING (CDR) OVERALL SCORE		RENAME	CDR0104	CDR01-Community Affairs		COMMUNITY AFFA

Figure 5: Excel file to track changes in iSDTM e.g., --TESTCD/--TEST/--CAT/--SCAT

From that file we were then dynamically generating SAS code to be applied in each individual findings class iSDTM domain (see figure 6). The same approach was used to other harmonization / recoding, such as SI unit conversion, and Race, end of treatment/end of study, arm, treatment exposure recoding as well selection and recoding of supplemental qualifiers.

Special Cases

Given the "age" of some studies, we did perform full review of certain key variables that in past submissions were the object of some observations from the FDA reviewers when some key variables were reported as 'OTHER' while the "Other, specify" was containing text that could have been classified in one of the expected "standard categories"; for example, end of study reasons classified as 'OTHER' while the specification of other was referring to an adverse event,

or Race classified as 'Other' while from the specification of other you could see a standard Race could have been selected, for example "SOUTH EAST ASIAN" could have been reported as "ASIAN". So, we did change the classification and kept the original value in the supplemental iSDTM qualifier datasets.
All this re-classification activities require of course the support of a clinical expert and regular review and for that reason having a way of sharing files with agreed changes and be able to automatically generate the code was of big help to streamline the double activity of specifications / programming.

```
/*---Controlled Terminology Harmonization---*/

    /*Generated by %RECODE on 24OCT22*/

    /*Source: AllFindingsClass.xlsx*/
    /*Output: qs-findings-class-recode.sas*/
/*----------------------------------------*/

/*Fake Variable*/
___Fake="";drop ___Fake;


IF upcase(compress(STUDYID)) EQ "001-001" AND  upcase(compress(QSCAT)) EQ "CLINICALDEMENTIARATING(CDR)OVERALLSCORE" AND
   upcase(compress(QSSCAT)) EQ "" AND  upcase(compress(QSTESTCD)) EQ "CDRTOTAL" THEN DO;
   QSTESTCD="CDRTOT";
   QSTEST="CDR01-Overall CDR Score";
END;

ELSE IF upcase(compress(STUDYID)) EQ "001-001" AND  upcase(compress(QSCAT)) EQ "CLINICALDEMENTIARATING(CDR)OVERALLSCORE" AND
   upcase(compress(QSSCAT)) EQ "" AND  upcase(compress(QSTESTCD)) EQ "CDR_COMM" THEN DO;
   QSTESTCD="CDR0104";
   QSTEST="CDR01-Community Affairs";
   QSSCAT="COMMUNITY AFFAIRS";
END;
```

Figure 6: Automatic Code generation from excel terminology harmonization file

Medical Coding up-versioning
As opposite to use case 2, in this submission we did apply the medical coding up-versioning in the iSDTM, by replacing original coding in the iSDTM AE and CM datasets. We decided not to store original coding in the supplemental qualifier, but to track changes in an internal documentation that was also reviewed by the clinical expert at the sponsor side to confirm they were happy for the changes proposed by our coding specialist.
To also give an idea of the effort required by the medical coding up versioning, table 3 reports some numbers specifically related to up-versioning of MedDRA for adverse events, from 11 different MedDRA versions, ranging from version 9.1 to version 25.0.

Item	N
Number of AE term to up-version	22129
Number of different MedDRA versions applied in individual studies	11
Number of AEs previously not coded	25
Wrong casing in original coding	13
Number of coding changes caused by the up versioning (any)	4676
Number of coding changes caused by the up versioning (SOC)	195
Number of coding changes caused by the up versioning (Preferred term)	1179

Table 3: Impact of AE MedDRA up-versioning

"Bridge" documents showing changes in medical coding caused by the medical dictionary up versioning, were also provided as an appendix of the icSDRG.

Handling "Unavoidable" conformance issues
Giving the fact all CDISC standards and their conformance rules, and as such tools implementing conformance rules such as Pinnacle21, are not built to handle specific data integration requirements, when validating iSDTM, as well as iADaM, you might need to handle a lot of conformance issues, you will have to either fix or most of the time provide a rationale in the icSDRG. I here report some conformance issues for which we had to provide a rationale in the icSDRG.

Missing TS dataset (SD1115)
As per communication with the FDA eData team, TS is not applicable for eCTD section "5.3.5.3".

Variable appears in dataset, but is not in SDTM model / SUBJID (SD0058)
The variable SUBJID has been copied onto all datasets to track the subject id assigned in the original study when the subject participates in more than one study.

Incorrect value for AESTDY variable (SD1090)
This occurs when subjects participated in more than one study. The Pinnacle check doesn't take into consideration the STUDYID in getting the reference start date from DM.

Inconsistent STUDYID (SD1349)

This is because we have pooled subjects from different studies, so STUDYID contains the original Study id.

EGTEST/EGTESTCD value not found in 'ECG Test Name' extensible codelist (CT2002)

Some terms were kept as per original study datasets / original terminology at the time of study database lock. Harmonization when needed is done in iADaM.

Value for MHDECOD not found in MedDRA dictionary (SD0008)

MedDRA was not up versioned for Medical History as per ISS SAP, so different versions are used. See table 1 in section 2.2 for more details about MedDRA versions used by individual studies

CONCLUSIONS

Regardless of which option you adopt for your next ISS/ISE, traceability and proper documentation are crucial, and so is early planning and discussion of your integration strategy with your relevant regulatory agency.

In general, option 2 can be considered the best option to choose. This option is slightly quicker and simpler than option 1 and 3 and it ensures everything is consistent with individual study ADaMs and so the individual study CSRs. We think a crucial point is gauging the client's expectation up front of how much standardization/derivations would be expected to be performed in the iADaMs, as this was sometime not explicitly discussed with the Sponsor(s) up front and caused us a few complications later when they changed their minds about things (this was also the source of several budget change orders).

However, our experience showed also that option 2 might not always be the right option to choose even when all concerned studies made use of CDISC standards for both source (SDTM) and analysis datasets (ADaM), unless the sponsor is able to make an appropriate surveillance of CROs work, if studies are outsourced, especially to multiple vendors. The sponsor needs to make sure not only the structure of ADaM datasets is consistent, but also definitions are consistent across studies e.g., derivations.

Upcoming Industry Standards

In 2019 the CDISC ADaM team released for public review the "ADaM Data Structures for Integration Document". However, the final release of this guidance is still on-hold and there is not a clear plan on when it will be released.

PHUSE has also started a new project to develop a template, completion guidelines and example documents for an integrated reviewer guide for ADaM (iADRG), for which a public review was completed last April.

REFERENCES

[1] "Integrated Summaries of Effectiveness and Safety: Location Within the Common Technical Document", FDA Guidance for the Industry

[2] "Integration Strategies in Support of ISS/ISE Submissions" – PHUSE White Paper 2020, https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Optimizing+the+Use+of+Data+Standards/Integration+Strategies+in+Support+of+ISS+ISE+Submissions.pdf

[3] "ADaM Structures for Integration: A Preview", W. Zhong, K. Minkalis and D. Bauer, PharmaSUG 2018

[4] "ADaMIG v1.2 & ADaM Integration", CDISC Webinar, 2019, https://www.cdisc.org/events/webinar/adamig-v1-2-adam-integration

[5] "Expert Answers to Community Questions" PHUSE webinar Friday March 26th, 2021 https://event.on24.com/eventRegistration/EventLobbyServlet?target=reg20.jsp&referrer=&eventid=3023348&sessionid=1&key=15BC66BD7480E5D87BBC2202C798E76C&regTag=&V2=false&sourcepage=register

[6] "Study Data Standardization Plan" https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Optimizing+the+Use+of+Data+Standards/SDSP.zip

[7] "The PHUSE Recommendations for Pooled Submissions with WHODrug B3 Format Data White Paper", https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Optimizing+the+Use+of+Data+Standards/Recommendations+for+Pooled+Submissions+with+WHODrug+B3+Format+Data.pdf

RECOMMENDED READING

Background

[1] "Navigating FDA requirement: ISS/ISE build strategies", PhUSE-EU Connect, 2019

[2] "Processes and Techniques for creating Integrated Summary of Safety (ISS)", G. Jangid and J. Marrer, PhUSE-US Connect 2021

SDTM and ADaM

[3] "Challenges involved with using legacy clinical trials study data during integration", A. Opasanya and A. Bonwick, PhUSE-EU 2019

[4] "Simplifying the Integration Riddle", A. Sri Krishna Mani, PhUSE-EU Connect, 2019

ADaM

[5] "Integrated ADSL", Nate Freimark and T. Clinch, PhUSE 2012

[6] "Choosing the right path to follow when integrating ADaM", PhUSE-EU Connect, 2019

[7] "Best Practices for ISS/ISE Dataset Development", B. Donthi, PharmaSUG 2019

[8] "Integrated Summary of Safety and Efficacy Production Strategies", N. Agnihotri and S. Pradham, PhUSE-EU Connect, 2021

[9] "ADaM Integration for Summary of Clinical Safety: The 'Unique Patient'", T. Sheman, A. Shah, B. Fairfield-Carter, PharmaSUG 2018

[10] "Ensuring USUBJID is Unique for an Individual within an Application", M. Mattson, M. Sullivan, and G. Walker, PhUSE, 2015

CONTACT INFORMATION

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

Angelo Tinazzi
Cytel Inc.
Geneva, Switzerland
Email: angelo.tinazzi@cytel.com
Web: www.cytel.com
The "Good Data Submission Doctor" Cytel Blog: https://www.cytel.com/blog/author/angelo-tinazzi

APPENDIX I: Data Gap Analysis - Inventory

Item	Study nr 001	Study nr 002
Study Characteristics		
Planned in the ISS/ISE Pool?		
Trial Title (brief)		
EudraCT Number / Clintrials.Gov Number		
Phase		
Population		
Randomized		
Randomized Subjects in CSR		
Objective (1)		
Objective (2)		
Objective (3)		
Objective (4)		
Additional Notes		
CRO		
Study Material		
Protocol Version		
Sample/ Blank CRF		
EDC aCRF		
SAP		
CSR		
TLFs		
Data Transfer Specifications		
External Data		
Randomization Schema		
Other DM documents		
Analysis Datasets Specs		
(Comments)		
Datasets		
Data Available?		
Source Data (SDTM/Legacy)		
(Comments)		
Analysis Datasets (ADAM/Legacy)		
(Comments)		
SDTM Package		
SDTM Package (N=Only Datasets but no xpt, define, sdrg,		
SDTM lg		
SDTM CT		
SDTM define.xml		
SDTM csdrg		
SDTM acrf		
ADaM Package		
ADaM Package (N=Only Datasets but no xpt, define, sdrg,		
ADaM lg		
ADAM CT		
ADaM define.xml		
ADAM adrg		
ADAM Programs		
Medical Dictionary		
MedDRA Version		
WHO-DD Version		
Datasets Checks		
Datasets Readable		
Datasets vs aCRF or other specs		
Datasets vs Protocol e.g. primary/sec endp		
Datasets records (nr. In DM) vs Planned in the protocol / CSR		
First Patient In / Last Patient Visit (CSR)		
Comments		
Datasets Analysis / Review		
Local Lab/Central Labs/SI Unit		
Protocol Deviations		

APPENDIX II: Data Gap Analysis – Overview of Data Packages Assessment

Activity	Description	Outcome (Summary)
Run P21 validation	Perform validation of SDTM/ADAM datasets with define and of define alone. Assess current issues and recommend actions	
SDTM aCRF	Review if aCRF was annotated correctly and if bookmarks are present (see Cytel aCRF-Good-Annotations-Rules)	
Define.xml and reviewer gudie Review	Check if the define.xml are of good quality with particular focus on the way comments/methods are described i.e., use of SAS code, check if all P21 issues are described in the RG	
SDTM mapping heterogeneity	Assess if any major differences between studies, with particular focus on the following aspects: - Use of SI unit for Labs - FDA main requirements: presence of EPOCH, Treatment Emergent flags, Race, Screen Failures and subject randomized not treated, use of special characters - Presence of Subject Initials	-
Legacy Conversion	Assess effort required for legacy datasets migration to SDTM	See section 4
FDA Feedback	Check for historically feedback received from FDA by other sponsors	See section 5
All datasets can be opened	Check if all datasets can be opened and if they all have the expected (count) subjects	
aCRF vs received datasets	Make sure all expected datasets in the aCRF or eDC CRF in case of legacay datasets are available. Check if any external datasets are available, tipically for Pk check availability of both concentration and Pk estimates data, protocol deviaton, etc.	See section 4

APPENDIX III: Data Gap Analysis – Detailed Review Checklist

Section	Item	Data set	Topic	Study 1	Study 2
Standard Checks					
FDA	PDF Documents	All	PDF Properties / eCTD filename		
FDA	aCRF	All	Proper annotations / bookmarks aCRF		
FDA	Datasets	All	Presence of special characters		
FDA TCG	Datasets	All/ Many	Presence of EPOCH		
FDA TCG	Datasets	AE	Seriousness Criteria presence		
FDA TCG	Datasets	AE	TEAE in SUPPAE		
FDA TCG	Datasets	DM	Screen failure should have ARMA/ACTARM blank		
FDA TCG	Datasets	DM	ACTARM blank for randomized not treated		
FDA TCG	Datasets	DS	Use of DSCAT/DSSCAT		
FDA TCG	Datasets	TS	Trial Start Date (PARAMCD=SSTDTC)		
Other SDTM	Datasets	DD	Check against DM and AE		
Other SDTM	Datasets	DM	Race Terminology		
Other SDTM	Datasets	EC	To be used in situation when NOT DONE admin. EX needed anyhow		
Other SDTM	Datasets	Findings	Check Imputations e.g. in LBSTREN when LBSTRESC=<xxxx		
Pool	Datasets	All	CT Consistency		
Pool	Datasets	All	VISIT/VISITNUM Consistency		
Pool	Datasets	LB/VS/EG/….	--TEST/--TESTCD/--STRESU consistency		
Pool	Datasets	Many	--CAT/--SCAT Consistency		
Pool	Datasets	Many	SUPP Consistency		
Pool	Datasets	Many	Old Ig/Old CT vs new Ig/New CT		
Pool	Datasets	Many	Presence of domain by study		

Section	Item	Data set	Topic	Study 1	Study 2
Pinnacle 21 Reports					
P21	SDTM	All	Define.xml		
P21	SDTM	All	Datasets+define.xml		
P21	ADaM	All	Define.xml		
P21	ADaM	All	Datasets+define.xml		
FDA Mock Submission					
SDTM	Define	Many	Merged Codelists across different variables		
SDTM	Define	Many	Missing Codelist for some variables		
SDTM	Define	Many	Missing Codelist for all Value Level Metadata		
SDTM	Datasets	CM	Many collected terms in CM domain were not converted into standard CT terms		