Paper Number: DH12

Validation consistency and conformance checking of SEND datasets

Mohit Mathew, PointCross Lifesciences, Silver spring, MD USA
Dr. Karen Porter, PointCross Lifesciences, Silver spring, MD USA
Dr. Laura Kaufman, PointCross Lifesciences, NJ USA
Karuna Polavarapu, PointCross Lifesciences, Bengaluru India

ABSTRACT

Validators can check if nonclinical studies submitted in CDISC SEND format meet FDA business rules, FDA validation rules, CDISC conformance rules and PMDA's Define.xml rules as applied for SEND. However, these validators cannot check if the SEND Trial Design is modeled correctly according to the Study's design with sufficient granularity. For example, if a Latin square study is coded in SEND as a parallel design, the SEND dataset cannot be reviewed or analyzed. Since SEND datasets are prepared separately using a CDISC trial design, the SEND data set should be checked to ensure it will generate the same summary results as listed in the audited study report. Manual "spot checks" are insufficient and risky. This paper discusses the best practices for sponsors to make sure that their SEND data set is of high quality and is consistent and conformant with the Study Report using special tools and validators.

INTRODUCTION

Since the FDA mandated that NDA and IND submissions use the CDISC SEND data standard for non-clinical studies, there has been a steady increase in SEND submissions.

Reviewers have always used the Study Report to conduct regulatory reviews independently. SEND data offers reviewers unlimited opportunities to select subject groupings and timing for testing any hypothesis directly with a computer and analysis scripts. To do that with confidence they need to have confidence in the SEND subject level data. Since the Study Report is the only other report that is signed and audited, it is natural that they will look to it as a "calibration" device to make sure that the SEND data does in fact generate the same summary data if the same assumptions are applied. If that consistency cannot be demonstrated or easily proven, or if the consistency is suspect, then the reviewer cannot depend on the SEND data set, even with its richness, to make regulatory decisions. This paper will discuss specific ways that sponsors can ensure that the consistency of the SEND data with the Study report can assured during preparation, and demonstrable after the SEND data set is generated.

CONFORMANCE AND COMPLETENESS

The FDA has set up multiple checkpoints for the submitted data to be assured that the SEND dataset is fit for review. It begins with the check by the eData group where the submission and the data sets within are checked to make sure that the FDA's published rules are met and that none of the technical rejection criteria are met at the eCTD level. Validators are rules-based engines that can verify if the data set being validated conforms to the rules supported by the validator. At this time these rules are a composite of the conformance rules published by CDISC for SDTM and SEND; or the business rules published by FDA or the rules defined by PMDA. The Technical conformance rules published by the FDA may also be tested at the study level or at the submission level. Together these check for conformance and completeness.

QUALITY ISSUES SEEN DURING FITNESS-FOR-REVIEW ASSESSMENTS

The SEND data sets are then passed through a Data Fitness assessment to ensure that the data is reviewable - i.e. to answer the question “Does it meet the needs of the Reviewer”. During the November 2019 SEND face-to-face meeting and the September 2019 FDA webinar, most commonly found issues in the data that prevent review or use of the data in the Janus nonclinical repository were listed. These issues were mostly related to inconsistency of the SEND data with the Study Report.

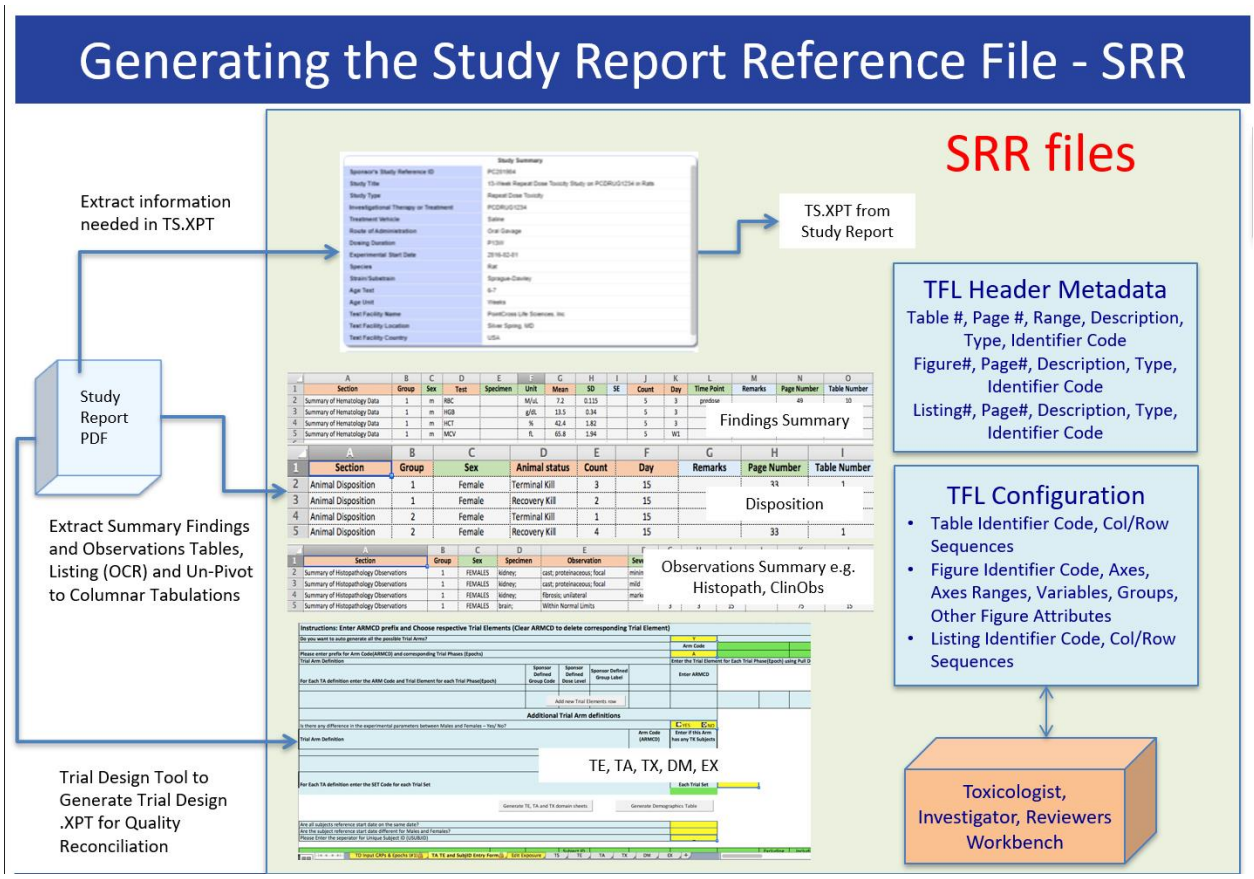
Ref: https://www.fda.gov/drugs/cder-small-business-industry-assistance-sbia/most-common-issues-cdisc-send-data-fda-toxicology-review-sep-12-2019

Consider some of the root causes for these types of inconsistencies. Experienced Toxicologists and Principal Investigators, who are solely focused on the analysis of the collected data and the report writing, often generate the Study Report soon after the study is closed and audited. On the other hand, teams skilled in SEND standards pick up the same GLP data from the LIMS and generate the SEND data set, the Define.XML and nSDRG – separate from, later than, and without systematic or deliberate reference back to the Study Report. Very few toxicologists are deeply familiar with CDISC standards, and SEND data is prepared by data and standards experts who see the GLP source data ex-post-facto. SEND preparers depend on spot-checks. But spot checks are just that – a statistically plausible step, but as the data shows, cannot deliver the kind of confidence that the reviewer needs to solely use the SEND data for review. This is not an issue for clinical studies submitted with SDTM and ADaM files because as-collected SDTM is used as a source for the analysis data sets by ADaM, where all the imputations, analysis assumptions and selection of cohorts are done before the final statistical analysis is run on the ADaM sets to generate tabulations, figures and listings for inclusion in the study report. In SEND, the ADaM step is by-passed, thereby creating twin processes for study report tables and SEND data sets that have no material ways to be compared. In order to eliminate these inconsistencies, we propose creating a “reference” file that represents the Study Report and is used to ensure that the subject level data that goes into the SEND data set is coded appropriately for use by downstream analytics applied on the SEND data.

The underlying need for a reference to the LIMS tables published in the Study Report used to analyze a study has been creeping into the standards deliberations in the past year. The need for a new variable, ASTAB, has been discussed in the recent SEND face-to-face meetings. ASTAB is “As-Tabulated” in the Study Report and refers to the subject level records that must be combined in order to regenerate an incidence count or a group summary that was published in the final report.

Nonclinical studies have pre-defined cohorts in the form of Trial Sets or Dose Groups. Summary calculations based on SEND datasets may require additional clubbing or combining of study days to replicate published summaries in the Study Report. Ensuring consistency requires the ability to easily map the SEND data to the published cohorts or dose groups in the Study Report. Steps being taken in the newer versions of the SEND IG such as the “nominal day” provides a placeholder for combining Visit Days reported in the report. Without a reference that represents the Study Report there will always be doubt about the SEND data set. That doubt can be eliminated with the use of a Study Report Reference file.

SRR - STUDY REPORT REFERENCE GENERATION



Since the key reference is the Study Report, which is available as a published, signed and usually audited PDF document, the proposed approach begins with electronically extracting the summary tables from the PDF report for the summarized findings (group means and SD) and incidence counts. The summary tables printed in the PDF report are intended to be human readable and the data is therefore pivoted to present the means, standard deviations and count against groups on their study days for each of the reported findings parameters.

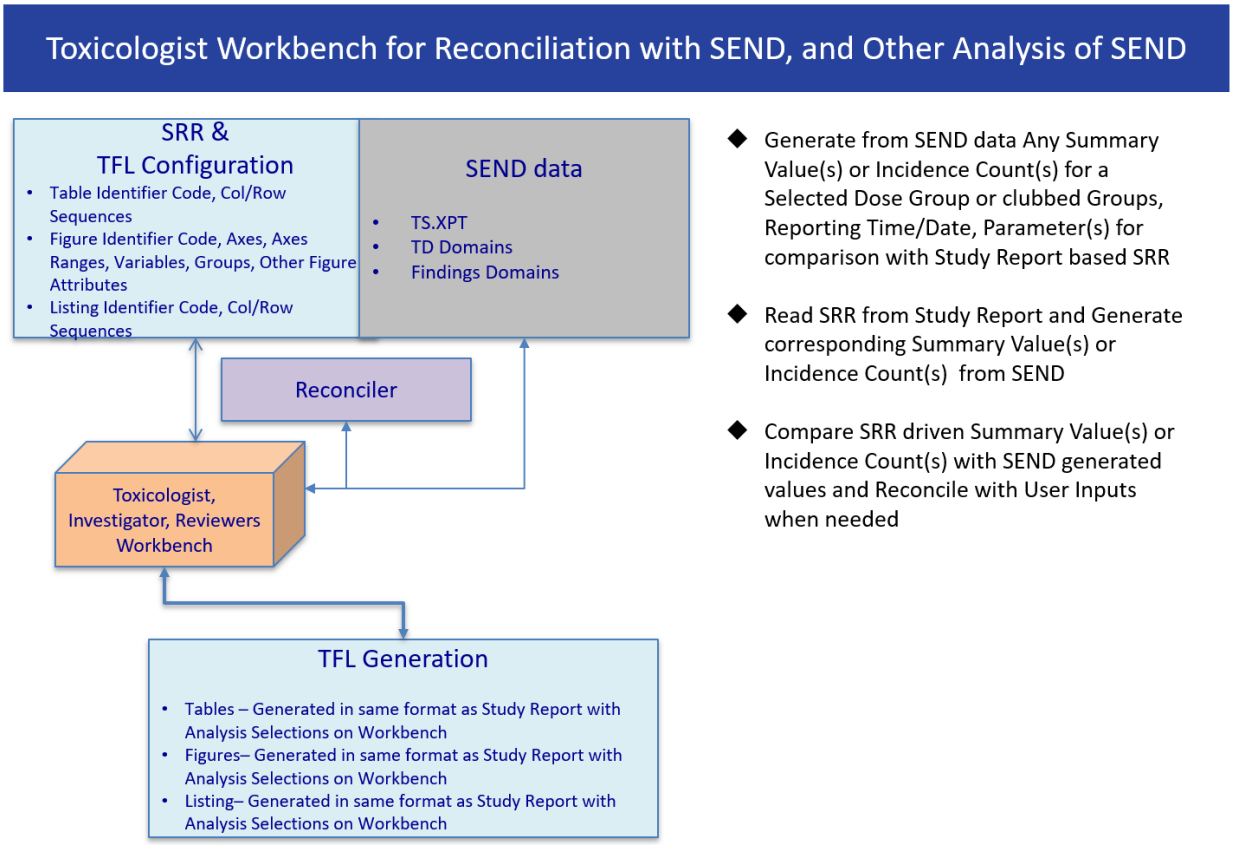
Extracting the published value, by itself, will not provide the necessary information that an analysis script can use to recalculate the summary results from the SEND individual subject data. Therefore the Study Report Reference (SRR file) needs to contain both the data and the decisions that went into calculating the published summaries.

Certain parameters such as Food and Water or Body Weight gain are reported according to investigator specified day-ranges. The same can be true for clinical observations. Therefore, the printed tabulated data is extracted retaining the provenance, including page number, table number, the sequence of the columns, parameters and rows. The extracted data is then un-pivoted so that they can be formatted into a columnar tabulation of records. This is the Study Report Reference file (SRR) that can be used by software for verification that the SEND data set can regenerate the same means, standard deviations and counts for each parameter on each study day reported. The provenance information associated with each record allows the data visualization tools to also regenerate the original tables and figures in the original Study Report using the data in the SRR records along with the Sponsor defined Group identifiers.

TDR - TRIAL DESIGN REFERENCE GENERATION

The Study Report describes the study design, the use of subjects and how they were organized into dose groups, and how certain subjects were chosen for additional PK/TK evaluations, or which ones were subjected to a recovery period, among other special conditions. Our approach uses the Study Report to generate an independent Trial Design that complies with the CDISC SEND IG. These include Epochs, Trial Arms (TA), Trial Elements (TE), Trial Sets (TX), Exposure (EX), and Demography (DM) using a Trial Design Automation tool. The trial design generated from the Study Report is the Trial Design Reference file (TDR). The TDR provides a reference generated only from the Study Report. TDR may then be used to verify that the SEND dataset is organized consistently with the same number of elements and inter-relationships within TAs, TEs, and TXs. The list of subjects and their IDs should be checked against all the SEND Domains including DM and TX to ensure that no additional subjects that may have subsequently been replaced are still included in the SEND data set while it is not in the summary report.

COMPARATOR - ENSURING THE QUALITY AND CONSISTENCY OF THE SEND DATASET



The Comparator, or Reconciler, is a toolset that can digitally and automatically compare the SRR and TDR files against the SEND data set. Each summary value from the SRR and the analysis rules are used to run a script against the SEND data set to make sure that the recalculated summary from the SEND data set matches the statistics or incidence counts published in the Study Report. The Reconciler can, at a minimum, report errors in the SEND datasets. When also used as a tool to aid SEND preparation, it can help sponsors or reviewers in their analysis:

- Report errors in the SEND data sets;
- Generate computer readable instructions to correct errors in the SEND data sets by the SEND preparers;
- Auto-generate text, with references to the provenance that can be used for explanations in the SDRG.

The comparator uses the mean, standard deviation and count for every sponsor defined dose group for every

collected parameter by reported visit day reported in the SRR file. These parameter names are semantically normalized using the same CT that is reported in the TS.XPT to make an easy comparison with the regenerated group summary calculated from SEND data sets.

The TDR reference file, generated directly from the Study Report, helps the comparison with the SEND data set's trial design and provides a way to validate the following:

1. For each sponsor defined dose group in the Study Report, the assignment specific subjects to a sponsor defined dose group in SEND is verified.
2. The assignment of Trial Elements to the right Epochs is verified.
3. Each unique Trial Arm is verified against the TA domain and any discrepancy must be reconciled, fixed, or explained in the nSDRG.
4. Trial sets in the TDR are verified against those in the TX.XPT domain in the SEND data set to ensure that all expected TXPARMs and TXVALs are entered in the SEND TX.XPT.
5. The list of Subjects in the TDR are the subjects listed in the Study Report. A comparison of this list against all the domains including, but not limited to, the DM domain, flags any subjects that are carried over from the LIMS system even though they may have been replaced. The subject IDs are also reconciled in this process.

The next step is the comparison of the SRR to the SEND data sets. The SRR contains the summary data presented in the Study Report tabulations organized in terms of the sponsor dose groups. Since the sponsor dose group assignment to subjects has already been verified in the TDR comparison, and the variable names in the SRR have been mapped to the chosen CT of the SEND data set, the SRR comparator automatically regenerates the group summaries from SEND data for all quantitative and categorical data incidence counts using the date ranges or groupings used in the Study Report and as listed in the SRR in machine readable form. The Means, Standard Deviation and Count are compared between the SRR and the SEND regenerated group summaries. Similarly, the expected combinations and filters used to report incidence counts for MA and MI are used by the comparator to regenerate the same counts using the same combination of findings and modifiers as reported in the summary report. In the case of CL domain, the date ranges must be picked up from the SRR to ensure that the incidence counts can be correctly regenerated. The checks run across all the rows of the SRR will identify:

- A complete match of Mean, Standard Deviation and Count for any dose group at any reported study day for any parameter means that the underlying subject level data in the SEND data sets for the subjects is correct and consistent with the data published in the Study Report. In certain data domains, the group means may have to be calculated over a range of study days, such as Body Weight Gain, if reported. In this example the verification of the Body Weight ensures that the SEND dataset carries the right body weight. However, if the Body Weight Gain happens to be reported, this comparison still needs to be done to ensure that the time intervals for calculation of Body Weight Gain are correct.
- A minor mismatch of Mean and Standard Deviation that represents the difference in the significant digits printed in the Study Report versus the floating-point number with more significant digits is checked.
- A mismatch of the Count points to a potential incorrect grouping for the reported time point or study day and any clubbing of "visit-days" to achieve the right count is first done before human intervention is needed. This results in a series of instructions to correct the timings of the reported data for the purpose of summary calculations. These instructions are used to generate a human readable instruction for SEND preparers; or machine-readable scripts that are used by the SEND preparation and visualization tools so that the SEND data sets can be corrected automatically.

Where differences are detected an explanation could be provided if the SEND data set has made assumptions about how to organize the Trial Sets. Differences in TA and TE are more likely to be serious or an error in the creation of the TDR.

CONCLUSION

SEND is a completely different way to present the same data that is also published in the Study Report. The CDISC SEND IG provides clear rules for implementing the data formatting into standard, or custom, domains. Conformance rules published by CDISC, and business rules published by FDA and PMDA provide rules that can be used to validate the Study, thereby ensuring that the SEND data set meets the SEND IG's de jure standard.

Ensuring that the SEND data set will generate the same signals and trends in the data as derived from the Study Report requires that SEND data is checked for consistency against the published, signed, audited Study Report. Creating a machine-readable data set (SRR and TDR) from the human-readable printed Study Report is an essential step for being able to compare the Study Report and the SEND data set structurally, and in its elemental data values.

A SEND data set that is checked 100% against the Study Report can be confidently considered ready for the Reviewer. Byproducts of the processes detailed in this paper also provide opportunities for a high level of automation for use by data preparers, quality checkers, data fitness assessors, and provide a clearer and unambiguous nSDRG.

REFERENCES

Most Common Issues with CDISC-SEND Data in FDA Toxicology Review – Sep. 12, 2019:

- *https://www.fda.gov/drugs/cder-small-business-industry-assistance-sbia/most-common-issues-cdisc-send-data-fda-toxicology-review-sep-12-2019*
- *Slides: https://www.fda.gov/media/131181/download*

The Impact of SEND data on FDA review of Nonclinical studies a PhUSE Presentation by Matthew Whittaker (FDA)

- *https://www.phusewiki.org/docs/2019_CSS/Presentations_Posters/CSS%20Presentations/3%20-%20MWhittaker%20PhUSE%20CSS_NONCLINICAL_5-31-19.pdf*

Data Consistency: SEND Datasets and Study Report Wiki:

- *https://www.phusewiki.org/wiki/index.php?title=Data_Consistency:_SEND_Datasets_and_Study_Report_Wiki*

Study Data Technical Conformance Guide:

- *https://www.fda.gov/media/131872/download*

Technical Rejection Criteria:

- https://www.fda.gov/media/100743/download

CONTACT INFORMATION

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

Mohit Mathew, Software Architect, PointCross Inc. (mohit@pointcross.com)
Dr Karen Porter, Senior Toxicologist, PointCross Inc. (karen@pointcross.com)