

Data Consistency and Quality Issues in SEND Datasets

PointCross has reviewed numerous SEND datasets prepared for test submissions to the FDA and has worked with the FDA on their KickStart program to ensure that received datasets are suitable for pharm/tox review. What we are observing is that creating a dataset that is *technically compliant* (i.e., one that passes the validation rules) is a low bar – it is necessary, but not sufficient to meet FDA review requirements.

In our experience, sponsors and CROs are having a much more difficult time ensuring that SEND datasets have no *data quality issues* that impact “fitness for review”. We have routinely observed cases where SEND datasets are inconsistent with the accompanying study report and have insufficient data for pharm/tox review despite the fact that these data often have been collected and tabulated in the study report.

These issues are serious and can result in delays in the review and approval process.

Technical Compliance Issues

Violations of the CDISC SEND standard or FDA Specific SEND Validation Rules in a SEND dataset prevent it from being loaded into the FDA’s NIMS. For example, a critical piece of missing information can disrupt data loading. We believe that simple formatting issues can be resolved with a little experience and self-run validations.

Some examples of technical compliance issues beyond simple formatting that we have observed include:

- Syntax errors in datasets or Define.xml that prevent validation or loading into FDA’s NIMS
- Coded values that are not defined within the SEND dataset

Data Consistency, Quality and Sufficiency Issues

Being technically conformant to the SEND standards should not be the only concern for sponsors when creating SEND datasets. Frequently, we are seeing that the greatest challenge for companies is ensuring the SEND dataset is *consistent and sufficient* for FDA review.

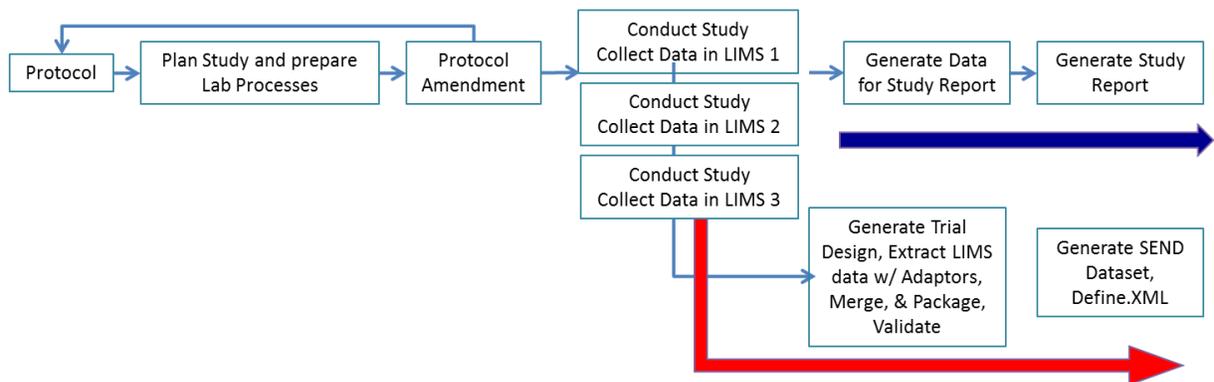
These types of issues may be the result of relying on two separate processes to generate the study report and the SEND dataset (see figure below). Because data typically originates from multiple LIMS systems with proprietary data models, sponsors or CROs creating SEND datasets are required to reconcile disparate terminologies, units, groupings, and coded data sources. As a result, there can be inconsistencies between the study report and SEND datasets.

If FDA reviewers see a signal of interest or calculate a group summary that does not match what is reported in the study report, these discrepancies will result in reviewers questioning the trustworthiness of the dataset. The review process may be disrupted until such data issues are resolved. In cases of extreme discrepancies, a submission could be put at risk.

Examples of data quality and consistency issues we have observed include:

- Inconsistencies in trial sets labels and study report groups, which can be confusing to reviewers.
- Incorrectly parsed modifiers for qualitative data, which can result in the misinterpretation of a finding.
- The SEND dataset lacks data present in the study report that can be reported in SEND. This can cause the reviewers to question why the data was not included, and delays the review process.
- Individual data elements in the SEND dataset missing or misaligned from the matching data in the study report, which may cause the reviewer to question the quality of the SEND dataset or the report.
- Specimen information missing in SEND domains like Laboratory Test Results and Macroscopic Findings, which makes the data difficult to review and interpret.

Data quality and consistency issues introduced by separate processes for the study report and SEND



- Keeping up with SEND and Controlled Terminology Releases, & FDA updates is complex
- SEND datasets are generated by a **parallel process** possibly by different people using Trial Designs needed for SEND. The data is reported using “preferred” terminology of the toxicologists. These do not match the study report’s dose groupings - see 
- Study Reports are signed by the Study Director. *The SEND Dataset must also be signed by the Study Director.* Because the SEND data and study reports are constructed using disparate processes they need NOT be consistent - to prove that they are consistent is **difficult, expensive, and time consuming**

Common Causes for Data Inconsistency in SEND Datasets

There are several points in the process of generating SEND datasets where issues related to consistency, quality and sufficiency can occur. Examples include:

- Establishing the SEND trial design domains from the study protocol (TS, TA, TE, TX, EX).
- Transforming data terminology from LIMS terminology to SEND Controlled Terminology.
- Standardizing qualitative findings into SEND.
- Merging laboratory findings data from multiple providers into a single domain (LB for example).
- Harmonizing naming conventions, data and terminologies from multiple data providers.
- Re-packing required reportable data into SEND tabulation format.
- Generating the Define.xml and Study Data Reviewer's Guide (SDRG).

Except for the extraction from the LIMS systems, all of these processes are different from those used for tabulating data to support the study report. Even the process of LIMS extraction is fraught with risks.

For example, data managers may not extract all of the data from the LIMS system required to create the SEND dataset such that it is equivalent to the data in the study report. Transformation of LIMS terminology to SEND controlled terminology (CT) may result in errors due to the lack of familiarity of the study director and other toxicologists or pathologists with the SEND CT.

Finally, since many sponsors are unaware that the same data can be modeled differently by their different CROs, they may not question decisions made by CROs until it is too late in the process. Ultimately, accountability for the SEND dataset rests with sponsors, as they are the ones signing off on the submission to the FDA.

Trial Design Variations

The study design, sponsor defined groups, and group summary data in the study report are designed for readability by a toxicologist or reviewer. The trial design of a SEND dataset is designed to be machine-readable and it is very granular. It takes into account all variations in the trial arm of each dose group leading to more trial sets of fewer and more narrowly similar subjects.

The granularity of the SEND trial design domains makes reconciling subject IDs used across various LIMS systems even more challenging. The increased granularity in SEND trial design domains compared to sponsor groups can introduce further challenges when trying to ensure the SEND dataset is equivalent to the study report, and for conducting data quality checks.

Terminology Differences

CROs and labs are familiar with the terminology in their study reports and in their LIMS systems. However, SEND controlled terminology (CT) can be unfamiliar, and converting LIMS data to SEND can inject new errors or cause difficulties when comparing the SEND data to the study report. This increases the quality assurance burden for both CROs delivering the standardized SEND data, as well as for sponsors receiving them.

Standardizing Qualitative Findings Domains

Histopathology domains such as MA and MI are not adequately covered by CDISC CT, but the modifiers and qualifiers must be parsed into a form not found in the LIMS data. Without a semantically enabled tool, accompanied by expert human oversight, this process is prone to generate inconsistent incidence summaries between the SEND dataset and study report. This in turn can alarm the reviewers and delay the review process.

Ensuring that Required Reportable Data is Reported in SEND

If the business decisions about what domains and variables in SEND are to be reported – or not – is not clear in the SDRG, the resultant data that is prepared for submission will likely not meet reviewer expectations.

Computing Derived Values included in SEND 3.1

SEND IG V3.1 imposes calculations for certain derived values to support machine readability and presentation of data. These values will not be available in LIMS systems, but must be calculated from the collected data.

An example is the calculation of Nominal Day (NOMDY) which is intended to group measurements that may have been taken over a range of study days into *a single nominal day* that can be used for group summary calculation, graphing and tabulation purposes. NOMDY is an example of a calculation made from the date/time stamps in the LIMS system. Calculating these manually can be a source of error that is best avoided by automation.

Although these are not yet required variables, most sponsors will insist this data be provided by CROs or by their internal data providers. It is also important that these be aligned with the presentation of data in the study report.

Approaches for Resolving Data Quality in SEND Datasets at CROs

Consistency, quality, and sufficiency in SEND datasets can be enforced in two basic ways: by “inspection,” or by ensuring it through the process that generates it. Controlling the process is less expensive over time, and will produce better quality SEND datasets.

Since consistency with the study report is a major concern, this process should focus on a single data repository, or source of truth, with a familiar terminology that is easily accessible and that may be repurposed for SEND datasets, tabulations for study reports, or interim study data reporting and monitoring. This is a function that simply *cannot* be satisfied by a GLP LIMS system.

By Inspection

The gold standard of a study is the study report generated by a Principal Investigator toxicologist and the study team. In order to compare SEND datasets to this standard, it is necessary for these datasets to be read and available to a qualified toxicologist who can then re-constitute the trial sets to represent the same sponsor-defined groups with exclusions and derive the summary data.

In practical terms, this is still a very time and labor-intensive task (requiring additional FTEs) because of the number of groups, number of visit days, and the number of findings in a typical study. However, this is precisely what a reviewer will be able to do using the FDA NIMS data visualization and analysis tool, ToxVision™.

This inspection can be performed by the pharma sponsor independent of the CRO. The use of ToxVision available from PointCross for data QA by sponsors or their CROs will reduce the time and effort to perform such inspections.

By Process

CROs and labs have a well-established process to extract collected data and prepare them in a form that allows them to tabulate and print the individual subject data into tables that will become part of the study report. In addition, they provide the ability to pivot and generate datasets for any specified sponsor-defined group for any set of visit days for a specific lab test or finding, so that group summaries may be calculated for the summary report sections.

This data can be directly imported into a “single source of truth” reportable data repository, with a universally accessible data model (UDM), and with a standard global terminology, that can then be automatically transformed to the preferred terminology of the toxicologist or the SEND CT. The required tabulations and the CDISC SEND datasets can then be automatically generated directly from this single truth reportable data store. This approach is summarized in the figure below which also includes a preceding step to specify the study data packages required from each provider at the outset as discussed in the next section of this paper.

By Planning and Specifying the Study Data Plan

CROs and other data providers begin data collection after completion of the study planning process and acceptance of the protocol. During this early stage of the study, sponsors and CROs should make

decisions about what portions of the study will be included in the SEND dataset, and how to use this information to make better informed decisions about how the protocols are applied into their data collection system. By doing this, sponsors and CROs can ensure that all contracted labs will collect all of the data necessary to make a complete SEND dataset.

We recommend that a Nonclinical Study Data Specification (NSDS), consistent with the protocol, be generated by the Study Director's team at the study start. This will ensure that all data collected during the study, or as modified by protocol amendments, will be present in the final SEND dataset. PointCross provide a software solution, NSDS, to generate master study data specifications that define all components of the study, including DMPK/TK, histopathology, and other types of data.

Use of NSDS provides the *least cost, least risk, and maximum quality assurance* while making periodic interim monitoring of studies easier. In this approach SEND datasets are completed contemporaneously with the study reports.

