PharmaSUG 2025 - Paper SS-348

Common Issues in BIMO Clinical Site Dataset Packages

Michael Beers, Pinnacle 21 by Certara

ABSTRACT

When preparing a submission package for the Bioresearch Monitoring (BIMO) Clinical Site (CLINSITE) Dataset, it is important to ensure compliance to the FDA's specifications and guidance. It is often the case, however, that issues exist with the CLINSITE dataset and the associated documentation. This paper will review some of the most common issues seen across the industry, and how the issues should be addressed.

INTRODUCTION

WHAT IS BIMO AND WHY IS IT IMPORTANT?

The FDA’s Bioresearch Monitoring (BIMO) program is used to ensure the quality and integrity of data from clinical trials and protect the rights of and safety of trial participants. This is done through inspections and audits of the entities involved, such as clinical investigators, sponsors/CROs, institutional review boards, laboratories, bioequivalence facilities, and clinical research organizations.

There are a few different things that a sponsor needs to provide to facilitate the agency’s selection of sites to inspect, such as clinical study-level information, subject-level data line listings by clinical site, and a summary-level clinical site dataset. The scope of this paper is to focus just on the issues related to the Summary-Level Clinical Site Dataset (CLINSITE). The FDA’s BIMO Technical Conformance Guide (TCG) also provides specifications, recommendations, and general considerations for Clinical Study-Level Information and Subject-Level Data Line Listings by Clinical Site, but these will not be discussed in this paper.

Due to finite resources and limitations of time, the FDA uses an approach to efficiently determine which sites to conduct inspections based on certain factors, which correspond to some of the data requested as part of the CLINSITE dataset. The CLINSITE dataset is then used by the FDA’s clinical investigator site selection tool.

WHAT IS A BIMO CLINSITE DATASET?

The Summary-Level Clinical Site Dataset (CLINSITE) summarizes data for the clinical investigator sites from all major/pivotal studies that are used to support safety and efficacy. It is a single dataset for the application.

The CLINSITE dataset is used by the FDA’s clinical investigator site selection tool to facilitate selection of sites for inspection. Since this dataset is used by this agency tool, it is important to strictly adhere to the specifications listed in the FDA’s guidance, to avoid any issues when the dataset is loaded into and used by the tool. Care should be taken to avoid incorrect implementation of the dataset; any issues should be identified and corrected prior to submission.

BIMO CLINSITE GUIDANCE TO FOLLOW

Preparers of CLINSITE datasets should follow the guidance provided in the FDA’s Standardized Format for Electronic Submission of NDA and BLA Content for the Planning of Bioresearch Monitoring (BIMO) Inspections for CDER Submissions, which was finalized December 2024. This document states that: *Technical specifications associated with this guidance are provided as a separate document and are updated periodically: Bioresearch Monitoring Technical Conformance Guide.*

The most current version of the Bioresearch Monitoring Technical Conformance Guide (BIMO TCG) is version 3.1 at the time of this writing. The BIMO TCG provides specifications regarding the organization and formatting of the dataset, the data it should contain, the documentation that should accompany it, for which types of studies should or should not be included, the expected location of the dataset/documentation in the eCTD structure, and examples of the dataset to refer to.

MISSING INFORMATION

The first category of common issues seen in CLINSITE datasets is missing information. This includes missing documentation, missing variables, and missing data values.

MISSING DOCUMENTATION

A critical component of a BIMO submission to the FDA is the documentation that accompanies the CLINSITE dataset, which should clearly describe the contents of the dataset, including the origin and derivation of the variables in the dataset.

Missing define.xml

Occasionally, only a clinsite.xpt file is submitted, with no define.xml (and associated stylesheet). This is problematic, because:

a) The define.xml is specifically requested by the FDA, as the BIMO TCG states: *The summary-level clinical site dataset should be accompanied by a data definition file.*

b) There are several variables in the CLINSITE dataset for which information needs to be provided in the define.xml. Some examples include:

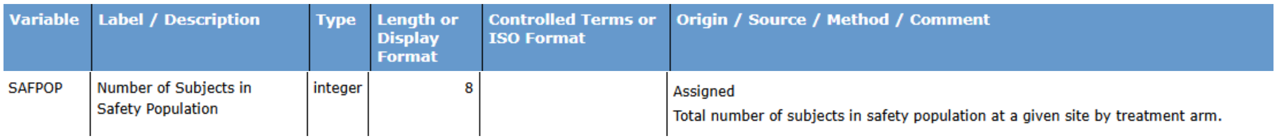
    a. For the Safety Population (SAFPOP) variable, an explanation is needed for how subjects that transferred from one site to another were handled (if applicable),

    b. For Primary Efficacy Population (EFFPOP), the analysis population flag variable (Full Analysis Set, Per Protocol, Intent to Treat, etc) that was used to determine this population needs to be specified, and

    c. For Treatment Efficacy Result One (TRTEFFR1) and Treatment Efficacy Result Two (TRTEFFR2), the analysis datasets and variables that were used to derive these variables should be specified.

These are just some examples of why a define.xml is critical and should always be provided. The absence of this document results in lack of clarity of where exactly the CLINSITE information comes from, as sponsors' each have their own differences when providing this information.

In addition, the comments/methods in the define.xml should be specific as well. Below is an example of a very general comment for the SAFPOP variable. For this study, it was seen that all records in the CLINSITE dataset had SAFPOP = 0, which seems unusual. This comment in the define.xml does nothing to show how the SAFPOP variable was determined or to explain why this situation exists (and there was no reviewer's guide provided to clarify). Care should be taken to provide transparency into variable origins, comments, and methods.



Variable	Label / Description	Type	Length or Display Format	Controlled Terms or ISO Format	Origin / Source / Method / Comment
SAFPOP	Number of Subjects in Safety Population	integer	8		Assigned Total number of subjects in safety population at a given site by treatment arm.

Missing Reviewers Guide

It seems relatively common for sponsors to not include a BIMO Data Reviewer's Guide (BDRG) file in the BIMO submission package, as the BIMO TCG doesn't (at the moment) state that it should be included. The recommendation, however, is to provide this document, for the following reasons:

a) The BIMO Technical Conformance Guide does mention using it to further explain and document information in the CLINSITE dataset, although it is clearly not a requirement, as it is discussed in terms of *"…the BIMO Data Review Guide when one is provided…"*.

b) The BIMO TCG provides a recommendation for a template, which is the template provided by PHUSE.

c) The PHUSE BIMO Data Reviewer's Guide provides a section to explain any conformance issues that exist for the CLINSITE dataset. As with any types of data (SDTM, ADaM, SEND, etc), if any conformance issues exist (and cannot be corrected), it is always best to explain these issues in the reviewer's guide preemptively, to potentially avoid any possible delays in submission timelines.

MISSING VARIABLES

Occasionally preparers of CLINSITE datasets drop certain variables. It seems that is typically due to the variable(s) not being applicable.

An example of this situation can be found with the Censored Observations in SAFPOP (CENSOR1) and Censored Observations in EFFPOP (CENSOR2) variables. The CENSOR1/CENSOR2 variables are only to be populated when the Primary Endpoint Type (ENDPTYPE) variable contains the value "time to event". The guidance for these variables state: *If not applicable, leave blank.*

Other times, although less frequent, preparers drop critical variables such as Primary Endpoint (ENDPOINT), ENDPTYPE, Treatment Efficacy Result for SAFPOP (TRTEFFR1), Treatment Efficacy Result for EFFPOP (TRTEFFR2), etc.

The recommendation is to include all of the variables specified in the BIMO Technical Conformance Guide. If a variable is not applicable, it should still be included and guidance should be followed (such as leaving it blank, or using values of 'NA', whichever is specified in the TCG for that variable). The define.xml should include the reason why the variable is blank (or 'NA') for all records.

MISSING DATA VALUES

ARM Variable

When Description of Planned Treatment Arm (ARM) is left null, it is usually due to there only being screen failure subjects at a site. In this case, ARM is null, but Number of Subjects Screened (SCREEN) is populated with a value. ARM should not be left null in this case however, as the BIMO Technical Conformance Guide states: *When no arm or treatment group is available due to only screen failure subjects at site, use label "Screen Failure."*

Other Variables

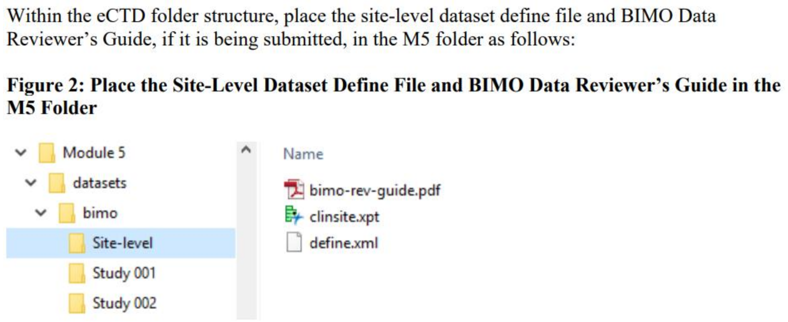
Examples of other variables where missing data values are common include:

- CITY, STATE, Postal Code (POSTAL)
- Financial Disclosure Amount (FINLDISC)
- Investigator Last Name (LASTNAME), Investigator First Name (FRSTNAME)

It is assumed that when this information is not provided, it is just an incorrect implementation, or the information is unable to be obtained for some reason. The recommendation for this situation, if effort to find and populate this information is unsuccessful, is to note this in the define.xml and provide a detailed explanation in the reviewer's guide.

UNEXPECTED LOCATION IN ECTD

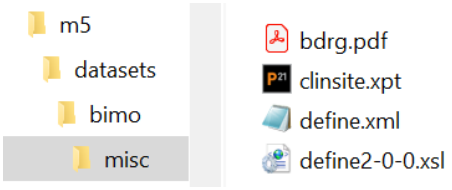
The BIMO Technical Conformance Guide provides guidance on how exactly the CLINSITE dataset and documentation should be represented in the eCTD. It states to use the following folder structure in the eCTD to store the CLINSITE file and documentation:



Within the eCTD folder structure, place the site-level dataset define file and BIMO Data Reviewer's Guide, if it is being submitted, in the M5 folder as follows:

Figure 2: Place the Site-Level Dataset Define File and BIMO Data Reviewer's Guide in the M5 Folder

However, it is common for preparers of CLINSITE submission packages to use other (unexpected) folder structures to store these files instead. One commonly used eCTD folder that differs from guidance is this:

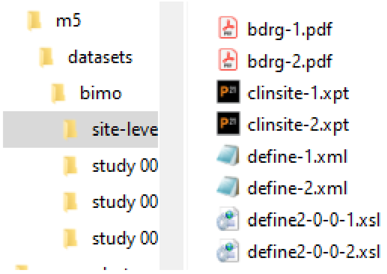


Instead of using a folder named 'site-level', the folder has been named 'misc'.

Other examples of eCTD folder structures that preparers sometimes use are:

- m5/datasets/bimo/misc/site-level
- m5/datasets/bimo/listings
- m5/datasets/bimo/misc/*study* (where separate CLINSITE datasets/documentation are created for different studies, and put in separate folders labeled with study names)

An even worse example of incorrect implementation of the eCTD folder structure is having multiple versions of the CLINSITE dataset and multiple versions of the associated documentation in the same folder, but each file appended with something like "-1" and "-2", etc., as is shown in this screenshot:



The recommendation for this issue is to use the exact folder structure listed in the BIMO TCG, so that interested parties do not have to search for information, to avoid confusion of why it differs, or to have multiple datasets when a single dataset was requested, and to avoid complications where automated processes (using expected folder structure) might exist.

INCORRECT USAGE OF VARIABLES

USING OLD VERSIONS OF BIMO GUIDANCE

The latest version of BIMO TCG should always be used when preparing a CLINSITE dataset and documentation for submission. The latest version, at the time of this writing, is 3.1 (or 3.0, as there were no changes to the specification in 3.1). The specifications for the CLINSITE dataset have been changed in the subsequent version of BIMO TCGs, so it is important to be aware of the latest guidance when preparing a CLINSITE dataset.

Changes from BIMO 1.0 to BIMO 2.0:

- STUDYTL (Study Title) variable changed to TITLE
- SPONNAME (Sponsor Name) variable changed to SPONSOR
- COHORT (Description of Planned Cohort) has been added as a new variable
- SITEEFFE (Site-Specific Treatment Effect) and SITEEFFS (Site-Specific Treatment Effect Standard Deviation) variables have been removed
- PROTVIOL (Number of Protocol Violations) variable has been removed
- IMPDEV (Number of Important Protocol Deviations) and NOIMPDEV (Number of Non-Important Protocol Deviations) have been added as new variables

Changes from BIMO 2.0 to BIMO 3.0:

- EFFPOP (Number of Subjects in Efficacy Population) has been added as a new variable
- TRTEFFR (Treatment Efficacy Result) and TRTEFFS (Treatment Efficacy Result STD) variable have been removed
- TRTEFFR1 (Treatment Efficacy Result for SAFPOP) and TRTEFFR2 (Treatment Efficacy Result for EFFPOP) have been added as new variables
- CENSOR (Number of Censored Observations) variable has been removed
- CENSOR1 (Censored Observations in SAFPOP) and CENSOR2 (Censored Observations in EFFPOP) have been added as new variables
- COUNTRY variable now uses the Geopolitical Entities, Names and Codes (GENC) codelist

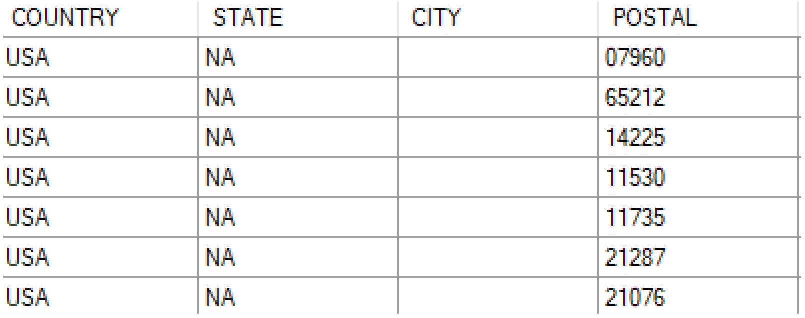
INVALID DATA VALUES

Populating CLINSITE variables in a way that differs from BIMO TCG specifications and guidance is unfortunately a somewhat frequent occurrence.

STATE Variable

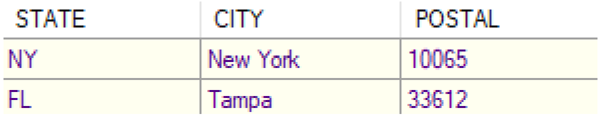
It is not uncommon for the STATE variable to have nonconformant values. Common scenarios for when STATE is not populated correctly include:

- Missing values when COUNTRY = USA
  - STATE should always be provided when the COUNTRY is USA
- COUNTRY = USA, but STATE is populated with “NA”
  - In this example below, we see that USA = COUNTRY, STATE = NA, CITY = null, POSTAL is populated. Note: STREET is also populated. This just seems like an incorrect implementation, as STATE can obviously be populated with a valid value if a postal code is available.



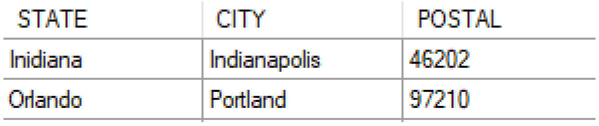
COUNTRY	STATE	CITY	POSTAL
USA	NA		07960
USA	NA		65212
USA	NA		14225
USA	NA		11530
USA	NA		11735
USA	NA		21287
USA	NA		21076

- STATE values not using GENC codelist values
  - The most common scenario here is the two-letter state abbreviation is used instead of the GENC subdivision unabbreviated preferred name, as is seen in this example:



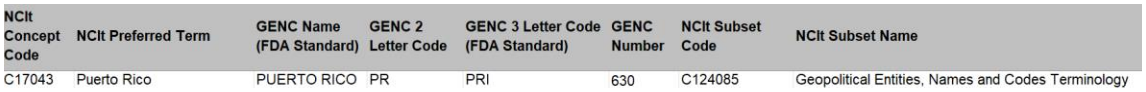
STATE	CITY	POSTAL
NY	New York	10065
FL	Tampa	33612

- Incorrect values in the STATE variable, such as city mistakenly used instead, or misspellings, or leading spaces
  - This example shows a misspelling for STATE, and a value of city used mistakenly:



STATE	CITY	POSTAL
Inidiana	Indianapolis	46202
Orlando	Portland	97210

- Populating the STATE variable with value of Puerto Rico
  - Puerto Rico is listed on the GENCCOUNTRIES tab of the GENC file, and has a GENC 3 Letter Code (FDA Standard) = PRI. This is how Puerto Rico should be represented in the CLINSITE dataset, with COUNTRY = PRI and STATE left null.



NCIt Concept Code	NCIt Preferred Term	GENC Name (FDA Standard)	GENC 2 Letter Code	GENC 3 Letter Code (FDA Standard)	GENC Number	NCIt Subset Code	NCIt Subset Name
C17043	Puerto Rico	PUERTO RICO	PR	PRI	630	C124085	Geopolitical Entities, Names and Codes Terminology

Variables that should be integer

There are variables listed in the specifications as integer, such as:

- Variables that show counts of subjects like DEATH, DISCSTUD, DISCTRT, SAFPOP, EFFPOP, SCREEN, or
- Counts of deviations or events like IMPDEV, NOIMPDEV, NSAE, SAE, or
- Other information like SPONCNT, IND, NDA.

Somewhat frequently, preparers attach ".0" to these values. This causes a number of different validation rules to fire, could possibly cause issues loading or using the data since integer values are expected, and it just generally doesn't make sense, as decimal points to do not apply to values for these variables.

SCREEN Variable not populated according to guidance

The SCREEN variable is sometimes populated by preparers in a way that differs from guidance in the BIMO Technical Conformance Guide, which is: *Total number of subjects screened (and consented) at a given site (overall number per site as subjects have not yet been assigned to treatment arm).*

However, it is sometimes seen that the values in SCREEN differ for records for the same SITEID. It seems that in this situation preparers use the number of screened subjects for each ARM, which differs from the guidance.

Another situation where SCREEN is incorrect, is when screen failure records are not included in the count for SCREEN. Sometimes this situation is connected with the scenario above…because if SCREEN values are populated per ARM (instead of SITE), it is likely that the preparer won't include a separate record for ARM = Screen Failure, and therefore the number of screen failure subjects won't be included in SCREEN for any records for that SITEID in the CLINSITE dataset.

SAFPOP is less than DISCSTUD

There are certain variables that should be populated with the "Number of subjects in the safety population…" that meet a certain condition. For example, the BIMO TCG, for the Number Subjects Discont. Study (DISCSTUD) variable, states: *Number of subjects in the safety population who discontinued from the study by treatment arm at a given site.*

However, it is often the case that prepares of CLINSITE datasets, for these types of variables, will provide the number of subjects that meet the condition, but ignore the "…in the safety population…" portion of the guidance. Therefore, the number of subjects listed in these variables might be greater than the number of subjects in the safety population, which might introduce confusion.

Other examples of variables that should be limited to subjects in the safety population include: DISCTRT (Number Subjects Discont. Study Treatment), NSAE (Number of Non-Serious Adverse Events), SAE (Number of Serious Adverse Events), DEATH (Number of Deaths), IMPDEV (Number of Important Protocol Deviations), and NOIMPDEV (Number of Non-Important Protocol Deviations). The recommendation here is to follow the guidance exactly, and where specified in the guidance, limit the counts to subjects that are in the safety population.

TRTEFFR1 value is populated when SAFPOP=0 &

TRTEFFR2 value is populated when EFFPOP=0

Typically, these issues occur when a preparer populates the Treatment Efficacy Result for SAFPOP (TRTEFFR1) or Treatment Efficacy Result for EFFPOP (TRTEFFR2) variables with a value of 0. These variables are meant to contain actual efficacy results, though, so populating these variables with values of 0 may lead to confusion of whether or not the 0 should be interpreted as a result, or as the lack of a result since SAFPOP/EFFPOP equals 0.

The recommendation for this issue is to only populate these variables with actual results, and to not use values of 0 to indicate a lack of a result.

TRACEABILITY ISSUES

It is important to be sure that information provided in the CLINSITE dataset can be traced back to the appropriate SDTM and ADaM datasets for each study.

MISSING SITES

All sites from each study must be listed in the CLINSITE dataset. Occasionally it is seen that sites are missing from CLINSITE, and the reason is likely that there were no randomized subjects from those sites for the study. For these sites, they still must be listed in CLINSITE, likely just having one record for that SITEID, and the BIMO TCG, for the ARM variable, states: *When no arm or treatment group is available due to only screen failure subjects at site, use label "Screen Failure."*

COUNTS NOT MATCHING

Care should be taken to ensure that variables are populated with correct counts from the SDTM or ADaM datasets, and that the documentation very clearly explains where the information comes from.

An example is the Number of Non-Serious Adverse Events (NSAE) variable. The guidance from the BIMO TCG, for this variable, states: *Total number of nonserious adverse events at a given site by treatment arm for subjects in the SAFPOP. This value should include multiple events per subject and all event types (i.e., not limited to only those that are deemed related to study drug or that are treatment emergent events). When events with the same preferred term have occurred on different dates for a subject, each event should be counted separately in event count.* Occasionally, the value in the NSAE variable isn't just a straightforward count from the SDTM AE domain, and if there is a valid reason for filtering out some events based on the data collection for the study, this needs to be very clearly defined in the define.xml and possibly explained in the reviewer's guide.

Many variables have somewhat obvious predecessors in SDTM or ADaM datasets. Examples are the Number of Subjects Discont. Study (DISCSTUD) and Number Subjects Discont. Study Treatment (DISCTRT) variables. This information is typically already included in the ADSL dataset in the End of Study Status (EOSSTT) and End of Treatment Status (EOTSTT) variables, respectively. If the derivation/origin used to populate these variables differ from the obvious datasets/variables from the SDTM or ADaM data, this should be clearly identified in the define.xml.

CONCLUSION

PROBLEM

There are often issues with BIMO CLINSITE datasets. These issues tend to include:

- Missing documentation
- Missing variables
- Missing data values
- Unexpected location in the eCTD
- Using old versions of BIMO guidance
- Invalid data values
- Missing sites
- Counts not matching study-level data

As the CLINSITE dataset is used by the FDA's clinical investigator site selection tool, strict adherence to specifications and guidance is important in order to avoid issues with loading and using the dataset, to avoid hindrances to the agency in this important work of safeguarding trial participants through site investigations. In addition, adherence to specifications and minimizing issues with the CLINSITE is likely to reduce the possibility of delays in submission timelines (having to potentially deal with potential information requests from the agency).

SOLUTION

Preparers of CLINSITE datasets should do the following to ensure the dataset and documentation meet FDA's expectations:

- Develop a standard process for creating these submission deliverables,
- Strictly adhere to specifications and guidance from the FDA,
- Clearly document the origins and derivations of the data,
- Verify that these deliverables are of high quality with no issues

Regarding the last point, there are no officially published validation rules from the FDA at the time of this writing, however Pinnacle 21 Enterprise does support validation of CLINSITE datasets against the specifications listed in the BIMO TCG, drastically reducing the amount of time it might take to identify and correct issues prior to submission of the CLINSITE dataset and documentation to the FDA.

REFERENCES

U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER). December 2024. Standardized Format for Electronic Submission of NDA and BLA Content for the Planning of Bioresearch Monitoring (BIMO) Inspections for CDER Submissions.

U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER). September 2024. BIORESEARCH MONITORING TECHNICAL CONFORMANCE GUIDE. Version 3.1.

U.S. FOOD & DRUG. Bioresearch Monitoring Program Information. https://www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/fda-bioresearch-monitoring-information/bioresearch-monitoring-program-information.

PHUSE. Bio-research Monitoring Data Reviewer's Guide. https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Optimizing+the+Use+of+Data+Standards/BDRG+V3.0+Package.zip

CONTACT INFORMATION

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

Michael Beers
Pinnacle 21 by Certara
michael.beers@certara.com