PhUSE US Connect 2018

Paper DS12

CRF Design for Data Standards

David A. Scocca, Rho, Inc., Chapel Hill, NC, USA

Abstract

The CRF is the primary source for the standardized data that goes to regulatory agencies. Ideally, a CRF can be designed to support the requirements and expectations for standardized data in ways that go beyond simply using standardized forms such as CDASH. This paper discusses some specifics including: capturing key results and timepoints necessary to identify study epochs, elements, and dispositions; structuring forms for unscheduled and rescreening visits to clearly identify the source of observations; collecting or recording sufficient context to avoid ambiguous data; and focusing data collection on fields that will actually be used for analysis.

Introduction

The Study Data Tabulation Model (SDTM) is the standardized data structure in which you submit tabulated clinical data to a regulatory agency (FDA or PMDA).

Understanding its requirements reduces the cost of converting from the raw clinical data to the standardized SDTM data. CDASH is a good start but is not complete or sufficient; it provides examples of how to collect data, but does not provide as much guidance about what data should be collected.

Controlled Terminology

Standardization can apply to both the structure of the data and the values in the data; controlled terminology is about standardizing the values in the data. Wherever possible, the CRF should use the standard terminology that will eventually appear in the tabulated data. CDISC maintains lists of controlled terminology through the National Cancer Institute (NCI). A controlled terminology release is organized into codelists, each of which is made up of items or terms. These lists are available in multiple formats from:

- https://www.cancer.gov/research/resources/terminology/cdisc

The primary controlled terminology list for clinical trial data is the SDTM list, which as of March 2018 consisted of 710 codelists containing 25,270 items. Additional codelists that are specific to data collection (CDASH) and analysis (ADaM) are available, but these controlled terminology lists do not duplicate the codelists and items found in the main SDTM list. So for data collection with CDASH-based forms, both the CDASH and the SDTM controlled terminology lists should be used.

Among the kinds of lists contained in the SDTM controlled terminology are:

- Standard values for several specific data fields, including race and ethnicity for demographics; severity, location, action taken, and outcome for adverse events; and dose form, frequency, and route for medications.
- Standard results for subject dispositions, ECG findings, oncology response, reference range indicators, and other results.
- Standard categories for questionnaires, functional tests, disposition events, and other kinds of data.
- Units, including a general list of units as well as specialized lists of units for age, vital signs, and pharmacokinetic results.
- Relationships between a time point and a reference period, such as the duration of the study or the period during which the subject was receiving study treatment.
- Test codes and names for many different evaluations and instruments. These can be used as variable names and labels when collecting data horizontally that will be stored vertically in SDTM.

Many codelists are extensible. If a codelist does not contain either the term you want to use or a synonym for that term, you can extend the codelist with the additional term. However, if there is a synonym for your term on the codelist, your CRF should use the value from the codelist rather than extending the codelist. For example, you should represent milliliters or cubic centimeters with “mL” from the unit codelist rather than extending the list with “cm3” or “cc”.

Timing — Epochs

The concept of a trial epoch is specific to the CDISC data standards, and is defined as “a period of time that serves a purpose in the trial as a whole” (SDTM IG 3.2, section 7.1.2). A subject’s participation in a particular trial epoch does not depend on that subject’s treatment assignment; the epoch for a particular data point can be determined even

while the trial is blinded. The SDTM controlled terminology list for EPOCH illustrates the intent: SCREENING, BASELINE, TREATMENT, FOLLOW-UP, RUN-IN, WASHOUT, INDUCTION TREATMENT, CONTINUATION TREATMENT, BLINDED TREATMENT, OPEN LABEL TREATMENT, and LONG-TERM FOLLOW-UP.

Not all of these will be used in a single study, of course; many studies are simply broken down into SCREENING, TREATMENT, and FOLLOW-UP. Because trial epochs are intended to be non-overlapping, a trial that used the epochs BLINDED TREATMENT and OPEN LABEL TREATMENT would not have an overall TREATMENT epoch. The SDTM codelist of trial epochs is extensible; a crossover trial might use the epochs TREATMENT PERIOD 1 and TREATMENT PERIOD 2 to identify the two periods of treatment.

The FDA has indicated that it expects the EPOCH variable to be used for any “clinical subject-level observation” (FDA SDTCG 4.1, section 4.1.4.1). Because trial epoch is not typically collected on the CRF, it is important to consider how it will be derived in the process of tabulating the raw clinical data.

The first consideration is identifying the specific points in the study at which a subject transitions between trial epochs. For example, a subject might transition from SCREENING to TREATMENT at the time of the first dose, and might transition from TREATMENT to FOLLOW-UP at 24 hours after the time of the last dose. In those cases, the CRF must capture the relevant dose times in order to accurately assign events to epochs. Wherever an event marks the transition between trial epochs, it is necessary for the CRF to capture the date of that event. If the time component will be necessary to assign events to epochs, then it is necessary to capture the time as well.

A second consideration is associating an epoch with each observation. If the transition event dates and times are accurately captured on the CRF, then it is possible to derive the epoch from the date and time of the observation. If there are observations or events on the same day as a transition event, collecting times is necessary to properly assign epoch in this way.

While using dates and times to derive epoch is possible when times have been collected, it is far easier to simply define study visits on the CRF so that each visit falls entirely within a single epoch. In cases where the subject has initial treatment with pre-treatment and post-treatment evaluations in a single visit to the clinic, this means setting up the CRF so that the day consists of two visits, with the second visit beginning with the administration of the initial treatment. For example, the visits might be named “Day 1 Pre-Treatment” and “Day 1 Treatment”.

Timing — Dispositions

One of the uses of the trial epoch is to structure the presentation of dispositions. In a simple trial, with a single treatment phase, you can collect a disposition that applies to the subject’s participation in the entire study. In a study with long term follow-up after treatment, or where there are multiple treatment epochs, it may be preferable to collect a subject disposition at the end of each epoch. This will indicate whether the subject completed participation in that epoch or provide a standardized reason why the subject did not complete participation in that epoch. The standardized reasons are found in the controlled terminology; not every reason will apply in every epoch, so it can be necessary to have distinct disposition forms for each study epoch.

For example, at the end of participation in the screening epoch you might offer a choice of FAILURE TO MEET RANDOMIZATION CRITERIA but not NON-COMPLIANCE WITH STUDY DRUG, but at the end of the subject's participation in the treatment epoch you would make the opposite decisions, while after a follow-up epoch you would offer neither option. Disposition forms often have names specific to the associated epoch, e.g. “End of Screening” (or an assessment of whether subject qualifies for randomization), “End of Treatment” (if there is a significant follow-up period), or “End of Study”.

For disposition, the important question is whether the subject completed participation in the epoch as defined by the protocol. In a pain study, for example, the protocol might call for ending treatment after completing a certain number of doses, but stopping earlier if the subject is no longer in pain. In that case, asking if the subject had completed the full number of doses does not accurately identify all subjects who completed the treatment phase as defined by the protocol.

Timing — Unscheduled Visits

Unscheduled subject visits are common in clinical trials. To usefully incorporate the data from these visits into SDTM, the CRF must capture sufficient context about the timing of and reason for the unscheduled visit. The CRF must also distinguish between multiple unscheduled visits in such a way that observations from each of those visits can easily be grouped together.

Because EPOCH should be populated for observations in SDTM data, unscheduled visit pages should collect enough information to easily identify the trial epoch. In a simple screening/ treatment/ follow-up trial, this could be as easy as asking whether the subject has started or completed the study treatment; in a more complex trial, identifying the latest scheduled visit completed and the next scheduled visit expected may allow easy derivation of EPOCH.

The SDTM subject visits domain (SV) includes the variable SVUPDES that can be populated with the reason for an unplanned visit; for this field to be useful, the CRF must capture the reason for the visit. While there is no defined codelist of standard reasons for unplanned visits, the collection of this reason will be more straightforward if you provide a selection list of possible reasons.

Also, the structure of the CRF should make it possible to know what was collected during the same unplanned visit. If vital signs and lab tests were both collected on an unplanned visit, it should be possible to assign the same VISIT and VISITNUM to the associated records in VS and LB. To this end, it is useful if the electronic CRF is configured so that each unscheduled visit is recorded on a distinct section of the CRF, with enough distinct unscheduled "visits" available so that no subject uses them all.

Timing — Rescreening

In trials in which subjects may be rescreened, rescreening visits can be treated as a special kind of unscheduled visit, where the CRF pages at rescreening visits reflect the forms that are to be completed when a subject returns for another screening visit. In tabulating the data, it is far easier if rescreening is treated as an additional unscheduled visit for an existing subject than if subjects undergoing rescreening are treated as new subjects. As with other unscheduled visits, it is useful if the electronic CRF offers enough distinct rescreening visits to accommodate the maximum number of rescreening visits allowed for the trial.

Collecting Dates and Times

In addition to epoch transitions, there are a couple of additional considerations to keep in mind when collecting dates and times on the CRF.

Records for events should be sure to capture the date of the event, or the start and end dates for a multi-day event, particularly when this differs from the visit date at which the event was recorded. The forms should clearly distinguish visit dates from event dates. For findings data, make sure to capture the most relevant date--for example, the date on which a radiography image was taken rather than the date on which it was interpreted.

Times should be collected along with dates when it will make the data more useful or easier to interpret. If you have multiple records on a single day, as in the case of vitals taken before and at intervals after dosing, or sample collections for pharmacokinetic studies, then the times are clearly necessary. But even if a given evaluation is only done once on a particular day, if there is any chance you will need to determine the order of distinct events and evaluations on that day, then collecting the time makes it possible.

Free Text Fields

Free text is in some sense the opposite of standardized data. To analyze free text data, it is necessary to somehow standardize the values. With the fields that identify adverse events or medications, the range of possible values is so broad that you essentially have to allow the entry of free text and then have an experienced medical coder standardize any values that do not appear in the appropriate dictionary. While medical coding is the best practice for handling free text, it takes time and effort and can only be used on fields for which an appropriate dictionary is available.

An alternative to having a professional coder is to structure the CRF to obtain both a free text and a standardized response. For example, a disposition form might both allow free text to describe the reason for non-completion and require the selection of a value from the controlled terminology list. This approach likely requires data management review to be sure the free text is consistent with the selected term, but it will guarantee that standard values are available for analysis while capturing additional context that may be useful when looking at specific cases or crafting narratives.

When offering a free text field to go along with a value selection, it is not necessary to enable the free text for every possible selected value. A CRF value selection field may have an "Other, Specify" choice which activates a free text field to specify the alternative value. An "Other, Specify" option should only be added to a selection field when the existing values reflect all the choices that will be needed for analysis. In addition, data management should review the specified values that are entered to identify cases where a user should have used one of the other values and cases where the CRF needs to be updated to add an additional choice that was not foreseen when the original list of values was developed. (Note that offering an "Other" option does not require you to offer a "Specify" field. If you do not feel that the information gained from the free text data will be useful, you can simply offer the "Other" with no associated text field.)

This can be a particular issue when CRF values require units of measurement to be specified on the form. Handling free text entry for units can be a very time-consuming part of standardizing the data; because the values in a unit field need to match controlled terminology, any synonyms or alternative capitalizations must be mapped to the correct version of the unit for the standard field. In many cases an additional supplemental qualifier field must be added to capture the raw value as a way of documenting the mapping.

Free text fields specifically labeled as "Comments" are particularly problematic and should be avoided if possible. While SDTM provides the CO domain for the storage of comments (in 200-character chunks), it is rare for this data to be used for analysis. The software systems that analyze free text such as web content or a stream of tweets are generally not using algorithms that produce results appropriate for a clinical study report. Because there is no reliable way to analyze text content for a clinical trial, the CO domain ends up serving as a dumping ground. Someone may search the CO domain for values of particular interest with respect to a single subject, but on the whole it is not used.

The greatest risk of comment fields is that of overlooking data that is important but is incorrectly entered in a comment rather than as part of the appropriate CRF field. Reviewing the comment data and checking its consistency

with the rest of the data—for example, whether an adverse event mentioned in a comment field is also reported on an adverse event form—can require a significant amount of CRA and data manager time and effort.

Unwanted Data

You can improve the efficiency of your data collection and standardization by taking care to collect only the data you actually need. Collecting, cleaning, and tabulating data is expensive—avoid this expense for data you do not actually need. When using standardized forms like CDASH, be sure to review what you actually need. In creating a standardized form, the usual goal is to provide ways to collect a wide range of information. If you do not actually need all this information, you want to remove fields from the standardized form so that your CRF reflects only what you are actually interested in capturing.

Concomitant medication forms are an example of this. Because concomitant medications are an intervention, standard CM forms have many fields including the treatment name, dose, units, route, frequency, form, indication, start date, and stop date. It is good that the form makes all of these fields available and standardized, but if you do not expect to use the information it is better to remove the field from your CRF. For most general concomitant medications, analysis will only require the medication name, start date, stop date, and possibly some form of indication. The medication name will usually be a free text field that goes through medical coding; some additional fields, including route and indication, may be useful to provide context for the medical coder.

For the indication field, it may be better to offer a small list of standardized choices (treat the underlying condition; treat an adverse event, other) rather than a free text field which would also need to be coded.

If there are cases where more specific information is needed for some kinds of medications, consider having two CRF pages with a simpler page for general medications and offering more fields only in the cases where it is required. For example, a pain study might want information about specific doses of non-study pain medications, where the indication (“pain” or “other”) could be used to distinguish whether the drug should be entered on the simpler form or the more detailed form. Similarly, in an oncology trial where subjects have standard chemotherapy in addition to the study drug, you might want to collect more details about the timing and doses of the concurrent cancer treatments than about the other medications the subject happens to take during the course of the trial. Using an indication field with a small number of choices can let you set up validation rules to check that a medication is recorded on the correct form.

Collecting Necessary Context

There are times when an additional field may need to be added to a CRF to make the data easier to interpret. A CRF page for inclusion and exclusion criteria is generally fairly simple but has problems capturing changes to the criteria that frequently take place over the course of the study. Protocol changes that affect inclusion or exclusion criteria may not be simultaneously adopted across all sites. An electronic CRF must either be frequently updated to offer choices based on a site’s current protocol version or provide a place to indicate the version of the protocol under which the inclusion and exclusion criteria were evaluated. (Usually, this would be the same version of the protocol under which informed consent was obtained.)

Conclusion

Because standardized data is what will be submitted to the agency, CRFs should be designed to capture the data required for analysis while making data conversion for submission as straightforward as possible. CRF design decisions can have a major impact on the amount of time and effort required in later parts of the study, both from data managers validating and reviewing CRF data and from programmers and statisticians producing the standardized data. Involving data standards experts in the review of the CRF at study startup can more than pay for itself over the course of the study.

References

FDA Study Data Technical Conformance Guide, v4.1 (March 2018), available from:
https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm

SDTM Implementation Guide, v3.2 (November 2013), available from:
https://www.cdisc.org/standards/foundational/sdtmig

Contact Information

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

David A. Scocca
Rho, Inc.
6330 Quadrangle Drive
Chapel Hill, NC 27517
dave_scocca@rhoworld.com