PharmaSUG 2014 - Paper IB06

Good versus Better SDTM – Why “Good Enough” May No Longer Be Good Enough When It Comes to SDTM

Henry Winsor, Relypsa, Inc., Redwood City, CA

ABSTRACT

While companies are finally making strong efforts to use and provide SDTM, mostly because of serious encouragement from the FDA, there seems to be some misunderstandings about why this work needs to be done. Although originally strictly intended to replace paper Case Report Tabulations with data that is electronically accessible, SDTM is being used – and abused -- with other purposes in mind. Problems arise when people forget what SDTM is for and tailor their implementations for something other than easing the Reviewer’s task.

The author provides a gentle reminder of SDTM’s real purpose in the FDA submission world and some suggestions of how and why to maximize your company’s benefit from having SDTM data sets available for use, preferably long before a CSR or eCTD is complete.

INTRODUCTION

It has been more than ten years since the CDISC Consortium published the first standard for what they call the Study Data Tabulation Model, or SDTM for short. If you can trust Wikipedia, FDA has been allowing sponsors to submit SDTM data since July of 2004, although FDA has yet to actually insist that they will only consider submissions filed that include SDTM data.

Ten years is a pretty long time, even in the world of regulatory submissions, and one would think that there would be little controversy about the standard and that it would be part of the background knowledge set of almost all people working with clinical data in the pharmaceutical/biotech/medical device industry. However, a cursory scan of social media on the topic of SDTM will leave you with a different impression, that people are unsure of the contents of the standard, even asking if they have to implement parts for which they apparently see little use.

Thus, I felt compelled to write this paper, not from any position of authority other than that of a user who has worked for over 20 years in the industry and has been both a user and provider of clinical data for most if not all of that time. Although I do possess a Statistics degree, my work has solely been in the realm of what is referred to as SAS Programming; a person who works with clinical data from the data acquisition stage through the providing of clinical data to others, in both aggregated and non-aggregated form. I have been working with several models of standardized data since long before the CDISC consortium existed, with the idea of making my job and that of others more efficient and less devoted to spending hours of time writing code to account for seemingly arbitrary differences between data aspects in different clinical trials.

WHAT IS SDTM?

It might be easier to start with what SDTM is not. SDTM does not exist to make life easier for people creating analysis data sets, no matter what many people think. This opinion is understandable if one works solely within the average statistics group, as the important work within the group will seem to be oriented towards the outcome of a trial. However, clinical data does not exist only to provide a hopefully meaningful p-value. This is a regulated industry in which we work, and there are plenty of people who use clinical data for things other than assessing the difference between two populations. Among them are the medical reviewers at FDA, who will want to see the whole data picture of the trial, not just the aggregated results that the sponsor has determined will show the efficacy and safety required for an approval. Also, while a FDA statistical reviewer may focus on the results provided by the sponsor, you can be pretty confident that he will want to use the entire data to assess whether the sponsor is appropriately including data from the entire study population or whether there are certain exclusions being done that make the results appear better than they are.

This may come as a complete surprise to some of the audience, but FDA reviewers do not necessarily trust everything a sponsor tells them. With their experience, they know what to look for in the data that can lead them to assess whether the sponsor ran the trial according to the rules that they set up prior to the trial (normally stated in the protocol). Following the rules is very important, as those rules lay the foundation that allows an analyst to actually use statistical methods in analyzing the results. For example, subjects aren’t randomly allocated to different treatments just of the heck of it. Without the assignment process being conducted randomly, the underlying assumptions necessary for inference are no longer met, and the use of statistical methods can no longer be assumed to provide meaningful results.

So, SDTM is merely a description of how to provide clinical study data in a standard format in a submission. It supplies naming conventions and rules for data formatting, so that a reviewer will have a consistent view of the data, no matter what study or company provides the data.

I will emphasize two things about SDTM, that it is supposed to include all of the data and that it should be assumed that somebody will actually want to look at it.

WHAT WAS USED BEFORE SDTM?

One of the pillars of conducting statistical analysis is that there are certain assumptions that have to be present in the data that allow the use of statistical methods. Sadly, these underlying assumptions are usually mentioned briefly in the onset and assumed to be in place thereafter, so people who haven’t been adequately trained in the analysis process often don’t realize how important those rules are. The point is, it isn’t just important what your results are, how you conducted your trial is as important if not more so.

The easiest way for a FDA reviewer to assess how the sponsor conducted the trial is through review of the collected data. While data is made available upon request in copies of each subject’s case report forms, the normal process of providing the data has always been through providing the data in a tabular form, as it is much easier to compare and contrast subjects’ data when there it is on the same page. Back in the Dark Ages, when a huge desktop monitor had 12 inches of diagonal space, the data was provided in what were called Case Report Tabulations or CRTs for short.

If you aren’t familiar with the term, CRTs were printouts of the data where the values were laid out in columns, much like a data listing (and I am going to assume that people still do data listings these days, so I won’t explain the term), except CRTs had to include every variable collected, whether it was used in analysis or not. At its simplest form, this would merely be a data dump of the trial database; although more enlightened companies would make an effort to organize the data into as meaningful a grouping as possible, because CRTs were meant to be read. Layout design did matter; I remember making a few alterations to the design of a few diary listings in a 300 subject trial and reduced the size of the CRTs in one study from over 20,000 pages to less than 5,000, or from four standard boxes of paper down to less than one.

Yes, paper filings could get pretty big and quite unwieldy, encouraging a transition from paper to electronic submissions. As FDA moved towards electronic submissions, it was recognized that the CRTs had to be provided in an electronic form as well. I worked on one filing where we supplied pdf versions of the paper CRTs, indexed by subject number. A nice and easily implemented concept, until your number of subjects was too big to easily find the subject number in the bookmark list. The next step, after the Agency stopped accepting CANDAs as too hard to use, was to request that the data be provided as SAS Version 5 transport files, following a certain set of rules issued in a January 1999 Guidance. CRTs were Item 11 on the Table of Contents; hence they were called Item 11 data sets. Item 11 data sets can be considered the direct ancestor of SDTM; there were general guidelines as to form and content, but the provided specifications were never as detailed as the SDTM specifications.

Shortly after this, CDISC was formed as an independent body and has been taking the lead in publishing data standards for the industry, so FDA got out of the homegrown standards business and marked their seal of approval on the new SDTM standard in July of 2004. SDTM is currently on its fourth version, with the release of version 1.4 in December of 2013.

Although FDA has been recommending use of SDTM since mid-2004, it still isn’t officially a requirement. However, anyone who thinks that there’s no need to make up SDTM data sets for their filing will quickly learn otherwise at the Pre-NDA meeting, where the sponsor’s representatives are supposed to meet with the FDA review team and work out the material handoffs and come to an agreement as to the details of what is expected to be provided. While much of this appears in Guidance Documents, the actual details are up to the reviewers and they can ask to be provided items that are not referenced in the Guidance Documents. The Sponsor doesn’t have to roll over and agree to every little request, but it behooves them to come to an agreement with the review team; otherwise they should look forward to a bumpy review process.

WHY THE FDA WANTS SDTM DATA

FDA reviewers have to work with what is provided to them, and as users of the data, they benefit greatly from the things that SDTM gives them, which is browseable data in an electronic form, with standard names and content, so they don’t have a big learning curve for every study. Think about what it would be like for a reviewer, if he was responsible for reviewing say a five study filing, where one study was supplied in Excel, one showed up as a flat text file, and the rest in some varying version of SAS data sets, with the names for commonly anticipated variables named differently in each study.

Lack of standardization is the reason why the early efforts with CANDAs were a failure; a reviewer had to learn a new environment on new hardware with a new library of tools, with no guarantee that any of that learning would ever be useful again, as there was no guarantee that even two CANDAs from the same sponsor would be remotely close to the same. The reviewers tried and ultimately went back to using paper, as it was easier to use and made them more productive.

With a standard data structure, it is a different world. They know that an adverse event verbatim term will be called AETERM, no matter who provided the data, so they can use their own computers, their own tools and be able to easily share programs between themselves, all contributing to making the reviewer’s job easier, so they can continue to meet the productivity goals that the US Congress sets for them.

WHAT CAN HAPPEN WHEN REVIEWERS DON’T GET SDTM

As stated before, while FDA encourages the inclusion of proper SDTM in a submission, there is no actual requirement. Some view this lack of a requirement as license to supply whatever they feel like doing, and up to now reviewers have been relatively tolerant and made considerable efforts to work with the “SDTM-like” data that they have received. However, anyone who thinks this situation will continue indefinitely should expect some unpleasant surprises in their future.

There has been at least one filing returned to the sponsor marked RTF (for Refusal to File) for SDTM deficiencies that I know of, not to mention more than a few known cases where the sponsor has been requested to provide data reworked to a certain specific standard in information requests, ones that effectively put the review process on hold until the request is satisfied. While it may seem like a prudent business decision at the time to take shortcuts during submission preparation and prepare an abbreviated data product to meet internal timelines, these decisions can come back to haunt you when a reviewer realizes that you haven’t supplied complete SDTM in support of a CSR. Does anyone think that having your work highlighted as the reason for delaying a review is a good career move in this industry?

The expectations of the reviewing community for comprehensive and quality SDTM is only going to grow as they get used to having SDTM data sets for their work, and as other sponsors willingly provide the data to them.

USING SDTM DATA LONG BEFORE YOUR COMPANY IS READY TO FILE

From the previous discussion, it should be obvious why some companies treat SDTM as a filing specific requirement and do it as one of the last things to be done. This view is short sighted for a number of reasons. The primary reason is FDA has recently made it very clear that they expect that SDTM data will be integrated into the CSR workflow, if only to minimize the possibility of having contradictory data sources in the submission. Simply put, when a reviewer finds something in submission data that cannot be explained, like three death Adverse Event reports in the analysis data but four in the SDTM data, the reviewer is going to send a request for an explanation, one that the sponsor is going to have to quickly resolve and explain away, hopefully in a fashion that does not necessitate even more questions.

Furthermore, a reviewer finding a continuing pattern of discrepancies between your data sources is going to cause him to question at least the competence if not the honesty of the sponsor, leading to a more detailed and systematic (and time-consuming) review than would have otherwise been thought necessary. The easiest way for a sponsor to avoid such problems is to produce the CSR results directly from analysis data sets that were created using the SDTM data sets, even if it requires the sponsor to make available SDTM data sets long before the decision to file has been made. In any event, a company should have a standard data structure in place for normal operations and if your company hasn’t spent the necessary time and money already to have one in place, there are a lot worse structures than SDTM.

Combining data from multiple studies is a common task these days and can be a tedious nightmare unless the data is organized in an internally consistent manner. If you work with multiple EDC vendors, you should be well aware that the final raw databases from one vendor’s study will not be consistent with any other study’s databases (even when using the same EDC product), and that effort will need to be expended to make the data combine properly. Even if you don’t, you’ll still find that small inconsistencies can creep in into the final database of the most diligent Data Management group, as their focus is on a trial at a time, not having to combine data from multiple trials. Let Data Management do their job, take what they provide and make it standard.

SDTM is an industry wide standard, so not only are external training resources available, but it is more likely that you will find a new employee/contractor that is familiar with the data structure and need less training in order to be productive than someone who isn’t experienced in your company’s standard.

The focus of most sponsors is to eventually file a drug submission, so they are going to have to implement the SDTM standard sooner or later. Doing it earlier will permit employees to gain familiarity and experience with SDTM long before they are in the middle of a filing and rushed for time. There is always too much work to do at the end of a filing, it is better to move as much effort as early as possible in the filing process. Also, unlike the situation where the SDTM is prepared as a standalone product, working with it earlier means that the accuracy and completeness will be checked on an ongoing basis, so the validation work necessary will be minimal and easy. If you wait until the end to create SDTM, so that it is never used, you may be in for an unpleasant surprise when a reviewer gets ahold of your data.

Answering ad-hoc queries about the clinical data is an overlooked reason to have standardized data at your beck and call. More and more, the work of a SAS programmer involves finding the answer to questions others have about the data, and it's the sign of a good group if they are prepared and ready for anything that comes wafting their way. You'll also make a lot of people outside your immediate department happy, which is not a bad thing to have in times where budgets are tight and cutting bone is sometimes necessary. The more people who think your work is an asset and necessary to the proper functioning of the company, the more stable your position will be.

This does mean that an attitude will need to shift; people need to stop thinking that SDTM is a submission only item.

WHY ALL THE DATA?

A common question is does SDTM really have to include all of the data? The short answer is yes, a longer answer is yes, because it means a reviewer can more easily assess study conduct if everything is there. This is why the findings domains include the xxSTAT and xxREASND columns, so that the sponsor can report instances where some data point was to be collected but for some reason was not. This does matter. Reviewers are well aware that stuff happens during trials, that not everything scheduled to be done is done, not every expected data point is collected. Including the reports of when something didn't get collected assures the reviewer of two things, that you made an effort as the sponsor to make sure that something was supposed to be done at a given time following the study conduct rules, and that the sponsor isn't just selectively reporting results. Reviewers are well aware that clinical studies are conducted by human beings, that not everything is going to be perfect and so they would look on data that is too impossibly perfect as maybe having been "cleaned" a little too extensively and purged of unpleasant results. What happens when they find this? More questions for the sponsor…

A related problem can appear in data that is exclusively ordered and reported by the scheduled study visit. Take a look at the table below, which was selectively reproduced from a CRT done back in the paper days at a company that no longer exists.

Subject Number	Visit	Visit Date	Drug Name	Dose (mg)
XXX-XYZ	Visit One	10SEP1995	Whatever	30
	Visit Two	17SEP1905	Still Whatever	40
	Visit Three	20SEP1995	Still Whatever	40

A close look at the table shows a most interesting fact, that the subject went back in time and took a dose of a study drug approximately 85 years before the drug was first synthesized in a laboratory. How does this happen? The programmer used the visit number to order the dose record instead of the date, probably because it was easy. This is a problem; as ordering by visit is only appropriate in places like the protocol where you are still thinking in terms of what is supposed to happen. CSRs do include a section on what was supposed to happen, but the meat of the document concerns itself with what did actually happen during the trial. Not every subject comes back when scheduled for data collection; that is why analysis programming typically uses date driven windowing and rules to select observations for analysis, because visit driven programming will tell a different story if all subjects weren't perfectly compliant and didn't necessarily come back only when they were scheduled to return.

So, who does order data by scheduled visit? My best guess is this is due to the influence of Data Management groups, where data is reported and cleaned in terms of the case report form or the EDC screen, where visits figure prominently. Nothing wrong with that, but it should point out the necessity for checking the validity of collected dates against the visit schedule, and doing it at a time when the database isn't locked, so sites are available to answer queries and hopefully fix date issues while still early in the CSR preparation process. This is another reason why SDTM needs to be prepared earlier than the end of a filing.

I will also note that one of the uses of the SV module is in date checking. It is the only place within SDTM where all of a subjects' visits are reported and provides another reason why the SV domain should be populated early, so as to easily identify suspect dates and also prepare the background for narratives about any subject who grossly deviated from the trial's schedule. The reviewer will be looking at the SV provided to him and use it to identify problematic subjects, and guess what happens when there are things that cannot be explained using the available data? More questions for the sponsor…

WHY SE?

The SE domain can be considered one of the red-headed stepchild domains in SDTM, as it is probably the domain that comes to mind most often when people question whether all domains need to be included in a submission when you survey social media about SDTM. The question pops up every few months or so and never seems to totally die away, even though FDA is on record as saying that they explicitly expect the SE data set to be submitted as part of a filing. This leaves me somewhat bewildered, but maybe I was spoiled by my first employer. Way back in 1992-93, something that they called segment logic appeared on my desk and I was expected to start including it in all of my data listings. Not realizing that we could have whined about it, my colleagues and I started including the trial elements, duration intervals and element days into our work. I won't lie and pretend that it was a trivial exercise; we had to work through some considerable implementation issues, but after a while, the necessary code was standardized and inclusion did become a trivial task. Since then, I've carried that work to a number of companies and while often receiving initial resistance from people unfamiliar with the concept, that resistance has always melted away when people could see the end result and understand how useful it could be.

SE, as it currently exists, has all of useful items that segment logic ever had. What SE gives you is the ability to populate EPOCH (another FDA "request") in all of the domains where date is present and EPOCH would be meaningful, like AE. Even more helpful is if you construct the study elements so that EPOCH contains treatment information. While the Implementation Guide is not helpful enough to show an example of this (even where the elements contain treatment information, for some reason EPOCH does not), the specifications do not forbid it and it can provide useful context directly in the domain.

Here's an example of where EPOCH lends context. Below is a table of completely fabricated AE data, including the AETERM, AESTDTC and AESTDY along with the EPOCH, which is this case is taken from a five way crossover design, where the subject is dosed once weekly, with a washout period of six days between each active treatment element. The EPOCH value is one that indicates what the subject was taking, rather than one where all treatment is blinded.

AETERM	AESTDTC	AESTDY	EPOCH
Anaphylactic reaction	2010-08-10	15	Formulation C
Reaction to Medication	2010-09-11	22	Formulation C
Allergic Reaction	2010-10-12	29	Formulation C
Flushing and Trouble Breathing	2011-01-13	8	Formulation C
Anaphylactic shock	2011-03-30	15	Formulation C

Note that while AESTDTC provides a clear start date for the event, it provides no context, while AESTDY does make it clear that whatever is happening seems to be happening on the first day of a treatment period. EPOCH on the other hand, makes it pretty clear that there seems to be a problem with one of the formulations in this study.

Even further context is given by the EPOCH DAY, a relative day calculated from the start of each EPOCH. This variable is mentioned positively in the IG but sadly relegated to SUPPQUAL for the present. The column isn't present because the entire table did not really need repeating, but the EPOCH DAY in all instances is 1, take my word for it. This should provide further confirmation that Formulation C isn't ever likely to be marketed so long as tort lawyers are active and in need of settlement payments.

Does context really matter? It doesn't unless the data user is actually looking at the data, but that is something for which SDTM is intended. Even with multiple monitor setups, it is a pain to have to keep multiple windows of different data sets synchronized so that the context can be taken from them. It's a lot easier to see if every applicable column is present in one window.

Now, someone could argue that this context is very helpful, but doesn't really require a separate data set, that EPOCH and EPOCH DAY could be programmed separately within each domain. Think about that for a minute. Does anyone really want to be responsible for maintaining the code to determine EPOCH and EPOCH DAY in 15-20 separate programs, just so you don't have to provide one domain?

I suspect that most complaints about SE stem from unfamiliarity with the multitude of uses for the data set if it actually exists. SE is very useful for date checking across the other domains. Like SV, it is an easy place to assess study conduct, and you can expect any reviewer to be doing exactly that. It never hurts to do the same review, and be prepared to answer questions about any problematic subjects.

CONCLUSION

I hope that I have offered a sufficient explanation of SDTM’s place in the regulatory world and also provided some reasons why sloppy submissions may have been tolerated in the past but that toleration is not likely to continue in the future. FDA has been traditionally slow to enforce standards and mandate their implementation, but they eventually get there and the result will not be a pleasant one if companies don’t heed the semaphore flags being currently waved from government towers. I would urge you to go back to your companies and instead of asking whether some data has to be included in SDTM, ask yourself for a good reason why it should not, and act accordingly. I would also ask that you start thinking about the place of SDTM in your current workflow, and where it can be first done to maximize the benefit to your company.

REFERENCES

CDISC Study Data Tabulation Model (SDTM) v1.4 and Study Data Tabulation Model Implementation Guide

(SDTMIG) v3.2. http://www.cdisc.org/sdtm.

FDA. CDER Common Data Standards Issues Document.

http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/UCM254113.pdf

FDA CDER,CBER,CDRH. Providing Regulatory Submissions in Electronic Format ² Standardized Study Data.

http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM292334.pdf

Fred Wood, Mary Lenzen. Trials and Tribulations of SDTM Trial Design. PharmaSUG 2011 - Paper CD13

CONTACT INFORMATION

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

Name: Henry B. Winsor
Enterprise: Relypsa, Inc.
Address: 700 Saginaw Drive
City, State ZIP: Redwood City, CA 94063
Work Phone: 650-421-9585
Fax: 650-421-9785
E-mail: hwinsor@relypsa.com
Web: n/a
Twitter: n/a