PhUSE 2014

Paper CD14

Considerations in Creating Transparent SDTM-Based Datasets

Jerry Salyers, Accenture Life Sciences, Berwyn, PA., USA
Richard Lewis, Accenture Life Sciences, Berwyn PA., USA
Fred Wood, Accenture Life Sciences, Berwyn, PA., USA

ABSTRACT

One of the definitions that one finds for "transparency" emphasizes that anything judged to be transparent is "characterized by visibility or accessibility of information" and that it is "free from pretense or deceit". As one looks at the real issues of data transparency, and in much the same way as looking at data "traceability", the methods for the preparation of the SDTM-based submission datasets play an important role. The methods employed will impact a reviewer's level of confidence in how "transparent" the final datasets are and how well they reflect the overall clinical data flow from the point of collection to the point of submission.

The correct mapping of stored operational data to SDTM-based datasets represents an all-important layer in the overall quest to provide transparent and traceable data. This paper will look at a number of specific areas where incorrect or misleading mapping may compromise this goal.

INTRODUCTION

Data transparency is currently a much talked about topic among pharma companies. The pharmaceutical industry as a whole has come to understand that study subjects have a right to know and understand the results of the clinical trial in which they participated. Above that, companies also now realize that the entire research community can benefit from more and better access to not just, heretofore proprietary, clinical study reports (CSRs) but also to clinical trial data at the subject or "participant" level. This data can be used by other researchers to perform further analysis or to simply confirm the original sponsor's own analysis. Making data transparent should be done in such a way that it doesn't violate any subject's right to privacy, but allows others to use the data to further research on the disease under study.

The AllTrials initiative and petition in Europe has resulted in new laws being considered that require pharma companies to register their studies and to publish the results. In the US, several sponsors either have their own dedicated trial registries or have joined a larger group of sponsors in configuring a website (www.clinicalstudydatarequest.com) dedicated to sharing "anonymized" clinical trial data with researchers who have submitted research proposals to an independent review panel for consideration.

This all comes under the heading of "transparency" in clinical research, albeit at a fairly high level. At the data level, new draft guidances recently released by the FDA point to the importance of submitting study data in a standardized electronic format in accordance with CDISC standards and the new Technical Conformance Guide. This guide, once final, is to be paired with the Data Standards Catalog and will supersede both the Study Data Specification as well as the most recent version of the Common Data Standards Issues document. These documents, in total, emphasize the importance of "transparency" in the tabulation data. They can be found via the following link:

http://www.fda.gov/forindustry/datastandards/studydatastandards/default.htm

This paper will focus on several examples of how transparency at the data level may be compromised in the eyes of a reviewer. We will examine a number of issues involving several incorrect assumptions on the part of sponsors that may lead to a reviewer having serious questions or doubts concerning a submission's level of data transparency. Employing less than optimum or misleading mapping of collected data to the submission-ready (SDTM) datasets may affect how that data is accepted and analyzed by the regulatory agencies.

VARIABLE CORE DESIGNATIONS

In Section 4.1.2 of the Technical Conformance Guide, there is a short discussion on the SDTM variable "core designations". In short, these are:

- Required – Column must be in dataset and there can be no NULL values.
- Expected – Column must be in dataset, however, under some circumstances, some rows may have NULL values.

- Permissible – Variable to be included in the dataset *if collected*.

Some sponsors have the idea that "Permissible" variables, even if collected, don't have to be included in the dataset. An example might be something such as the "indication" for a concomitant medication. If the CRF collected why the subject took the medication, it's incumbent upon the sponsor to submit the information. Less clear, but just as valid, is if the concomitant medication CRF collected an "ongoing" box, and it was "checked". In this case, the sponsor must translate that "check box" to a Relative Timing variable, either based on the study reference period or another reference "time point".

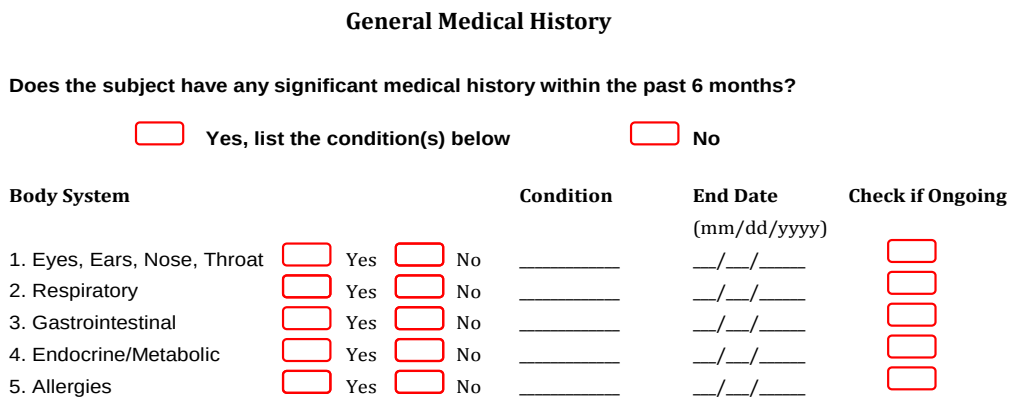
The Technical Conformance Guide also stipulates that whenever a date is collected (--DTC, STDTC, ENDTC), such as a study medication start date, EXSTDTC, that the "matching" Study Day variable, or EXSTDY, should also be provided. Similarly, in the LB domain, if the "expected" variable LBDTC is populated, then the matching LBDY variable should also be provided.

PROMPT QUESTIONS

Sponsors often attempt to submit data in SDTM-based datasets that fall under the category of a "prompt" question that is used mainly by study monitors or by data management in data cleaning. Prompt questions, at the request of the FDA, are not migrated to the tabulation data. A question can be considered a prompt question if:

- It's largely for data cleaning purposes such as "Does the subject have any relevant medical history in the 6 months prior to the Screening visit? A "No" response simply confirms the CRF page was seen by the site staff.
- The "No" response will not be part of any descriptive statistics or analysis
- The "Yes" response is confirmed by the presence of a record

Often, sponsors have a difficult time telling the difference between a prompt question and a valid use of the --OCCUR variable. Refer to the snippet of a Medical History page shown below:



General Medical History

Does the subject have any significant medical history within the past 6 months?

☐ **Yes, list the condition(s) below** ☐ **No**

Body System			Condition	End Date (mm/dd/yyyy)	Check if Ongoing
1. Eyes, Ears, Nose, Throat	☐ Yes	☐ No	_____	__/__/____	☐
2. Respiratory	☐ Yes	☐ No	_____	__/__/____	☐
3. Gastrointestinal	☐ Yes	☐ No	_____	__/__/____	☐
4. Endocrine/Metabolic	☐ Yes	☐ No	_____	__/__/____	☐
5. Allergies	☐ Yes	☐ No	_____	__/__/____	☐

The issue regarding the CRF page above is whether the "Body System" represents a "Pre-Specified" term. When Medical History terms are pre-specified, then the MHOCCUR variable comes into play. Recently, a sponsor delivered this MH dataset, based on the above CRF, for compliance review (Note, some columns removed for brevity):

DOMAIN	MHSEQ	MHTERM	MHCAT	MHSCAT	MHOCCUR	MHENRTPT	MHENTPT
MH	1	Allergies	GENERAL		N	U	SCREEN
MH	2	Endocrine/Metabolic	GENERAL		N	U	SCREEN
MH	3	Eyes, Ears, Nose, Throat	GENERAL		N	U	SCREEN
MH	4	Shortness of Breath	GENERAL	Respiratory	Y	ONGOING	SCREEN
MH	5	Pleurisy	GENERAL	Respiratory		ONGOING	SCREEN
MH	6	Decreased Appetite	GENERAL	Gastrointestinal	Y	ONGOING	SCREEN
MH	7	Constipation	GENERAL	Gastrointestinal		ONGOING	SCREEN

The "Body Systems" shown on the CRF do not by themselves constitute "Pre-Specified Medical History terms. Thus, MHOCCUR should not be used. When a particular body system did not have a history, the MHTERM was set to the Body System and MHSCAT is blank. Where a body system did have a history, the "Condition" as collected on the CRF became the MHTERM and the MHSCAT was set to the Body System. This was an instance where the dataset, although incorrect according to SDTM, would not fail validation checks.

Of course, while the SDTM Implementation Guide (SDTMIG) supports migrating the "body system" itself to MHSCAT where a history does exist, the "Yes/No" checkboxes next to the body systems are indeed "prompts" and should be annotated as "NOT SUBMITTED" in SDTM as should the over-arching question at the top of the page. Thus, in the absence of any true "Pre-Specified" terms, the MHOCCUR variable should be omitted from the dataset. Also of interest in the above dataset was the incorrect assigning of 'U' to the MHENRTPT variable for those body systems without a reported history.

In conclusion, the MH dataset should only contain records for those body systems with a reported history. The corrected dataset would look like this:

DOMAIN	MHSEQ	MHTERM	MHCAT	MHSCAT	MHENRTPT	MHENTPT
MH	1	Shortness of Breath	GENERAL	Respiratory	ONGOING	SCREEN
MH	2	Pleurisy	GENERAL	Respiratory	ONGOING	SCREEN
MH	3	Decreased Appetite	GENERAL	Gastrointestinal	ONGOING	SCREEN
MH	4	Constipation	GENERAL	Gastrointestinal	ONGOING	SCREEN

USING AN SDTM VARIABLE INCORRECTLY ("HIJACKING")

Often, sponsors may choose to use a variable for other than its intended purpose. This may be done for a number of reasons, but most often to either avoid creating a Supplemental Qualifier dataset or record or to simply "find a place" to map operational data "as collected" while also submitting the data as outlined by the SDTMIG. Consider the following snippet of a Substance Use (SU) dataset where the CRF captured the usual choices of 'Never', 'Former' and 'Current'.

DOMAIN	USUBJID	SUSEQ	SUTRT	SUCAT	SUSCAT	SUOCCUR	SUENRF
SU	001-0001	1	ALCOHOL	ALCOHOL HISTORY	NEVER	N	
SU	001-0001	2	TOBACCO	TOBACCO HISTORY	FORMER	Y	BEFORE
SU	001-0002	1	ALCOHOL	ALCOHOL HISTORY	CURRENT	Y	DURING/AFTER
SU	001-0002	2	TOBACCO	TOBACCO HISTORY	NEVER	N	
SU	001-0003	1	ALCOHOL	ALCOHOL HISTORY	NEVER	N	
SU	001-0003	2	TOBACCO	TOBACCO HISTORY	CURRENT	Y	DURING/AFTER

Again, this dataset would not fail any validation checks, however, it's clear that SUSCAT is being used to store the "data as collected" on the CRF, rather than using the value to populate the appropriate Relative Timing variable. When using the "Category" variables, it's important to remember that any assigned "subcategory" needs to be a subcategory of the value stored in the --CAT variable. We should also remember that categories and sub-categories are generally known and defined prior to data collection. In our example here, it's clear that the CRF choices of 'NEVER', 'FORMER', or 'CURRENT' are not true subcategories of the categories of Alcohol and Tobacco History.

The MH dataset below offers another example where a Timing variable is used for something other than how it's intended. Medical history was collected on both a "Parkinson's Disease" page and a "General Medical History" page.

MHSEQ	MHTERM	MHCAT	MHPRESP	MHOCCUR	MHDTC	MHSTDTC	MHENRTPT
1	ATYPICAL	PARK	Y	N	2006-05-01		
2	BALANCE	PARK	Y	Y	2006-05-01	2006-09-21	
3	BRADYKINESIA	PARK	Y	Y	2006-05-01	2006-09-21	
4	DYSKINESIA	PARK	Y	Y	2006-05-01	2007-02-12	
5	ABNORMAL GAIT	PARK	Y	Y	2006-05-01	2007-02-12	
7	L-DOPA TREATMENT	PARK	Y	Y	2006-05-01	2006-12-15	
8	RIGIDITY	PARK	Y	Y	2006-05-01	2007-08-21	
9	TREMOR	PARK	Y	Y	2006-05-01	2006-04-01	
10	BPH	GEN			2010-09-30		ONGOING
11	DEPRESSION	GEN			2010-09-30		ONGOING
12	GLAUCOMA	GEN			2010-09-30		ONGOING

In looking at the MHDTC column, it's clear that the dates mean different things. According to the SDTMIG, the --DTC variable in an Events observation class domain should be the date the information was recorded. Is that what we're

seeing here? According to the annotated CRF, the date reported in the --DTC variable is the date of original diagnosis of Parkinson's disease. For the spontaneously reported MHTERM values from the "General Medical History" page, the date in MHDTC can be seen to revert back to its intended use. Of course, the date shown in MHSTDTC is the date that the condition noted in the MHTERM variable was first diagnosed.

A better solution would have been to either create an additional record for overall diagnosis of Parkinson's disease or to create a Supplemental Qualifier record with an appropriate QNAM. Of course, the downside to the SUPPMH approach would be that the date of original diagnosis would appear on each record, similar to what we already have, although in a more aptly named variable.

A good thing to remember is that just because one can document the use of a variable in the Define file and even provide a CRF annotation as to how it's being used, doesn't mean that one can use a variable for other than its intended purpose.

TRANSPARENCY IN THE EXPOSURE (EX) DOMAIN

Our experience in data conversion confirms that despite the importance of providing a correct and thorough representation of how subjects are exposed to study medication, the EX domain remains one of the most vulnerable to poor or incomplete data collection and validation. Also, we've seen many occasions where sponsors may not even use all of the data points that are collected in order to complete the story of how a subject took the study medication. Consider the following scenario where subjects were taking 150 mg of study medication once a day for two weeks. The sponsor (or their CRO) supplied an EX dataset that looked like this (for two subjects):

DOMAIN	USUBJID	EXSEQ	EXTRT	EXCAT	EXDOSE	EXDOSFRQ	EXSTDTC	EXENDTC
EX	001-0001	1	DRUG A	AT SITE	150	QD	2012-01-08	2012-01-08
EX	001-0001	2	DRUG A	AT SITE	150	QD	2012-01-15	2012-01-15
EX	001-0001	3	DRUG A	AT SITE	150	QD	2012-01-22	2012-01-22
EX	001-0002	1	DRUG A	AT SITE	150	QD	2012-01-08	2012-01-08
EX	001-0002	2	DRUG A	AT SITE	150	QD	2012-01-15	2012-01-15
EX	001-0002	3	DRUG A	AT SITE	150	QD	2012-01-22	2012-01-22

This dataset provides us with the discreet doses the subject took at the study site as observed by the investigator. But does this tell a reviewer the whole story as to how the subject took drug? All we see are the doses that were observed. It appears as if the subjects took only three doses over a two-week period. What about the doses the subject took away from the clinic? An option would be to create a "blanket" record that covers the two-week dosing period, while also retaining these "AT SITE" dosing records. It bears mentioning that the frequency of 'QD' in the table above does not adequately reflect that the subject was taking drug outside of the clinic, apart from the doses observed by the investigator. The "blanket" records for the two subjects may look like this:

DOMAIN	USUBJID	EXSEQ	EXTRT	EXCAT	EXDOSE	EXDOSFRQ	EXSTDTC	EXENDTC
EX	001-0001	4	DRUG A	DOSING PERIOD	150	QD	2012-01-08	2012-01-22
EX	001-0002	4	DRUG A	DOSING PERIOD	150	QD	2012-01-08	2012-01-22

TRANSPARENCY IN THE TRIAL DESIGN DOMAINS

TA EXAMPLE

As the machine readable representation of a clinical trial and its design, the Trial Design datasets are of ever increasing importance to reviewers. Poor quality Trial Design datasets can lead to incorrect assumptions regarding the "planned" conduct of a clinical trial and how subjects progressed through the defined Elements of the study. Certainly, as the Trial Design datasets can be developed straight from the protocol, they can be used during CRF design to ensure that data points referenced in the Element Start Rules are indeed collected. The Trial Arms (TA) datasets allows a reviewer to easily see how subjects transition through the different periods or phases of a clinical trial. Element Start Rules in the Trial Elements (TE) dataset are of extra importance as they are used to develop and populate the dates in the Subject Elements (SE) dataset.

An issue that we often see is the inadequate differentiation of a study's treatment Elements. For a study with more than a single Arm, there must be an Element, most often a treatment Element, that lends uniqueness to each Arm. Consider the following two-Arm trial as illustrated in the TA dataset:

DOMAIN	ARMCD	ARM	TAETORD	ETCD	ELEMENT	TABRANCH	EPOCH
TA	A	DRUG A	1	SCRN	Screen	Randomized to Drug A	SCREENING
TA	A	DRUG A	2	TRT	Treatment		TREATMENT
TA	A	DRUG A	3	FU	Follow Up		FOLLOW-UP
TA	B	SOC + DRUG A	1	SCRN	Screen	Randomized to SOC + Drug A	SCREENING
TA	B	SOC + DRUG A	2	TRT	Treatment		TREATMENT
TA	B	SOC + DRUG A	3	FU	Follow up		FOLLOW-UP

It can be seen that subjects in both Arms proceed through the same set of Elements. Currently, there is not an Element that differentiates one Arm from the other. Again, in order to be a different ARM, there must be a unique Element at some point that subjects "branch" to. Therefore, we proposed the following to the sponsor:

DOMAIN	ARMCD	ARM	TAETORD	ETCD	ELEMENT	TABRANCH	EPOCH
TA	A	DRUG A	1	SCRN	Screen	Randomized to Drug A	SCREENING
TA	A	DRUG A	2	DRUGA	Drug A		TREATMENT
TA	A	DRUG A	3	FU	Follow Up		FOLLOW-UP
TA	B	SOC + DRUG A	1	SCRN	Screen	Randomized to SOC + Drug A	SCREENING
TA	B	SOC + DRUG A	2	SOCDRUGA	SOC plus Drug A		TREATMENT
TA	B	SOC + DRUG A	3	FU	Follow up		FOLLOW-UP

This representation provides a unique sequence of elements for each of the two Arms to which subjects may be assigned or randomized.

TA/TE/SE EXAMPLE

The following datasets represent an example where TA, TE and SE (Subject Elements) were submitted for "compliance" review. It bears mentioning, once again, that these datasets did not result in validation errors, even where there are obvious "gaps" between elements as shown in the SE dataset.

The study in question is a two-period crossover study where the same amount of study medication is administered via injection in one anatomical location during Period 1 and then another anatomical location in Period 2. Subjects are randomized to their assigned sequence, either "AS" or "SA".

DOMAIN	ARMCD	ARM	TAETORD	ETCD	ELEMENT	TABRANCH	EPOCH
TA	AS	TRT A/TRT S	1	SCREEN	Screening	Randomized to Treatment Sequence AS	SCREENING
TA	AS	TRT A/TRT S	2	PERIOD1	Period 1		PERIOD I
TA	AS	TRT A/TRT S	3	WASHOUT	Washout		WASHOUT
TA	AS	TRT A/TRT S	4	PERIOD2	Period 2		PERIOD II
TA	AS	TRT A/TRT S	5	EOS	End of Study		END OF STUDY
TA	SA	TRT S/TRT A	1	SCREEN	Screening	Randomized to Treatment Sequence SA	SCREENING
TA	SA	TRT S/TRT A	2	PERIOD1	Period 1		PERIOD I
TA	SA	TRT S/TRT A	3	WASHOUT	Washout		WASHOUT
TA	SA	TRT S/TRT A	4	PERIOD2	Period 2		PERIOD II
TA	SA	TRT S/TRT A	5	EOS	End of Study		END OF STUDY

As is apparent, similar to our earlier example, both Arms are shown to progress through the same sequence of Elements. Rather than having treatment Elements that are specific to the different anatomical locations, the treatment Elements just reflect the study period in which the treatment is administered. The TE dataset for this study looked like this:

DOMAIN	ETCD	ELEMENT	TESTRL	TEDUR
TE	SCREEN	Screening	Within 28 days prior to first treatment	P28D
TE	PERIOD1	Period 1	Period I Check-in (Day -1)	P3D
TE	WASHOUT	Washout	Washout for 5 days	P5D
TE	PERIOD2	Period 2	Period II Check-in (Day 6)	P3D
TE	EOS	End of Study	Immediately following 24-hour blood sample	P1D

When looking at this dataset, it looks like the Elements could only be in support of a single Arm, but we know that this a two-Arm study. There needs to be an Element that is specific for each administered dose. As we see, at the moment, the "treatment" Elements reflect nothing more than the study period. We also see an issue when examining the Start Rules for the Screening Element, the Washout Element, and the End of Study Element. These "rules" do not correspond to actual data points in the SDTM data. For instance, where is the data point that defines the start of the Washout Element? What is indicated is no more than a "duration" for the Washout Element. For the EOS Element, which 24-hour blood sample is the rule referencing? These issues become even more apparent when examining the SE dataset for the study's first subject:

DOMAIN	USUBJID	SESEQ	ETCD	ELEMENT	SESTDTC	SEENDTC	TAETORD	EPOCH
SE	ABC-101-001	1	SCREEN	Screening	2011-10-12	2011-11-07	1	SCREENING
SE	ABC-101-001	2	PERIOD1	Period 1	2011-11-08	2011-11-10	2	PERIOD I
SE	ABC-101-001	3	PERIOD2	Period 2	2011-11-15	2011-11-17	4	PERIOD II
SE	ABC-101-001	4	EOS	End of Study	2011-11-21	2011-11-22	5	END OF STUDY

We see that in the SE dataset, the Washout Element is missing altogether and that there is a wide "gap" between the end of Period 1 and the start of the Period 2 Element. Similarly, there is a gap between the end of the Period 2 Element and the start of the End of Study Element. Beginning with TE, we proposed the following back to the sponsor, replacing the "Period 1" and "Period 2" Elements with location specific treatment Elements: We also inserted an "end rule" column and removed the TEDUR column.

DOMAIN	ETCD	ELEMENT	TESTRL	TEENRL
TE	SCREEN	Screening	Date of Informed Consent	First dose, up to 28 days after the start of the element
TE	TRTLOCA	Treatment Location A	Date treatment administered to anatomical location A	Date of 24-hour blood sample
TE	WASHOUT	Washout	Date of 24-hour blood sample at Period 1	Date of study drug administration at Period 2
TE	TRTLOCS	Treatment Location S	Date treatment administered to anatomical location S	Date of 24-hour blood sample
TE	EOS	End of Study	Date of 24 hour blood sample at Period 2	Last contact with subject.

Making these changes to TE and correctly implementing the start rules for each element would have resulted in the following SE dataset for the study's first subject (the subject's Arm code being "AS"):

DOMAIN	USUBJID	SESEQ	ETCD	ELEMENT	SESTDTC	SEENDTC	TAETORD	EPOCH
SE	ABC-101-001	1	SCREEN	Screening	2011-10-12	2011-11-09	1	SCREENING
SE	ABC-101-001	2	TRTLOCA	Treatment Location A	2011-11-09	2011-11-10	2	PERIOD I
SE	ABC-101-001	3	WASHOUT	Washout	2011-11-10	2011-11-16	3	WASHOUT
SE	ABC-101-001	3	TRTLOCS	Treatment Location S	2011-11-16	2011-11-17	4	PERIOD II
SE	ABC-101-001	4	EOS	End of Study	2011-11-17	2011-11-22	5	END OF STUDY

OTHER IMPORTANT ASPECTS OF DATA TRANSPARENCY

During the legacy data conversion process, sponsors will sometimes request that "missing" or invalid data be "fixed" in the final dataset. An example may be the "hardcoding" of a weight unit when the unit is missing from the data. Another example may be when sponsors want to infer that "missing data" (data with no result) means that the observation was "NOT DONE, even though that wasn't expressly collected on the CRF. It's important to remember that SDTM-based datasets are the prescribed way to submit tabulation data in standard format, but the format cannot make up for poor data management practices or poor query resolution prior to database lock. This also applies to "missing data". Declaring an observation as "NOT DONE", in the absence of source data to support that designation, would be imputing data and should not be done in the submission datasets.

Data transparency needs to begin as early as possible in the process of developing study level documentation. We encourage sponsors to develop the study metadata and the Study Data Reviewer's Guide as early as possible so that it becomes an integral part of the study documentation from the very beginning. These are the best tools for allowing a reviewer to get up to speed as quickly as possible. Sponsors should look at these tools as more than just fulfilling a regulatory requirement.

For the Define file specifically, sponsors should begin compiling codelists and value-level metadata early in the data conversion or source variable mapping process. Care should be taken to avoid using source variable names where they may differ from the variable designation in the tabulation data. The reviewer may not have access to source variable names. This same reminder would also apply to the annotations as noted in the BlankCRF.

In terms of the Study Data Reviewer's Guide, sponsors should use the PhUSE developed template to explain all aspects of the data as well to provide a summary of the SDTM version used, the version of MedDRA used in performing coding, and the version of controlled terminology used. Again, this Study Data Reviewer's Guide can and should be developed early along with other study documentation. Later, of course, it can also be used to explain any validation issues.

CONCLUSION

The level of confidence on the part of a reviewer regarding a submission's degree of transparency or "traceability" may be compromised by poor or misleading mapping of operational source data to SDTM-based datasets.

On the other hand, the development and inclusion of well-designed metadata can only serve to enhance the overall appearance and confidence in a study's level of data transparency.

As always, it bears repeating that the SDTM and SDTMIG, as the standards for the submission of tabulation data, cannot make up for inadequate data management practices or poor query resolution during study execution. Data transparency and traceability should be the goal through all aspects of the clinical data flow, from CRF development, through data capture, to data management and data submission.

CONTACT INFORMATION

Comments and questions are valued and encouraged. Contact the authors at:

Jerry Salyers
Senior Consultant, Data Standards Consulting
Accenture Life Sciences
1160 W. Swedesford Road Bldg One
Berwyn, PA 19312, USA
513.668.8126
jerry.j.salyers@accenture.com

Richard Lewis
Senior Consultant, Data Standards Consulting
Accenture Life Sciences
1160 W. Swedesford Road Bldg One
Berwyn, PA 19312, USA
262.716.4167
richard.m.lewis@accenture.com

Fred Wood
Senior Manager and Lead, Data Standards Consulting
Accenture Life Sciences

1160 W. Swedesford Road Bldg One
Berwyn, PA 19132, USA
484-881-2297
f.wood@accenture.com