PharmaSUG 2020 - Paper DS-261

SUPPQUAL Datasets: Good, Bad and Ugly

Sergiy Sirichenko, Pinnacle 21

ABSTRACT

SUPPQUAL datasets were designed to represent non-standard variables in SDTM tabulation data. There are many recent discussions about whether the SDTM Model should allow the addition of non-standard variables directly into General Observation Class domains instead of using SUPPQUAL datasets? However, there is still a lack of implementation metrics across the industry to understand actual utilization of SUPPQUAL datasets. In this presentation we will summarize metrics from many studies and different sponsors to produce an overall picture of utilization of SUPPQUAL datasets by the industry. We will analyze commonly used SUPPQUAL variables for being potentially promoted to standard SDTM variables. Also, we will provide and discuss examples of correct and incorrect utilization of SUPPQUAL datasets in submission data to understand if the industry is ready to switch from SUPPQUAL datasets to non-standard variables?

INTRODUCTION

Supplemental Qualifiers (SUPPQUAL) are standardized representation of sponsor's non-SDTM variables. These non-standard variables may have all the properties of standard SDTM variables. For example, if necessary SUPPQUAL variables may contribute to domain keys.

As name "Supplemental Qualifiers" suggests, these datasets are intended to capture additional Qualifiers for an observation. Other type of collected information should not be stored in SUPPQUAL datasets. For example, data which represent separate observations should be handled as separate records in General Class Observation domains. There are other limitations in use of SUPPQAUL datasets. For example, according to SDTM-IG, SUPPQUALs cannot represent subject-level data, which should be provided in Subject Characteristics (SC) domain instead. Another example is findings which represent interpretations or require additional qualifiers, like units or normal ranges. Such information should be stored as separated records in Findings domains. Timing information or information about the non-occurrence events cannot be stored in SUPPQUAL variable as well. Finally, comments should be placed in the dedicated Comments (CO) domain.

Design of SUPPQUAL datasets allows merging non-standard variables to their parent domains. During transpose procedure, QNAM and QLABEL values serve as names and labels of new variables. Similar to --TESTCD and --TEST variables, they are limited to 8 and 40 characters respectively. There is a SAS® macro in PhUSE standard scripts repository that can help with this task [1, 2].

While merging SUPPQUAL variables back to their parent domains looks easy, sometimes there are technical problems. For example, in case of structural inconsistency when there is more than one QNAM record per referenced USUBJID and --SEQ. Some sponsors keep non-standard variables within SDTM domains ("SDTM+ structure") for ease of internal management and convert them into SUPPQUAL datasets at the time of regulatory submission.

Recently, CDISC team has proposed and is still considering a new approach for implementation of non-standard variables. Instead of creating SUPPQUAL datasets, the non-standard variables may be kept in their parent domains. It will simplify implementation and data review process. However, such approach may also encourage excessive use of non-standard variables with potential deviation from SDTM compliance.

SDTM-IG includes an "Appendix C2: Supplemental Qualifiers Name Codes" with 4 standard SUPPQUAL variables (Table 1).

QNAM	QLABEL	Applicable Domains
AESOSP	Other Medically Important SAE	AE
AETRTEM	Treatment Emergent Flag	AE
--CLSIG	Clinically Significant	Findings
--REAS	Reason	All general observation classes

Table 1. CDISC SDTM-IG 3.2 Appendix C2: Supplemental Qualifiers Name Codes

Some SUPPQUAL variables like AETRTEM and AESOSP are required by regulatory agencies. In pre-clinical data, --RESMOD (Results Modifier) in SUPPQUAL provides details about abnormal findings in MI and MS domains. This non-standard variable is utilized in almost every non-clinical study.

A major source of non-standard CDISC variables is Therapeutic Area User Guides (TAUG). These guides acknowledge lack of standard SDTM variables needed for specific therapeutic area. Often, CDISC TAUG is published as a provisional standard. It means that TAUG introduces new variables or special purpose domains which are not yet covered by CDISC SDTM model. When these new variables and special purpose domains become part of new SDTM model, then the provisional TAUG standard will become an official standard.

In some cases, it may take too much time. For example, CDISC Pharmacogenomics/Genetics (PGx) TAUG was published in 2015 as "Provisional" because the document includes many new variables and a relationship domain not supported by CDISC SDTM model. The intention was to include these new variables and domain in the next release of SDTM model, after which CDISC PGx TAUG would be promoted to official CDISC standard. Until then, a "Provisional" version is still a draft version with many expected changes. The CDISC team was planning to add new PGx variables into SDTM model but then pulled them back because it was decided that they were not well-defined. Therefore, today if you want to be CDISC compliant and use new PGx variables you should add them as SUPPQUAL variables.

CDISC has a project which tracks new non-standard variables introduced in TAUGs. There is a special ‘SDTM NSV Registry’ page on CDISC Wiki [3], which we will discuss later in this paper.

There is a non-official best practice on creation of new SUPPQUAL variables:

- SUPPQUAL variable names should start with <domain name> prefix like names of standard variables in domains. For example, AETRTEM, AESOTH, EGCLSIG, etc. (see Table 1 above). Exceptions are sponsor-specific variables which are utilized across domains like VISIT or USUBJID.
- QNAM values cannot use variable names which already exist in SDTM model.
- Utilization of SUPPQUAL variables should be consistent within a study and within a submission. It means that the same type of information should be represented by the same QNAM and QLABEL.
- Also, users should try to use existing non-standard variables from CDISC documentation.

Implementation of non-standard variables is driven by company and study-specific needs. So far, there are no industry-wide metrics which help to understand implementation of SUPPQUAL variables. Such metrics can be helpful for developing CDISC SDTM standard. We decided to run a pilot project to test methodology and potential use of findings for improving standard management processes including data validation.

METHODOLOGY

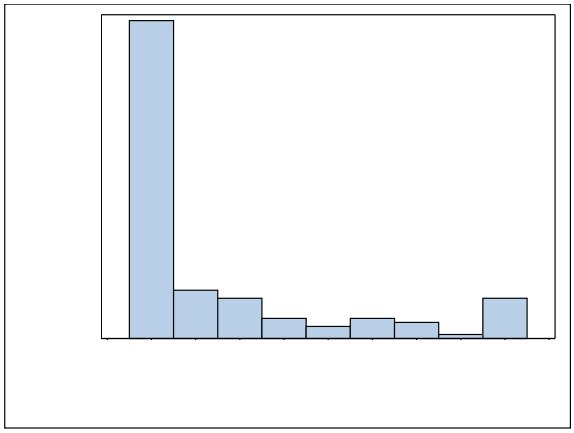
For our research we utilized metrics collected by Pinnacle 21 Enterprise. We planned to analyze clinical studies which are finalized or almost ready for regulatory submissions based on presence of define.xml file, TS domain and absence of data quality issues common for ongoing studies. To ensure diversity of collected data, sample studies were selected from different sponsors, phases and therapeutic areas. One sponsor may be represented by up to 3 studies within each phase and each therapeutic area. For example, it could be up to 3 phase II studies with different indications like Oncology, Antiviral and

Dermatology. Collected data include a list of SUPPQUAL variables, de-identified sponsor and study IDs, study phase and start date, version of SDTM, study indication collapsed into Oncology/Non-oncology categories and Pinnacle 21 Enterprise Validation Score.

We planned to analyze the most common SUPPQUAL variables, diversity in implementation of the same collected information, the industry utilization of CDISC standards and compliance with regulatory requirements.

RESULTS

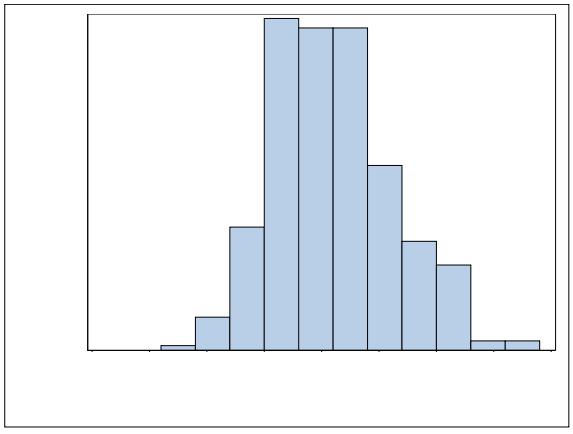
Analyzed SUPPQUAL data represent implementations of 325 studies from 124 sponsors. Most sponsors (60%) are represented by a single study with maximum 9 studies and mean of 2.5 studies (Graph 1).



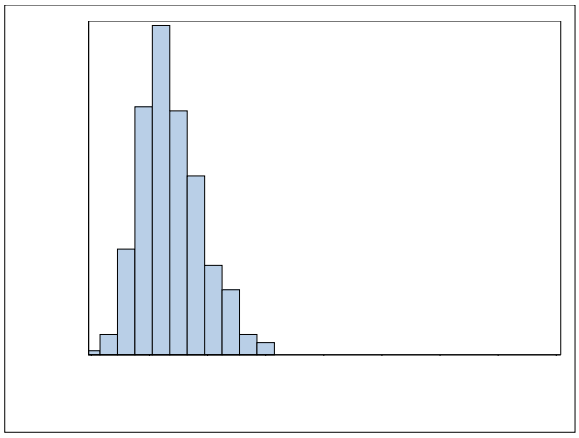
Graph 1. Distribution of number of sample studies (N) per sponsor

There are 91 (28%) oncology and 234 non-oncology studies. Most studies (82%) were started in 2015 or later. 248 (76%) studies were implemented based on SDTM-IG 3.2, 63 (19%) studies are based on SDTM-IG 3.1.3. There are a few studies utilizing other versions of SDTM-IG.

Analyzed studies have from 14 to 75 datasets with mean of 41.5 and median of 41 (Graph 2). There is one study with no SUPPQUAL datasets. Maximum number of SUPPQUAL datasets in studies is 30 while mean is 13.7 and median is 13. (Graph 3).



Graph 2. Distribution of number of all datasets per study



Graph 3. Distribution of number of SUPPQUAL datasets per study

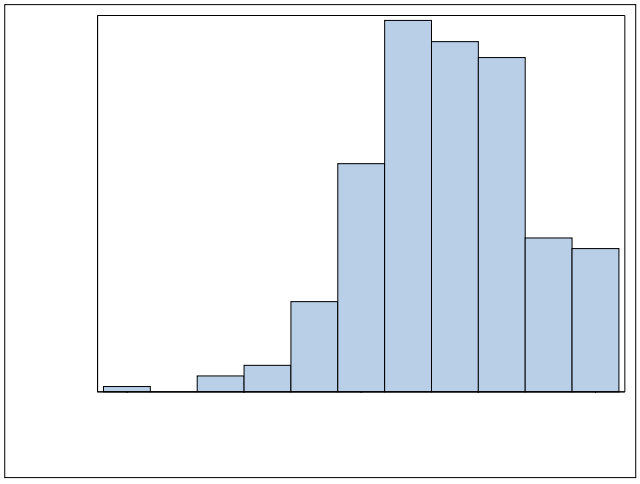
There is a difference between oncology and non-oncology studies. Oncology studies have more total number of datasets (mean 50.0 vs. 38.2) and more SUPPQUAL datasets (mean 17.5 vs 12.2).

We calculated a ratio of SUPPQUAL datasets per other domains in the study as

Ratio = number of SUPPQUALS datasets / number of qualified domains

<Qualified domains> are all General Class domains and DM domain plus any other domains which are not qualified to use SUPPQUAL, but still supplied by SUPPQUAL datasets incorrectly implemented by sponsors.

On average 68.4% of qualified domains have SUPPQUAL datasets across analyzed studies (Graph 4). This SUPPQUAL ratio is higher for oncology (72.3%) compare to non-oncology (66.8%) studies.



Graph 4. SUPPQUAL datasets Ratio across studies

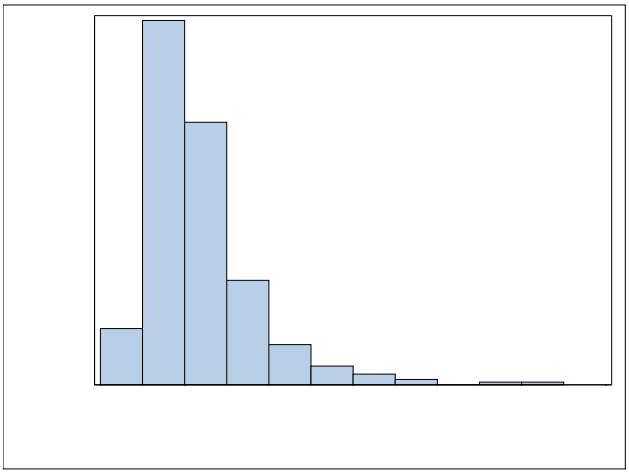
Collected data from 325 studies include 27,023 records of unique (Sponsor/Study/Dataset/QNAM) records which represent implementation of non-standard SDTM variables in SUPPQUAL datasets.

Number of non-standard variables in studies varies from 1 to 618 with mean of 84 and median of 65.5. For each study, we calculated average number of QNAM values per SUPPQUAL dataset as

N_of_NSV = number of unique DOMAIN.QNAM / number of SUPPQUAL datasets

<Number of unique DOMAIN.QNAM> represents all non-standard variables implemented in a study.

Average number of QNAM values per SUPPQUAL dataset varies from 1 to 26.4 with mean of 5.8, standard deviation of 3.4 and median of 5.0 (Graph 5). There is also a difference in this property for non-oncology (mean of 5.6) and oncology (mean of 6.6) studies.



Graph 5. Average number of QNAM values per SUPPQUAL dataset across studies

Table 2 shows the 30 most common SUPPQUAL datasets. SUPPAE is a leader with its utilization of 96% of all analyzed studies and 100% of all oncology studies. Such high use of SUPPAE dataset indicated a lack of standard SDTM variables for handling Adverse Events data. Almost every study requires additional non-standard variables. Another example is SUPPTU dataset with utilization in 76% of oncology studies compare to 52% and 49% of related SUPPTR and SUPPRS datasets. Possible interpretation of this difference could be due to lack of standard variables needed to represent Tumor Identification information.

Dataset	Number of studies with SUPPxx dataset	% (all studies)	% (non-oncology studies)	% (oncology studies)
SUPPAE	313	96.3	94.9	100.0
SUPPCM	298	91.7	89.4	97.8
SUPPDM	289	88.9	87.7	92.1
SUPPLB	251	77.2	74.2	85.4
SUPPDS	246	75.7	71.6	86.5
SUPPEG	222	68.3	64.4	78.7
SUPPMH	206	63.4	62.3	66.3
SUPPEX	205	63.1	56.8	79.8
SUPPDV	190	58.5	58.9	57.3
SUPPPR	141	43.4	33.1	70.8
SUPPPC	139	42.8	39.8	50.6
SUPPPE	120	36.9	39.8	29.2

Dataset	Number of studies with SUPPxx dataset	% (all studies)	% (non-oncology studies)	% (oncology studies)
SUPPVS	115	35.4	35.2	36.0
SUPPEC	112	34.5	28.4	50.6
SUPPFA	106	32.6	29.2	41.6
SUPPQS	105	32.3	33.1	30.3
SUPPDA	86	26.5	27.5	23.6
SUPPSU	69	21.2	21.2	21.3
SUPPTU	68	20.9	0	76.4
SUPPIE	67	20.6	18.6	25.8
SUPPSV	61	18.8	20.8	13.5
SUPPCE	58	17.8	15.7	23.6
SUPPRS	48	14.8	1.7	49.4
SUPPHO	47	14.5	12.3	20.2
SUPPTR	46	14.2	0	51.7
SUPPSS	42	12.9	5.1	33.7
SUPPDD	35	10.8	3.4	30.3
SUPPPP	35	10.8	10.2	12.4
SUPPRP	33	10.2	11.9	5.6
SUPPMI	30	9.2	2.1	28.1

Table 2. The 30 most common SUPPQUAL datasets

The next table represents the 30 most common values for QNAM variables across all SUPPQUAL datasets and studies.

QNAM	QLABEL	Number of studies	% (all studies)
AETRTEM	Treatment Emergent Flag	204	63.0
EGCLSIG	Clinically Significant	148	45.7
RACEOTH	Race, Other	129	39.8
DVTERM1	Protocol Deviation Term 1	98	30.2
LBCLSIG	Clinically Significant	64	19.8
PECLSIG	Clinical Significance	62	19.1
PRLLT	Lowest Level Term	61	18.8
PRHLGT	High Level Group Term	60	18.5
PRHLT	High Level Term	60	18.5
DVTERM2	Protocol Deviation Term 2	57	17.6
PRHLGTCD	High Level Group Term Code	55	17.0
PRHLTCD	High Level Term Code	55	17.0

QNAM	QLABEL	Number of studies	% (all studies)
PRPTCD	Preferred Term Code	53	16.4
PRLLTCD	Lowest Level Term Code	52	16.0
ATC3	ATC Level 3 Text	49	15.1
CMDECOD1	Standardized Medication Name 1	45	13.9
ATC2	ATC Level 2 Text	44	13.6
CMATC2	ATC2	43	13.3
CMATC3	ATC3	42	13.0
PROTVER	Protocol Version	42	13.0
ATC1	ATC Level 1 Text	41	12.7
PRSOC	System Organ Class	41	12.7
RACE1	Race 1	41	12.7
CMATC1	ATC1	40	12.3
CMDECOD2	Standardized Medication Name 2	39	12.0
CMCLAS1	Medication Class 1	38	11.7
RACE2	Race 2	38	11.7
CMATC4	ATC4	37	11.4
COHORT	Cohort	37	11.4
CMCLAS2	Medication Class 2	36	11.1

Table 3. The 30 most common QNAM values

Let's explore and discuss these findings.

AE TREATMENT EMERGENT FLAG

AETRTEM (Treatment Emergent Flag) information in SUPPAE dataset is requested by both FDA and PMDA agencies. There are special data validation rules to ensure that this information is populated for all records in Adverse Events domain. Therefore, it is a surprise that only 63% of studies are compliant with this regulatory requirement.

Some users believe that Treatment Emergent Flag should be populated only in Analysis rather than Tabulation data, since it's a derived value. However, Tabulation data may include other derived information like Study Days and Baseline Flag.

Sometimes, due to missing information like unknown or partial start datetime of adverse event, imputation algorithms may be required to compute Treatment Emergent Flag. SDTM model does not allow imputations. Therefore, users populate AETRTEM info only in Analysis data. However, such approach may reduce the quality of Tabulation data for regulatory review.

The major benefit of standardized data is to enable automation. SDTM structure and CDISC Controlled Terminology are very predictable and allow use of automated review and analysis tools. ADaM structure is flexible to develop datasets which are ready for analysis ("one proc away") in SAS or other tools. However, this analysis process is manual. Most automated review tools use only SDTM data and rely on presence of AE Treatment Emergent information in SUPPAE dataset. If this information is not available, then analysis results may be less predictable and confusing. For example, all records in AE domain may be considered as Treatment Emergent Events. Reviewers may try to derive AETRTEM information by themselves using different criteria compare to one utilized by the sponsor. Sometimes, Analysis Adverse Events dataset is not ideal source of AETRTEM due to unpredictable structure. Therefore, standard

SDTM SUPPQUAL variable AETRTEM is a good communication tool for sponsors to ensure that data review process is aligned with study protocol.

CLINICALLY SIGNIFICANT

The second the most common SUPPQUAL variable is EGCLSIG (Clinically Significant). This non-standard variable is present in 46% of all studies or 66% of studies with SUPPEG dataset. Some studies have SUPPEG dataset with the only non-standard variable being EGCLSIG. This information is also widely utilized in other Findings domains with PE as a runner-up to EG domain.

Dataset	QNAM	QLABEL	Number of studies with QNAM	% (all studies)	Number of studies with Dataset	% (studies with Dataset)
SUPPEG	EGCLSIG	Clinically Significant	148	45.7	224	66.1
SUPPPE	PECLSIG	Clinically Significant	62	19.1	121	51.2
SUPPLB	LBCLSIG	Clinically Significant	64	19.8	253	25.3
SUPPVS	VSCLSIG	Clinically Significant	24	7.4	116	20.7

Table 4. --CLSIG variables in common Finding SUPPQUAL datasets

However, this information in tables 3 and 4 is not very accurate. A problem is lack of CDISC conformance during the industry implementation of non-standard variables. While expected variable name or QNAM value to store “Clinically Significant” info in EG domain is EGCLSIG, there are many variations in implementation (table 5) which increase actual use of Clinically Significant flag in SUPPEG dataset to 50% of all studies or 73% studies with SUPPEG dataset.

QNAM	QLABEL
CLINSIG	CLINICALLY SIGNIFICANT
EGABN	Abnormal Clinically Significant
EGCHG	Clinically Significant Chg. from Screen?
EGCHGCS	ECG Changes Clinically Significant
EGCLIG	Clinical Significance
EGCLISG	Clinically Significant
EGCLISIG	Clinically Significant Result
EGCLSIG	Clinically Significant
EGCLSIG1	Abnormality 1 Clinically Significant
EGCLSIG1	Clinically Significant 1
EGCLSIG2	Clinically Significant 2
EGCS	ECG Clinically Significant

Table 5. ‘Clinically Significant’ variables in SUPPEG datasets

EGCLSIG is a standard CDISC SUPPQUAL variable from SDTM-IG Appendix C [3]. It looks like the industry successfully utilizes the correct variable name (QNAM = EGCLSIG) in 91% of cases. However, there are many variations in labels (QLABEL) for this variable. There are 31 different labels assigned to EGCLSIG variable (table 6). Some of these QLABEL values raise questions about valid utilization of EGCLSIG variable. For example, ‘*EG Clinically Significant, **Specify***’ and ‘*If Abnormal and clin. signf., **specify***’ looks like a Description for clinical significance rather than an expected flag.

QNAM	QLABEL	N
EGCLSIG	Abnormality Clinically Significant	1
EGCLSIG	Abnrml Interpretation Clin Significant?	1
EGCLSIG	CLINICAL SIGNIFICANCE	1
EGCLSIG	CLINICALLY SIGNIFICANT	1
EGCLSIG	CLINICALLY SIGNIFICANT OR NOT	2
EGCLSIG	CS/NCS	1
EGCLSIG	Clinical Significance	19
EGCLSIG	Clinical Significance Flag	1
EGCLSIG	Clinically Siginificant	1
EGCLSIG	Clinically Significant	94
EGCLSIG	Clinically Significant Abnormality	1
EGCLSIG	Clinically Significant for EG	1
EGCLSIG	Clinically Significant?	2
EGCLSIG	Clinically significance	1
EGCLSIG	Clinically significant	4
EGCLSIG	ECG Res. Abnormal Clinically Significant	1
EGCLSIG	ECG Res. clinically significant	1
EGCLSIG	ECG Result Abnormal Clin. Significant	1
EGCLSIG	ECG Test Result Clinically Significant	1
EGCLSIG	EG Clinically Significant, Specify	1
EGCLSIG	EG Clinically significant?	1
EGCLSIG	EG: If Abnormal, is it Clin Significant?	1
EGCLSIG	EGCLSIG	1
EGCLSIG	If Abnormal and clin. signf., specify	1
EGCLSIG	If Abnormal, Clinical Significance	1
EGCLSIG	If abnormal, clinically significant?	1
EGCLSIG	Interpretation Clinically Significant	2
EGCLSIG	Is the Result Clinically Significant?	1
EGCLSIG	SIGNIFICANCE OF ABNORMALITY	1
EGCLSIG	Was Abnormality Clinically Significant?	2
EGCLSIG	Was Finding Clinically Significant?	1

Table 6. Labels for EGCLSIG variable across studies

In some studies which do not have SUPPQUAL variable dedicated to 'Clinically Significant' flag, this information is still collected, but presented in less standardized way. A common example is 'Interpretation' variable in SUPPEG dataset populated with 3 terms "*Normal*', '*Abnormal, not clinically significant*' and '*Abnormal, clinically significant*'. In these cases, users mixed two potentially different types of information (Normal/Abnormal Result Interpretation and Clinically Significance) in a single non-standard variable. The industry needs a guidance and education on the correct implementation.

CDISC SDTM team is planning to add --CLSIG (Clinically Significant) variable to SDTM model. It should help with consistent implementation across the industry.

WHO DRUG CODING

Early versions of CDISC standards assumed usage of any coding dictionary. For example, Adverse Events may be coded utilizing MedDRA, SNOMED or COSTART. Therefore, AE domain in SDTM-IG 3.1.2 had only two generic coding variables AEDECOD and AEBODSYS. Later, CDISC acknowledged an exclusive use of MedDRA dictionary for FDA submissions and expanded SDTM model with variables specific to MedDRA coding to support regulatory review needs.

Unfortunately, it is still not the case for WHO Drug dictionary which is the primary coding dictionary for Concomitant Medications and has been recently added to FDA Data Standards Catalog [4] similar to MedDRA. So far, SDTM model supports only WHO Drug generic drug name in CMDECOD variable. SDTM-IG suggests utilization of SUPPCM dataset for additional coding information. However, no details or examples of implementation are provided. Lack of data standardization and guidance for WHO Drug coding results in huge diversity of implementations by the industry.

For example, 298 studies with SUPPCM datasets have 1,023 different values for QNAM / QLABEL or 667 unique QNAM values for records which represent WHO Drug Anatomical Therapeutic Chemical (ATC) Classification coding. There are 128 variations of QNAM values which include text ‘ATC1’ (table 7).

QNAM	QLABEL
ATC1	ATC Chemical Subgroup 1st Level
ATC1_C	ATC 1 Class Code
ATC1_T	ATC 1 Class Text
ATC1C	ATC 1 CODE
ATC1C_1	ATC1 Code 1
ATC1CD	ATC Chemical Subgroup 1st Level Code
ATC1CODE	ATC 1 CODE
ATC1ID	ATC Code Class 1
ATC1M14	ATC Level 1 Term for 14th Multiple Term
ATC1P	ATC Level 1 Term for Primary Term
ATC1P4C	WHO-DDE ATC1-MAIN GROUP-4 C
ATC1T	ATC 1 NAME
ATC1TERM	ATC 1 NAME
ATC1TM	ATC1 TERM
ATC1TXT	ATC1 TEXT
CMATC1C	ATC Code 1
CMATC1CD	ATC 1 CODE
CMATC1D	ATC Level 1 description
CMATC1TX	ATC 1 Text
DGATC1C	ATC1 CODE

QNAM	QLABEL
DGATC1D	ATC1 DECODE
MEDATC1	Eye Preparation ATC1
ORATC1	Original ATC Level 1 Term
ORATC1CD	Original ATC Level 1 Code
WHOATC1	WHO-DDE ATC1-MAIN GROUP
WHOATC1C	WHO-DDE ATC1-MAIN GROUP CODE

Table 7. Examples of implementation WHO Drug ATC1 coding variables

Sponsors implementations of labels for the same non-standard variable are also inconsistent. Table 8 shows an example of different values of QLABEL for QNAM='CMATC1' in SUPPCM dataset:

QNAM	QLABEL
CMATC1	ATC 1 Code
CMATC1	ATC 1 Term
CMATC1	ATC 1 TERM
CMATC1	ATC Level 1 code
CMATC1	ATC level 1 Text 1
CMATC1	ATC Text 1
CMATC1	ATC Text for ATC Level 1
CMATC1	ATC1
CMATC1	ATC1 Term
CMATC1	ATC1 Text
CMATC1	CM ATC 1
CMATC1	Level 1 ATC
CMATC1	Medication ATC1
CMATC1	Medication ATC1 Class
CMATC1	The ATC Level 1 Text

Table 8. Examples of implementation of QLABEL for records with QNAM='CMATC1' in SUPPCM dataset

The industry really needs help with standardization of WHO Drug coding in study data.

COMMENTS

Use of SUPPQUAL datasets instead of dedicated CO domain for collected comments is still a widespread violation of CDISC SDTM standard. However, there is difference in this violation across domains. Table 9 shows presence of Comments variables in most common standard domains. A leader is SUPPLB dataset with Comments records populated in 51 studies which represent 16% of all studies or 20% of studies with SUPPLB dataset. The second one is SUPPPC dataset with comments in 22 studies which represent 16% of studies with SUPPPC dataset. Note, that no Comments records were populated in SUPPDM or SUPPVS datasets. And, it's rarely populated in SUPPAE dataset.

Dataset	N	% (all studies)	% (studies with Dataset)
SUPPAE	1	0.3	0.3
SUPPCM	4	1.2	1.3
SUPPDM	0	0.0	0.0
SUPPDV	8	2.5	4.2
SUPPEX	3	0.9	1.5
SUPPLB	51	15.7	20.2
SUPPPC	22	6.8	15.8
SUPPVS	0	0.0	0.0

Table 8. Number of studies (N) with invalid implementation of Comments in SUPPQUAL datasets

Our understanding is that major drivers to store comments in SUPPQUAL datasets instead of dedicated standard CO domain are either due to convenience to keep all domain related information in SUPPQUAL datasets or lack of implementation experience. Additional educational efforts may be needed to promote CDISC conformance.

MAJOR VIOLATIONS OF CDISC CONFORMANCE

Looking across 325 analyzed studies you can find many possible examples of violation of CDISC SDTM conformance for implementation of SUPPQUAL datasets. Here are some examples of SDTM violations in analyzed data.

There are more than 1,000 variables which represent Timing information in SUPPQUAL datasets:

- 966 unique QNAM with QLABEL which include text “date”: *Date of Best Response, Subject Date of Birth, Data Entry Date, Last Contact Date, Report Date, Randomization Date*, etc.
- 492 unique QNAM with QLABEL which include text “time”: *Randomization Time, Time of onset, Time of blood draw, Actual Time*, etc. Note that some of these non-standard variables are overlapped with “date” variables. Few of them do not represent Timing info (e.g., *Ongoing at Time of Death*)
- Visit variables in datasets like SUPPEX, SUPPDV, SUPPCO, SUPPTR, SUPPLB, SUPPPC, etc.

There are unexpected variables which represent Normal Range information. However, such cases are quite rare. For example: SUPPLB.AGE_HIGH (*Normal Range Upper Limit-Age*), SUPPEG.EGORNRHI (*UPPER NORMAL RANGE VALUE*), SUPPLB.SINORMHI (*SI upper limit of normal range*), SUPPVS.SYSBPHI (*Sys BP Normal Range High*), etc.

Original, previous or supplemental results in Conventional or SI units. For example, in SUPPLB datasets: SIRESN (*SI Numeric Result*), CNVRESC (*Conventional Text Result*), LBORRES4 (*Result or Finding in Original Units*), LBSTRSCN (*Char. Result/Finding in Std Format (N)*), PSTRESC (*Previous Character Result in Std Format*), etc.

Subject Characteristics or other non-applicable information in SUPPDM dataset: DMEMPLO (*Current employment situation*), DEMMAR (*4.Current marital status*), DEMEDU (*6.Highest level of edu?*), P85BMI (*85th Percentile BMI (kg/m2)*), DMBLWT (*Baseline Weight (g)*), BLOODONR (*Blood Donor*), DTH_D (*Day of Death*), IEORRES (*Did subject meet eligibility criteria?*), INITDOSU (*Dose Units*), EXCONC (*Final Study Drug Concentration*), etc.

There are 61 (19%) studies with SUPPSV and 11 (3%) studies with SUPPCO datasets. It seems that incorrectly implemented SUPPSV datasets store all information collected on Subject Visit CRF. For example, SUBJID1 (*Subject Identifier 1 for the study*), TVISYN (*Is This a Treatment Visit?*), SVASSESS (*Assessments Performed*), SVUPDES1 (*Description of Unplanned Visit*), VISLB (*Lab Collection*), DOVDTC (*Date of Visit*), OTHERSP (*If Other, specify*), etc.

There are many studies where SUPPQUAL datasets store data management or tracking information which is not applicable for regulatory submissions.

For example, one of analyzed studies includes 618 non-standard variables in SUPPQUAL datasets. It looks like these variables represent raw data collected in EDC system. For example, SUPPAE.AESERN (*Serious Event (N)*), SUPPAE.AEST_Y (*Start Year of Adverse Event*), SUPPAE. EPOCHN (*Epoch (N)*) SUPPDM.RACEN (*Race (N)*), etc.

Only 44% of non-standard variables in SUPPQUAL datasets follow a good implementation practice to create SDTM variable name with prefix corresponding to domain value. For example, SUPP**AE**.**AE**TRTEM, SUPP**EG**.**EG**CLSIG, SUPP**CM**.**CM**ATC1, SUPP**XX**.**XX**ZYUN, etc.

UTILIZATION CDISC TAUGS

As we mentioned before CDISC has a special project which tracks new non-standard variables introduced in TAUGs. There is a special 'SDTM NSV Registry' page on CDISC Wiki [3]. It includes a list of 173 variables used as new non-standard variables across 40+ existing CDISC TAUGs. 142 of these variables are unique.

In analyzed 324 studies we found SUPPQUAL variables which match 24 (17%) out of 142 unique CDISC Non-Standard Variables (NSV). Such low ratio indicates a low utilization of existing CDISC TAUG by the industry.

Also, in many cases the industry implementation of non-standard variables in SUPPQUAL datasets is not consistent with CDISC. The most notorious variable is --SPEC with CDISC interpretation as '*Specimen Type*'. This --SPEC non-standard variable was implemented in 24 SUPPQUAL datasets. 11 of them have different interpretations of --SPEC variable as '*Other, **Spec**ify*', '*Other Symptom*', '*AE of **Spec**ial Interest*', '*Disposition Specifications'*, '*AE **Spec**ify*', '*Abnormal, **Spec**ify*', etc.

More active promotion of CDISC TAUGs is expected.

CONCLUSION

This study was run as a pilot to understand the potential use of industry metrics for improving standards management practices and to test methodology.

A major challenge in our research was the analysis of collected metrics which is partially manual. Expecting this issue, we limited the number of studies/sponsors to ~300/~100. Such approach still produced 27K records of non-standard variables. Due to lack of standardization of SUPPQUAL variables, we were limited in ability to automate the analysis. For example, WHO Drug coding was represented by at least 667 different SUPPQUAL variables (QNAM). Reviewing each of 27K records and grouping them into specific type of information cannot be fully automated at this stage. For example, our decisions were mostly relied on variable labels (QLABEL). In most times, labels have enough descriptive details to make decision about type of information stored in the variables. However, in some cases implementation of variable labels is not sufficient to understand content of non-standard variable. For example, QLABEL could be populated as a copy of QNAM value. That's why some of our numbers in this paper could be below actual values. For example, some WHO Drug coding variables may be missed in our analysis due to confusing or misleading labels (QLABEL).

Another challenge is that due to lack of standardization some information may be hidden. For example, a Flag of Clinically Significant finding is well adopted non-standard variable. However, in some studies this information is merged with information which represent Normal/Abnormal Result Interpretation by terms '*Normal', 'Abnormal, Not Clinically Significant' and 'Abnormal, Clinically Significant'*. Analysis of such cases is not possible in our investigation because collected metrics does not have study result details stored in QVAL variable.

We found other methodological challenges to address when performing our pilot. For example, some sponsors may implement SDTM+ structure and keep non-standard variables attached to standard domains. Study status (ongoing, completed, ready for submission, FDA vs. PMDA, etc.) is important when doing analysis of collected metrics. For example, the same study may have several data packages (or data cuts) with different implementation of SUPPQUAL domains.

This pilot research created suggestions for additional analysis of the industry implementation of non-standard variables. For example, how consistent is the implementation of SUPPQUAL within each company? During conduct of this pilot we saw both cases. Some companies are very consistent in implementation of SUPPQUAL variables, while others may have the same QNAM value, but different meaning within a submission or even within a study.

Some collected metrics are not included in study results. For example, majority of recent studies are implemented with only two versions SDTM-IG. Therefore, it's not enough data to reveal potential correlation of SUPPQUAL implementation with version of SDTM-IG.

A major goal of our study was to see how we can improve current standards management practices. We believe that existing standards and regulatory guidance documents are underutilized or ignored. For example, more than 1/3 of analyzed studies do not have AE Treatment Emergent Flag information in tabulation data requested by both FDA and PMDA. Implementation of Clinically Significant non-standard SUPPQUAL variable still vary despite a clear guidance in SDTM Implementation Guide Appendix C2. Additional educational efforts in promotion of data standards and regulatory requirements are expected.

Some information is utilized in almost every study but is not represented by standard SDTM variables yet. Clinical Significance and WHO Drug Coding are the most common examples.

There is still a common practice of misuse and incorrect mapping of collected data into SUPPQUAL datasets. While some companies manage SUPPQUAL metadata to ensure consistent implementation within organization, it's still an issue for others. Education efforts are expected to promote good SDTM mapping practices. New validation rules may also help.

We believe that collection and analysis of the industry implementation metrics can help identify global implementation issues and help with their eventual resolution.

REFERENCES

1. Dirk Van Krunckelsven. %Supp2Par_v1 Merging Supplemental Data onto Parent Domains. PhUSE EU Connect (2015). Available at https://www.lexjansen.com/phuse/2015/cs/CS01.pdf
2. Dirk Van Krunckelsven (2013). Supp2par_v1 as distributed in the PhUSE Standard scripts repository on GitHub: https://github.com/i-akiya/phuse-scripts/blob/master/lang/SAS/datahandle/supp2par/src/Supp2Par_v1.sas
3. SDTM NSV Registry. CDISC Wiki. Available at https://wiki.cdisc.org/pages/viewpage.action?spaceKey=GGG&title=SDTM+NSV+Registry
4. FDA Data Standards Catalog v6.1 September 9, 2019). Available at https://www.fda.gov/media/85137/download

CONTACT INFORMATION

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

Sergiy Sirichenko
Pinnacle 21
+1.908.781.2342
sergiy@pinnacle21.com
www.pinnacle21.com