PharmaSUG 2022 - Paper ss-079

eCOA and SDTM: Bringing Order to the Wild West

Vincent Allibone, YPrime, UK

Monali Khanna, YPrime, USA

Frank Menius, YPrime, USA

ABSTRACT

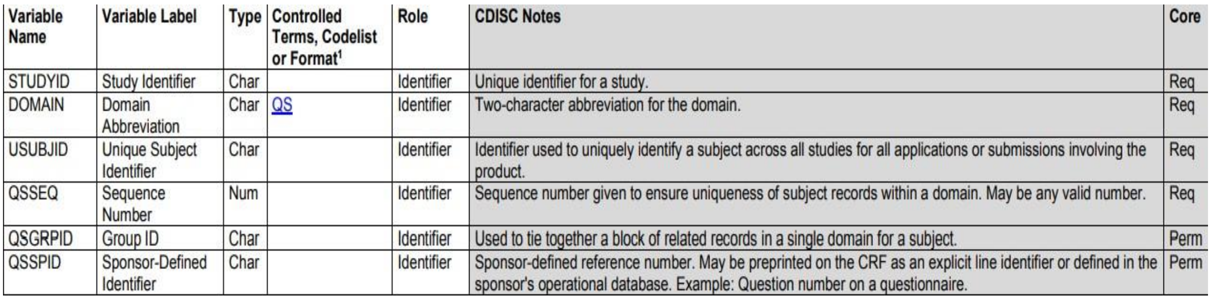
As the implementation of electronic Clinical Outcome Assessments (eCOA) continues to grow, the need for data standards is more apparent than ever. Data standards allow data to be processed and exported more efficiently, leaving us with high quality, organized data. The Clinical Data Interchange Standards Consortium (CDISC) have developed a model for organizing and formatting data in the form of the Study Data Tabulation Model (SDTM), which, when implemented, allows data to be submitted in an organized and clean way. In theory, the application of SDTM to eCOA data is straightforward, however, to the outsider or newcomer, it can seem a daunting and confusing process. With several different data classes, domains, variable roles and core variables that data could fall under, annotating screen reports for eCOA data seems to be straightforward for the experienced and an almost impossible task for the inexperienced. This paper will share tips and techniques of learning standards and applying them, specifically, to eCOA data by providing real-life examples, whilst summarizing some of the core guidelines of SDTM.

INTRODUCTION

Electric Clinical Outcome Assessments (eCOA) are incredibly prevalent in clinical trials, allowing subjects to answer health-based questionnaires from the comfort of their own home. Though the ease with which data is captured, is not always mirrored in how that data is mapped. eCOA produces a large amount of raw data and narrowing that down to conform with SDTM standards, for submission of data can be a minefield, especially to the beginner. This paper will aim to focus on how eCOA data can be mapped, using the QS (Questionnaire) domain outlined in the SDTMIG version 3.4 section 6.3.9.2. We will be using the European Quality of Life Five Dimension Five Scale Level (EQ-5D-5L) as an example of a questionnaire, in order to show how data can go from being captured to being submission-ready.

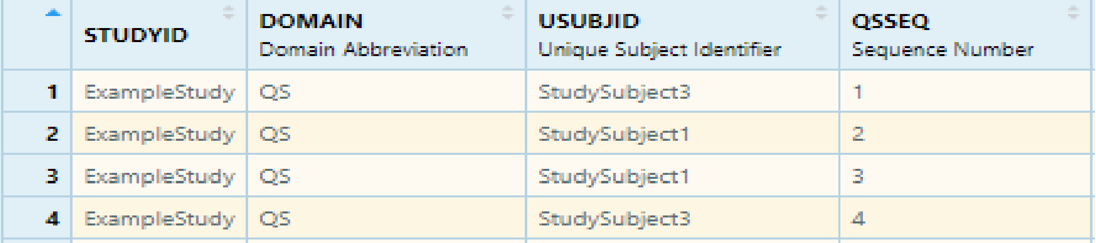
THE IDENTIFIERS

Initially, let us explore any variables that would need to be included in the QS domain, with Identifier roles. As the name suggests, these variables are used to identify where an observation, or row, of data originates from. The SDTMIG sets out 6 variables that would be considered Identifier variables (image below), with four of these being required in the data, with the other two being considered permissible, they can be included, but it is not necessary. Let us take a look at these in a bit more detail:



Variable Name	Variable Label	Type	Controlled Terms, Codelist or Format[1]	Role	CDISC Notes	Core
STUDYID	Study Identifier	Char		Identifier	Unique identifier for a study.	Req
DOMAIN	Domain Abbreviation	Char	QS	Identifier	Two-character abbreviation for the domain.	Req
USUBJID	Unique Subject Identifier	Char		Identifier	Identifier used to uniquely identify a subject across all studies for all applications or submissions involving the product.	Req
QSSEQ	Sequence Number	Num		Identifier	Sequence number given to ensure uniqueness of subject records within a domain. May be any valid number.	Req
QSGRPID	Group ID	Char		Identifier	Used to tie together a block of related records in a single domain for a subject.	Perm
QSSPID	Sponsor-Defined Identifier	Char		Identifier	Sponsor-defined reference number. May be preprinted on the CRF as an explicit line identifier or defined in the sponsor's operational database. Example: Question number on a questionnaire.	Perm

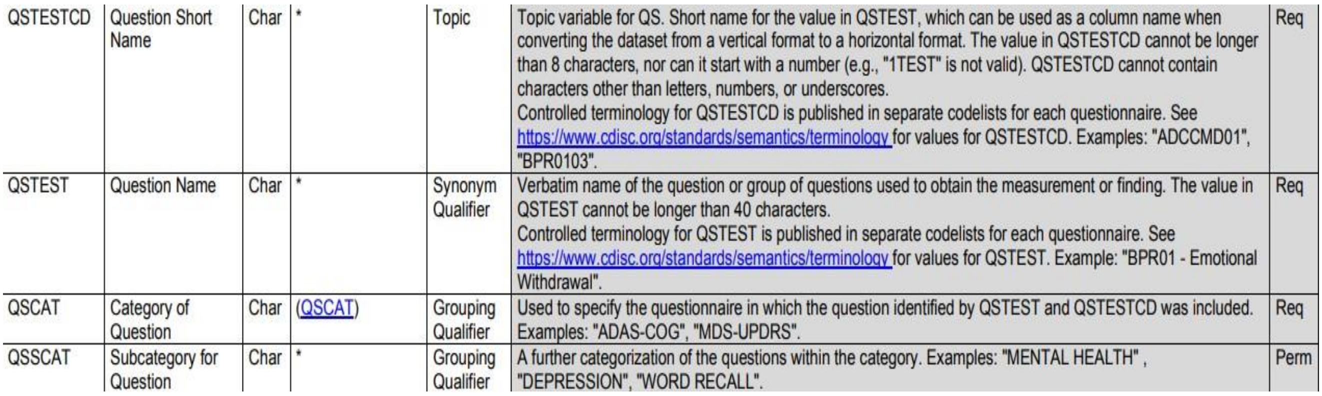
Firstly, we have the STUDYID and the DOMAIN – these two variables are included in all data that conforms to SDTM standards. The STUDYID is the unique identifier for a study, usually in character format and the DOMAIN is which domain is being used to map the data, in this case QS and is also in character format. The next two are the USUBJID and QSSEQ, the USUBJID is the unique subject identifier, in character format and would look something like this ‘StudyName1234’. The QSSEQ is the sequence number, this is included to make sure that all records are unique and is in numeric format. All these variables are required, whereas the QSSPID and QSGRPID are not. An example below shows how the 4 required variables would be displayed in submission data. These 4 are important for all data submission, not just data in the QS domain.



	STUDYID	DOMAIN Domain Abbreviation	USUBJID Unique Subject Identifier	QSSEQ Sequence Number
1	ExampleStudy	QS	StudySubject3	1
2	ExampleStudy	QS	StudySubject1	2
3	ExampleStudy	QS	StudySubject1	3
4	ExampleStudy	QS	StudySubject3	4

QUESTIONS CATEGORIES

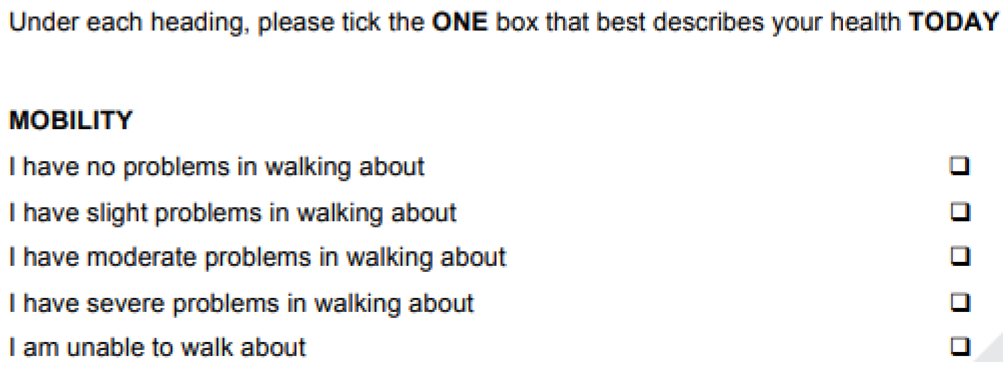
The next section that the SDTMIG sets out for data in the QS domain are the topic, synonym and grouping qualifiers. These variables define the questions being answered and the categories of those questions:



QSTESTCD	Question Short Name	Char	*	Topic	Topic variable for QS. Short name for the value in QSTEST, which can be used as a column name when converting the dataset from a vertical format to a horizontal format. The value in QSTESTCD cannot be longer than 8 characters, nor can it start with a number (e.g., "1TEST" is not valid). QSTESTCD cannot contain characters other than letters, numbers, or underscores. Controlled terminology for QSTESTCD is published in separate codelists for each questionnaire. See https://www.cdisc.org/standards/semantics/terminology for values for QSTESTCD. Examples: "ADCCMD01", "BPR0103".	Req
QSTEST	Question Name	Char	*	Synonym Qualifier	Verbatim name of the question or group of questions used to obtain the measurement or finding. The value in QSTEST cannot be longer than 40 characters. Controlled terminology for QSTEST is published in separate codelists for each questionnaire. See https://www.cdisc.org/standards/semantics/terminology for values for QSTEST. Example: "BPR01 - Emotional Withdrawal".	Req
QSCAT	Category of Question	Char	(QSCAT)	Grouping Qualifier	Used to specify the questionnaire in which the question identified by QSTEST and QSTESTCD was included. Examples: "ADAS-COG", "MDS-UPDRS".	Req
QSSCAT	Subcategory for Question	Char	*	Grouping Qualifier	A further categorization of the questions within the category. Examples: "MENTAL HEALTH" , "DEPRESSION", "WORD RECALL".	Perm

We’ll look at an example of a question from the EQ-5D-5L questionnaire, and then see how it may be mapped to these variables using an example data set. This is an example of a question on mobility of a subject:

Figure 1: EQ-5D-5L (UK English sample version)

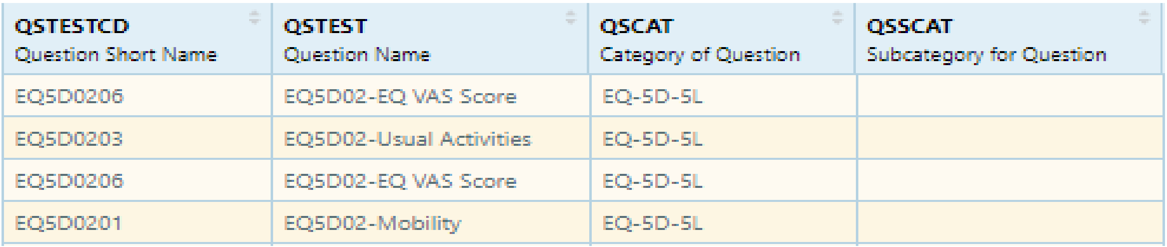


Under each heading, please tick the **ONE** box that best describes your health **TODAY**

MOBILITY

I have no problems in walking about	☐
I have slight problems in walking about	☐
I have moderate problems in walking about	☐
I have severe problems in walking about	☐
I am unable to walk about	☐

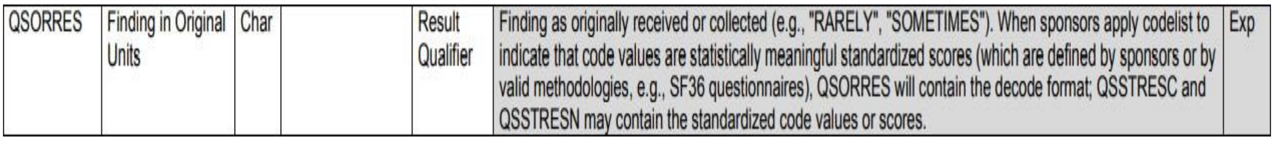
If we look at the four variables that we are trying to define, the first would be the QSTESTCD, which is the short name for the question. This can be a maximum of 8 characters. On the CDISC website, you can find an annotated version of EQ-5D-5L, which names the QSTESTCD for the mobility question as EQ5D020. The next is QSTEST, which is the full name of the question and can be no longer than 40 characters, for this question the QSTEST is EQ5D02 – Mobility. Moving on to the QSCAT, this is the name of the questionnaire, which in this circumstance is EQ-5D-5L. The final variable is the QSSCAT, this is a further categorization of the QSCAT, however, this is a permissible variable and is not necessary for the EQ-5D-5L, whereas the other three are all required. In a data set, these variables may look something like this:



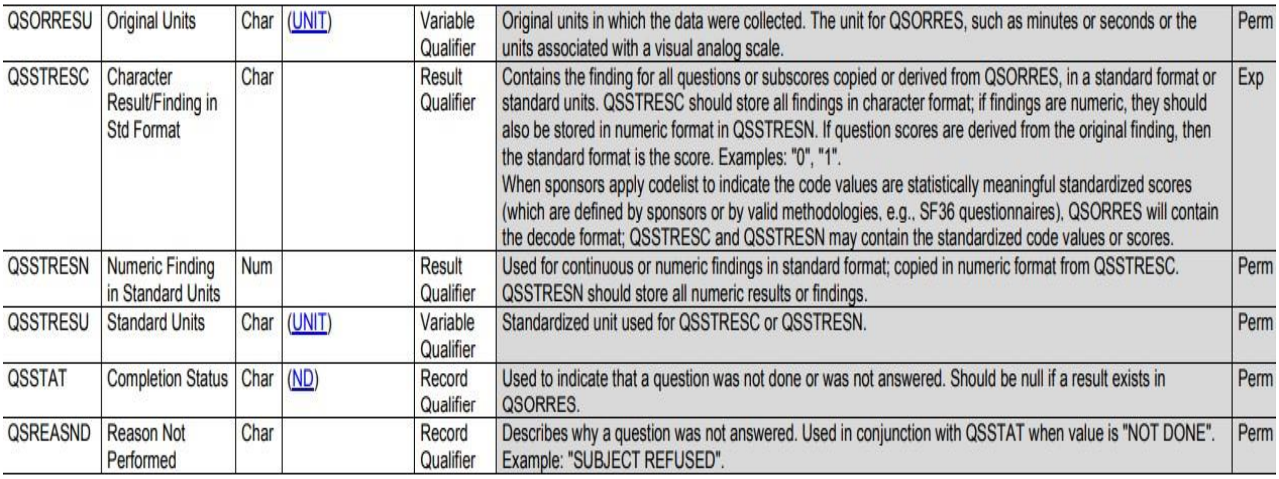
QSTESTCD Question Short Name	QSTEST Question Name	QSCAT Category of Question	QSSCAT Subcategory for Question
EQ5D0206	EQ5D02-EQ VAS Score	EQ-5D-5L	
EQ5D0203	EQ5D02-Usual Activities	EQ-5D-5L	
EQ5D0206	EQ5D02-EQ VAS Score	EQ-5D-5L	
EQ5D0201	EQ5D02-Mobility	EQ-5D-5L	

THE ANSWERS

The next category of variables that we are going to look at are the variables that are determined by the answer given by the user. These are comprised of the variable, result, and record qualifiers, and we will look at each of these in more detail, as I will break them down into smaller categories, and give examples in data for all of them:



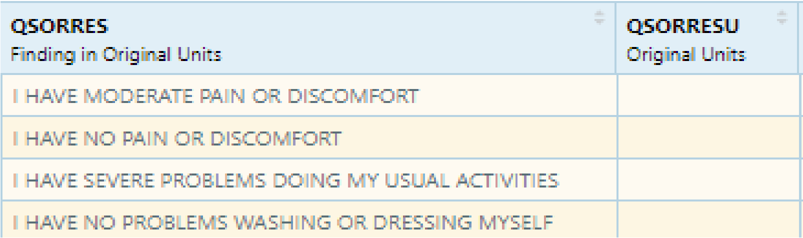
QSORRES	Finding in Original Units	Char		Result Qualifier	Finding as originally received or collected (e.g., "RARELY", "SOMETIMES"). When sponsors apply codelist to indicate that code values are statistically meaningful standardized scores (which are defined by sponsors or by valid methodologies, e.g., SF36 questionnaires), QSORRES will contain the decode format; QSSTRESC and QSSTRESN may contain the standardized code values or scores.	Exp



QSORRESU	Original Units	Char	(UNIT)	Variable Qualifier	Original units in which the data were collected. The unit for QSORRES, such as minutes or seconds or the units associated with a visual analog scale.	Perm
QSSTRESC	Character Result/Finding in Std Format	Char		Result Qualifier	Contains the finding for all questions or subscores copied or derived from QSORRES, in a standard format or standard units. QSSTRESC should store all findings in character format; if findings are numeric, they should also be stored in numeric format in QSSTRESN. If question scores are derived from the original finding, then the standard format is the score. Examples: "0", "1". When sponsors apply codelist to indicate the code values are statistically meaningful standardized scores (which are defined by sponsors or by valid methodologies, e.g., SF36 questionnaires), QSORRES will contain the decode format; QSSTRESC and QSSTRESN may contain the standardized code values or scores.	Exp
QSSTRESN	Numeric Finding in Standard Units	Num		Result Qualifier	Used for continuous or numeric findings in standard format; copied in numeric format from QSSTRESC. QSSTRESN should store all numeric results or findings.	Perm
QSSTRESU	Standard Units	Char	(UNIT)	Variable Qualifier	Standardized unit used for QSSTRESC or QSSTRESN.	Perm
QSSTAT	Completion Status	Char	(ND)	Record Qualifier	Used to indicate that a question was not done or was not answered. Should be null if a result exists in QSORRES.	Perm
QSREASND	Reason Not Performed	Char		Record Qualifier	Describes why a question was not answered. Used in conjunction with QSSTAT when value is "NOT DONE". Example: "SUBJECT REFUSED".	Perm

QSORRES and QSORRESU

The first of these variables is the QSORRES, this variable is expected, meaning that it has to be included but can be populated with NULL. The QSORRES is the finding as it was initially collected, so, if we go back to our example of the mobility question from earlier, the QSORRES would be any one of the five answers given. It does not have a word limit; however, most variables need to stay within a 200 characters limit maximum. The next one, QSORRESU, is permissible as it only needs to collect the units that QSORRES was given in, if applicable.

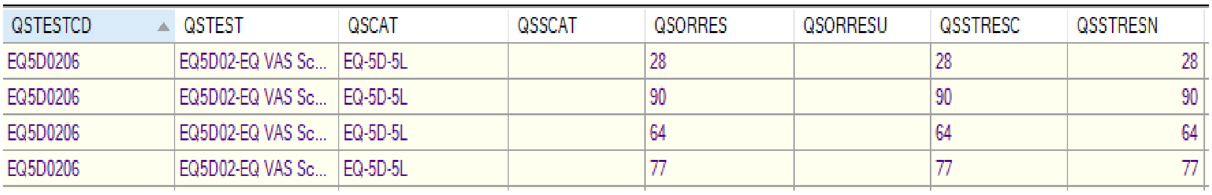


QSORRES Finding in Original Units	QSORRESU Original Units
I HAVE MODERATE PAIN OR DISCOMFORT	
I HAVE NO PAIN OR DISCOMFORT	
I HAVE SEVERE PROBLEMS DOING MY USUAL ACTIVITIES	
I HAVE NO PROBLEMS WASHING OR DRESSING MYSELF	

QSSTRESC, QSSTRESN and QSSTRESU

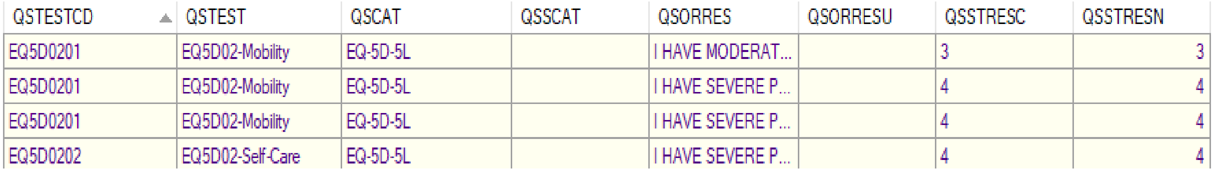
These three variables are dependent on each other, in terms of how they are populated. QSSTRESC is expected, whereas the other two are permissible. QSSTRESC will have the original findings, which can be copied from QSORRES. It is given in character format, however, if the original finding was in numeric form, it will still display the numeric form. QSSTRESN will display the answer in numeric format. This means each answer is assigned a specific number, in the mobility question answers go from 1 to 5, so if a subject answers ‘I have no problems walking about’, the QSSTRESN will be 1 whereas if a subject answers ‘I am unable to walk about’ this will be displayed as 5. The QSSTRESU will display the units for the answer, if applicable. If we look at the VAS Score section of the EQ-5D-5L, we can give ourselves a better idea of how these variables all marry together when an answer is numeric and below that is an example from the mobility question as this is in character format

VAS Score:



QSTESTCD	QSTEST	QSCAT	QSSCAT	QSORRES	QSORRESU	QSSTRESC	QSSTRESN
EQ5D0206	EQ5D02-EQ VAS Sc...	EQ-5D-5L		28		28	28
EQ5D0206	EQ5D02-EQ VAS Sc...	EQ-5D-5L		90		90	90
EQ5D0206	EQ5D02-EQ VAS Sc...	EQ-5D-5L		64		64	64
EQ5D0206	EQ5D02-EQ VAS Sc...	EQ-5D-5L		77		77	77

Mobility:



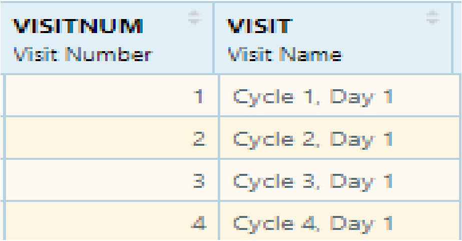
QSTESTCD	QSTEST	QSCAT	QSSCAT	QSORRES	QSORRESU	QSSTRESC	QSSTRESN
EQ5D0201	EQ5D02-Mobility	EQ-5D-5L		I HAVE MODERAT...		3	3
EQ5D0201	EQ5D02-Mobility	EQ-5D-5L		I HAVE SEVERE P...		4	4
EQ5D0201	EQ5D02-Mobility	EQ-5D-5L		I HAVE SEVERE P...		4	4
EQ5D0202	EQ5D02-Self-Care	EQ-5D-5L		I HAVE SEVERE P...		4	4

QSSTAT and QSREASND

The last two variables that we will look at in this section are the QSSTAT and QSREASND variables. These are used in the event that a question has not been answered. If a question has been missed, this will be recorded in the QSSTAT, and the QSREASND will give the reason for it being missed. If QSORRES is populated, the QSSTAT will be populated with NULL.

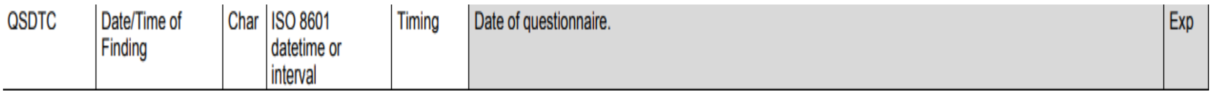
THE VISIT SCHEDULE

Two of the variables that I would like to discuss from a subject's visit schedule are the VISITNUM and VISIT variables. Although only the VISITNUM is expected, the VISIT variable also gives the reader of the data an indication of the visit, without having to have the correlation between visit number and visit name somewhere separate. The VISITNUM variable is in numeric format and is the numeric version of the visit's name. For sake of ease, when sorting through data, the VISITNUM is used as the visit's name can be somewhat length and can often look similar to other visit names. The VISIT variable is in character format and is the original name of the visit. Again, this is a permissible variable, but it would be my preference to use it. Here's how they may look in data:



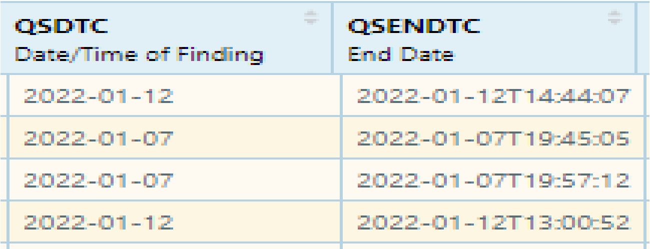
VISITNUM Visit Number	VISIT Visit Name
1	Cycle 1, Day 1
2	Cycle 2, Day 1
3	Cycle 3, Day 1
4	Cycle 4, Day 1

The final variable that I will be discussing, that is required as per the SDTMIG, is the QSDTC:



QSDTC	Date/Time of Finding	Char	ISO 8601 datetime or interval	Timing	Date of questionnaire.	Exp

This variable is the date of any observation. It is in ISO8601 format, meaning that it should look like – YYYY-MM-DD. If you want to take this further, although not specified in the SDTMIG, you can add the QSENDTC variable. This is the end date of an observation and marks the time of its completion. It is also in ISO8601 format and would look like this:

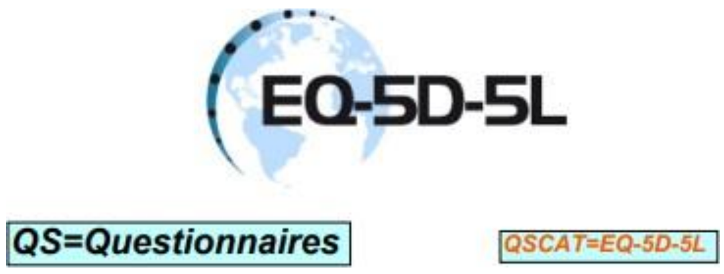


QSDTC Date/Time of Finding	QSENDTC End Date
2022-01-12	2022-01-12T14:44:07
2022-01-07	2022-01-07T19:45:05
2022-01-07	2022-01-07T19:57:12
2022-01-12	2022-01-12T13:00:52

ANNOTATED CRF

Now that we've had a look at the different variables that you may include in data collected from eCOA, let's take a peek at the annotated CRF of the EQ-5D-5L from the CDISC website, to see some of the variables first-hand, in order to be able to relate them back to the standards:

The first is a screenshot from the top of the report form, which shows that it is mapped to the QS domain, and gives us the QSCAT (the name of the questionnaire):

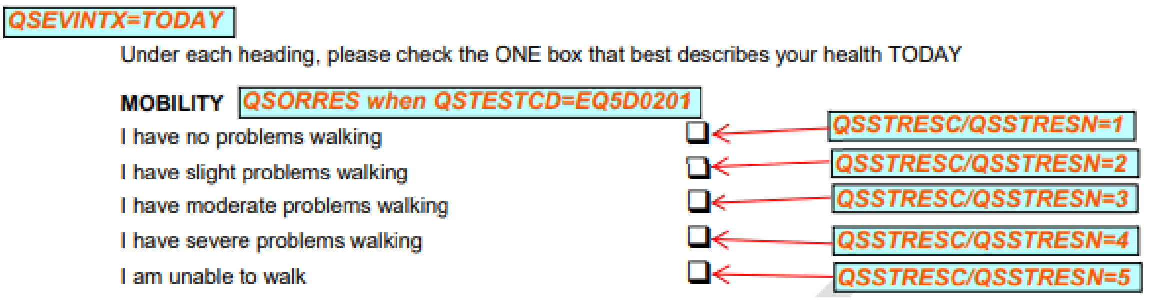


EQ-5D-5L

QS=Questionnaires

QSCAT=EQ-5D-5L

The next one that we'll look at is a screenshot from the Mobility question, so we can also see how that would be mapped:



QSEVINTX=TODAY

Under each heading, please check the ONE box that best describes your health TODAY

MOBILITY QSORRES when QSTESTCD=EQ5D0201

I have no problems walking □ QSSTRESC/QSSTRESN=1

I have slight problems walking □ QSSTRESC/QSSTRESN=2

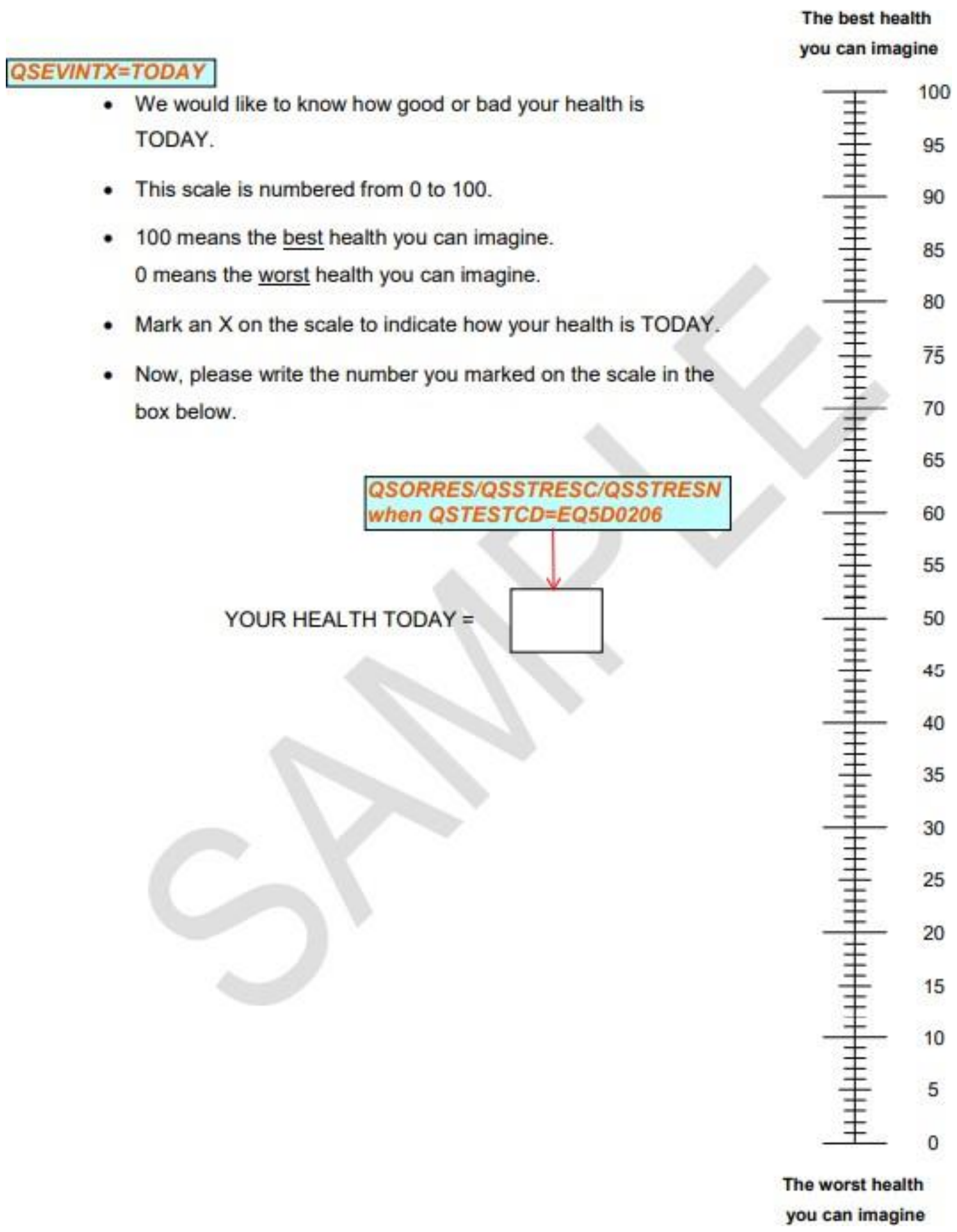
I have moderate problems walking □ QSSTRESC/QSSTRESN=3

I have severe problems walking □ QSSTRESC/QSSTRESN=4

I am unable to walk □ QSSTRESC/QSSTRESN=5

We can see here that it gives us the QSTESTCD which, again, is the question short name. It also tells us that when the QSTESTCD is equal to EQ5D0201 (the mobility question), the QSORRES can be any one of the answers above. The QSSTRESC will copy the QSORRES as the answer is in character format, whereas the QSSTRESN will be mapped with the first answer being 1, the second 2, and so on.

The final screenshot that we will be looking at is from the VAS score:



QSEVINTX=TODAY

- We would like to know how good or bad your health is TODAY.
- This scale is numbered from 0 to 100.
- 100 means the best health you can imagine.
  0 means the worst health you can imagine.
- Mark an X on the scale to indicate how your health is TODAY.
- Now, please write the number you marked on the scale in the box below.

QSORRES/QSSTRESC/QSSTRESN when QSTESTCD=EQ5D0206

YOUR HEALTH TODAY =

The best health you can imagine
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
The worst health you can imagine
SAMPLE

From this, we can also see the QSTESTCD, and we know that the QSORRES, QSSTRESC and QSSTRESN will all be the same, as the answer is always given in numerical form to this question.

CONCLUSION

In conclusion, it can be difficult to the untrained eye to map any data to SDTM standards, let alone eCOA data. However, by trying to understand how each data point can be mapped to their respective variables, we can begin to see how mapping is accomplished and, also, how important it is in keeping submission data orderly. Initially, what can seem a daunting task is made far easier by breaking down your data into its respective sections, along with the fantastic work done at CDISC.

CONTACT INFORMATION

Vincent Allibone
YPrime
Sandwich Discovery Park
Sandwich, Kent
vallibone@yprime.com
YPrime: https://www.yprime.com

Monali Khanna
YPrime
9 Great Valley Parkway, Malvern, PA 19355
Malvern, PA 19355
mkhanna@yprime.com
YPrime: https://www.yprime.com/
LinkedIn: https://www.linkedin.com/in/monali-k-a259a5178/

Frank Menius
YPrime
3301 Benson Dr.
Raleigh, NC 27609
919-412-6537
fmenius@yprime.com
YPrime: https://www.yprime.com/
LinkedIn: https://www.linkedin.com/in/frank-menius/