Mapping Nested e-CRF Data into SDTM Domains

Venkata Sekhar Bhamidipati, Merck & Co., Inc. & Kristi L. Garner, Merck & Co., Inc.

ABSTRACT

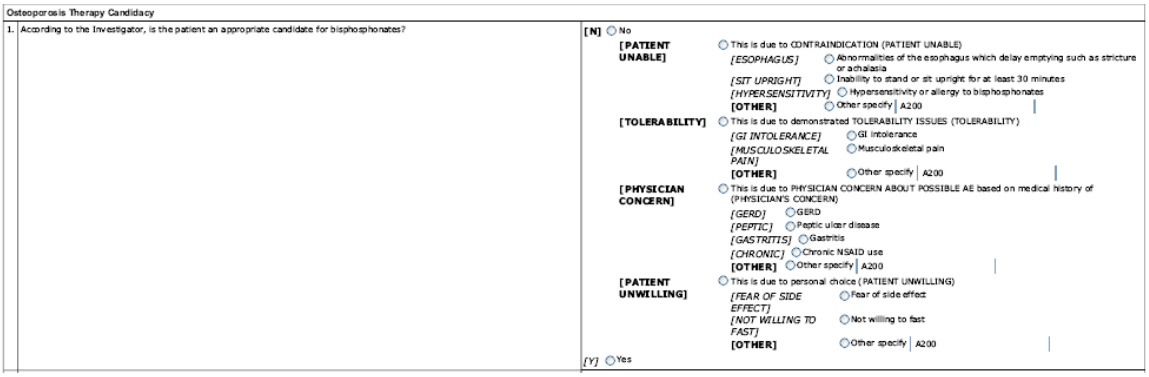
Mapping nested e-CRF data (e-CRF data points which trigger another set of data points to be collected) into SDTM domains is one of the most difficult and complex tasks in implementing the CDISC Study Data Tabulation Model (SDTM). This paper describes the complexities and considerations involved when mapping nested e-CRF data, as well as identifying three different approaches by which to map nested e-CRF data into SDTM domains. A comprehensive illustration of the advantages and disadvantages of each approach is presented. The intended audience of this paper is those people involved with SAS ® Programming, Data Management, and Statistics.

INTRODUCTION

A nested e-CRF is an e-CRF which contains questions that trigger the collection of subsequent data points based on the prior collected response. Mapping these nested e-CRF data points into SDTM domains is complex due to the dynamic nature of the data and the necessity of following the set of rules established by CDISC (Clinical Data Interchange Standards Consortium) SDTM mapping guidelines. The main objectives of this paper are to discuss some of the difficulties encountered when mapping nested e-CRF data, provide illustrations of the mapping of a nested e-CRF using three different approaches, and also identify the advantages and disadvantages of each approach for involved personnel, including SAS Programmers, Data Management, and Statistics. The sample mappings comply with the current SDTM implementation guide. CDISC standards, SDTM mappings guidelines, and specific SDTM domains identification for e-CRF forms are out of scope for this paper.

EXAMPLE OF A NESTED e-CRF

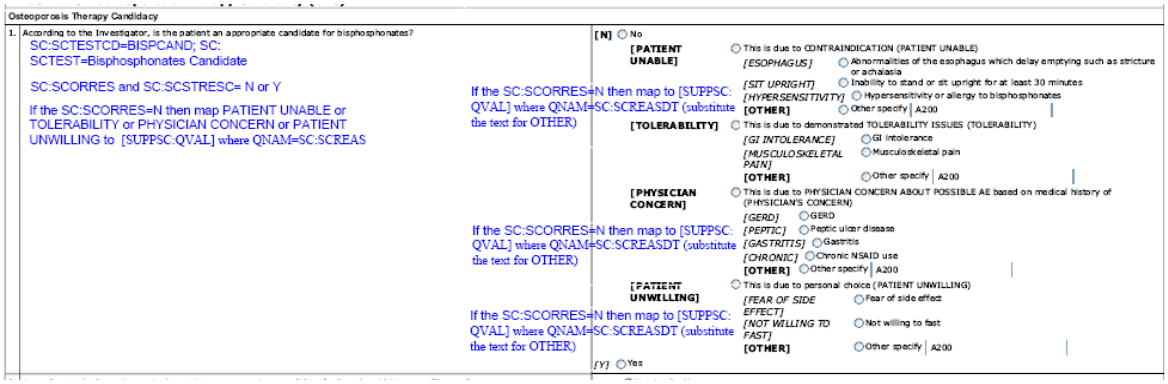
The following is a sample nested e-CRF. The question being posed is "According to the Investigator, is the patient an appropriate candidate for bisphosphonates?" If the answer is 'Yes', then the data collection is complete. However, if the answer is 'No', a selection is required from another set of radio buttons which, in turn, requires a selection from a third set of radio buttons. The final output data structure is entirely dependent on the series of responses for each subject. The dynamic nature of the data makes it difficult to choose the appropriate approach to map into the proper SDTM domains. Even when properly mapped, various difficulties may be encountered by programmers, statisticians, and data management personnel, depending on the approach used.



MAPPING NESTED e-CRF

In this section, let's examine three different approaches of mapping nested e-CRF into SDTM domains and the advantages and disadvantages of each approach. Mapping of the SDTM domains strictly follows the guidelines from CDISC. The sample e-CRF is identified as subject characteristic data which is mapped to the SC domain.

APPROACH 1 (WITH TWO SUPPQUAL VARIABELS)



Approach 1 Mock dataset with a subset of the possible data points with SUPPQUAL variables appended to the SC domain

SPID	SCCAT	SCTEST	SCTESTCD	SCORRES	SCSTRESC	SUPPSC: SCREAS	SUPPSC: SCREASDT
1	OSTEOPOROSIS THERAPY	Bisphosphonates Candidate	BISPCAND	N	N	PATIENT UNABLE	SIT UPRIGHT
2	OSTEOPOROSIS THERAPY	Bisphosphonates Candidate	BISPCAND	N	N	TOLERABILITY	OTHER : ENTER THE FREE TEXT
3	OSTEOPOROSIS THERAPY	Bisphosphonates Candidate	BISPCAND	N	N	PHYSICIAN CONCERN	GERD
4	OSTEOPOROSIS THERAPY	Bisphosphonates Candidate	BISPCAND	N	N	PATIENT UNWILLING	FEAR OF SIDE EFFECT
5	OSTEOPOROSIS THERAPY	Bisphosphonates Candidate	BISPCAND	Y	Y		

Mapping approach

The first approach aims to capture every data point on the e-CRF into a single record for each patient for each visit. However, the SC domain does not have all the suitable variables needed to map into one record for each patient for each visit since SDTM does not allow additional variables to be added to core domains. Therefore, the Supplemental Qualifiers special purpose dataset (SUPP--) can be used to capture non-standard variables and their association to parent records. Using this methodology, two supplemental qualifier variables SUPPSC:SCREAS and SUPPSC:SCREASDT are created and utilized to capture data for the dynamically triggered responses.

Advantages

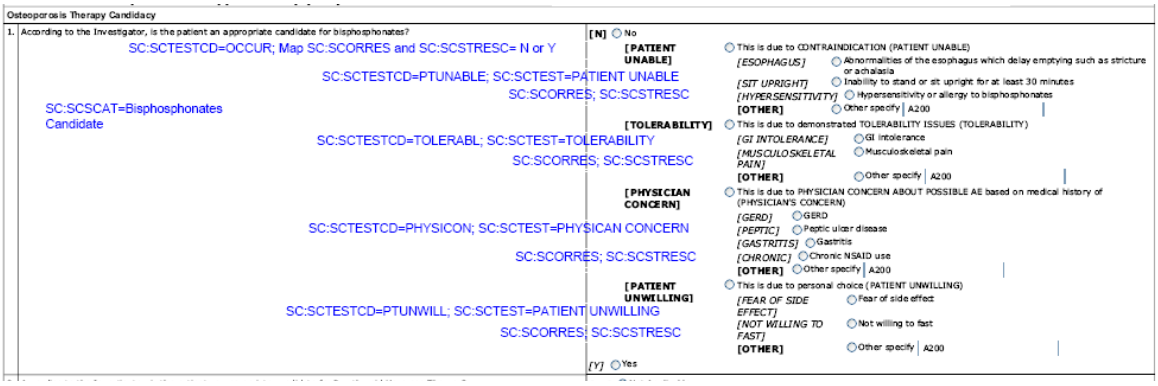
- The main advantage of this approach is that all the data can be captured in a single record for each patient for each visit (i.e. after merging the SUPPSC domain variables to the SC parent domain).
- For statisticians and SAS programmers, it becomes easy to utilize the data and apply programming logic without implementing further transformations.

- For data management, writing data quality checking reports and edit checks are simpler and relatively straightforward, since all data is contained in a single record.

Disadvantages

- The primary disadvantage is mapping and handling the SUPP-- dataset. Before adding each additional SUPPQUAL variable, the programming team must look for consistency across projects and seek approval from the standard mapping team.
- Depending on the e-CRF, this SUPP-- dataset may become extremely large. Each additional SUPPQUAL variable creates a record, per subject per visit, in the SUPP-- dataset and this may introduce data management difficulties due to the huge number of records.
- The amount of time needed to extract the volume of data in the SUPP-- dataset becomes longer.
- Additional programs need to be written and maintained to combine the SUPP-- domain with the parent domain.

APPROACH 2 (NO SUPP-- DOMAIN)



Approach 2 Mock dataset with a subset of possible data points

SPID	SCCAT	SCSCAT	SCTEST	SCTESTCD	SCORRES	SCSTRESC
1	OSTEOPOROSIS THERAPY	Bisphosphonates Candidate	OCCUR	OCCUR	N	N
1	OSTEOPOROSIS THERAPY	Bisphosphonates Candidate	PATIENT UNABLE	PTUNABLE	SIT UPRIGHT	SIT UPRIGHT
2	OSTEOPOROSIS THERAPY	Bisphosphonates Candidate	OCCUR	OCCUR	N	N
2	OSTEOPOROSIS THERAPY	Bisphosphonates Candidate	TOLERABILITY	TOLERABL	OTHER : FREE TEXT	OTHER : FREE TEXT
3	OSTEOPOROSIS THERAPY	Bisphosphonates Candidate	OCCUR	OCCUR	N	N
3	OSTEOPOROSIS THERAPY	Bisphosphonates Candidate	PHYSICIAN CONCERN	PHYSICON	GERD	GERD
4	OSTEOPOROSIS THERAPY	Bisphosphonates Candidate	OCCUR	OCCUR	N	N
4	OSTEOPOROSIS THERAPY	Bisphosphonates Candidate	PATIENT UNWILLING	PTUNWILL	FEAR OF SIDE EFFECT	FEAR OF SIDE EFFECT
5	OSTEOPOROSIS THERAPY	Bisphosphonates Candidate	OCCUR	OCCUR	Y	Y

Mapping approach

The second approach is trying to capture every data point on the e-CRF without using SUPP-- domain. Since the SC domain does not have all the suitable variables to map all the data into a single record for each patient for each visit, this alternative approach captures every data point in either one or two records for each patient each visit. If the response to the question is 'Yes' then only one record is needed because there are no dynamically triggered responses, otherwise, two records are created. This method also requires that test codes (OCCUR (this test code captures the proxy answer), PTUNABLE, TOLERABL, PHYSICON, PTUNWILL (these last 4 test codes capture the other data points)) be created for all possible responses.

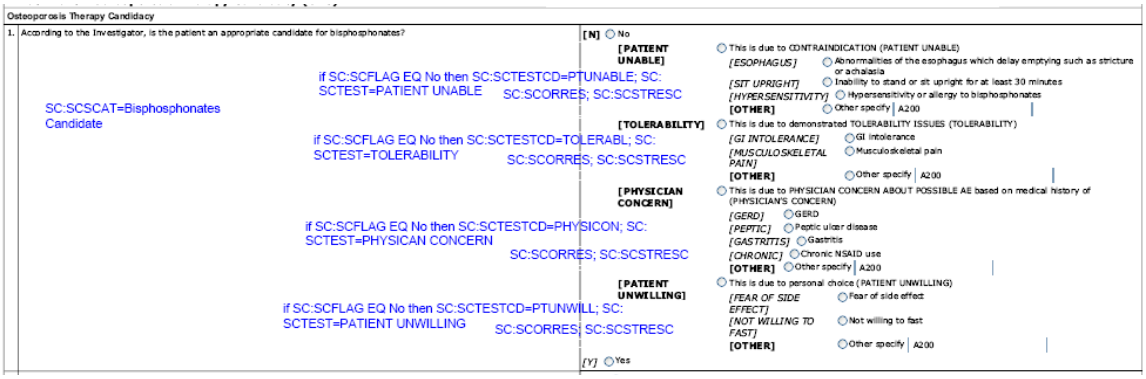
Advantages

- The main advantage of this approach is avoiding the SUPP-- domain and the associated mapping and coding requirements.
- For data management purposes, reducing SUPP-- domain volume is very helpful
- No merging of data records is required.

Disadvantages

- The main challenge of mapping every data point into SDTM core domains without using SUPP-- domain is finding suitable variables that follow the CDISC guidelines yet achieve the desired mapping results.
- For the statisticians and programmers, it becomes very difficult to utilize the data and applying appropriate programming logic is greatly more complicated, due to the parent child relationship of the records in the database.

APPROACH 3 (NO PROXY QUESTION, NO SUPP-- DOMAIN)



Approach 3 Mock dataset with a subset of possible data points

SPID	SCCAT	SCSCAT	SCTEST	SCTESTCD	SCORRES	SCSTRESC
1	OSTEOPOROSIS THERAPY	Bisphosphonates Candidate	PATIENT UNABLE	PTUNABLE	SIT UPRIGHT	SIT UPRIGHT
2	OSTEOPOROSIS THERAPY	Bisphosphonates Candidate	TOLERABILITY	TOLERABL	OTHER : ENTER THE FREE TEXT	OTHER : ENTER THE FREE TEXT
3	OSTEOPOROSIS THERAPY	Bisphosphonates Candidate	PHYSICIAN CONCERN	PHYSICON	GERD	GERD
4	OSTEOPOROSIS THERAPY	Bisphosphonates Candidate	PATIENT UNWILLING	PTUNWILL	FEAR OF SIDE EFFECT	FEAR OF SIDE EFFECT

Mapping approach

The third approach involves capturing a subset of identified data points (not every data point) on the e-CRF in a single record for each patient for each visit without using SUPP-- domain. This approach avoids storing the result of the proxy question "According to the Investigator, is the patient an appropriate candidate for bisphosphonates?" and only captures the data points when

the answer is 'No', using test codes (PTUNABLE, TOLERABL, PHYSICON, PTUNWILL) for all possible responses.

Advantages

- The main advantage of this approach is that the data is captured in a single record for each patient for each visit without using the SUPP-- domain.
- The volume of data is kept minimal since there is only one record per subject for this e-CRF and no records are written to the SUPP-- domain.
- For data management, writing data quality checking reports and edit checks are simpler.

Disadvantages

- The main challenge is identifying various future implications that may later arise as a result of not mapping all collected data points.
- For the statisticians and programmers, it is very difficult to identify the total N of evaluated subjects for the avoided proxy question.
- Data management would need programming assistance to create a query to confirm (against the collected data) that every subject selected either a 'Yes' or 'No' response for the proxy question, since we can not confirm this utilizing the SDTM mapped data

CONCLUSION

While minimizing the use of nested e-CRF data may be the best approach, this is not always possible. Each of the approaches illustrated for mapping nested e-CRF data has both advantages and disadvantages and one approach can not be selected as a standard approach. By understanding the inherent challenges and complexities in each approach, the most appropriate data structure for a particular purpose can be identified and implemented.

REFERENCES:

CDISC Inc., 15907 Two Rivers Cove, Austin, Texas 78717
© 2008 Clinical Data Interchange Standards Consortium, Inc. All rights reserved

SAS User Guide (link: http://support.sas.com/rnd/scalability/tools/fullstim/index.html)

CONTACT INFORMATION:

Your comments and suggestions are valued and encouraged.

Contact the authors at:
Venkata Sekhar Bhamidipati
Merck & Co., Inc.
Scientific Programming Analyst
Email:venkata_bhamidipati@merck.com
Phone:(267) 305-1632

Kristi L. Garner
Merck & Co., Inc.
Scientific Programming Analyst
Email:kristi_garner@merck.com
Phone:(267) 305-1322