PhUSE 2010

Paper CD02

How Do I Map That? - SDTM Implementation Challenges

Chris Price, Roche Products Ltd, Welwyn Garden City, UK

ABSTRACT

In many cases the mapping of a Case Report Form (CRF) page to the appropriate Study Data Tabulation Model (SDTM) domain and variables is relatively straightforward, especially if the CRF has been designed with SDTM in mind. However, there are situations where the data that is collected within a clinical study does not have an obvious mapping to either a standard or sponsor defined domain. This paper will look at some data collection scenarios encountered on a legacy Phase III Rheumatoid Arthritis study and examine some of potential mapping solutions considered, as well as the reasons behind the decision for the chosen option. It will also describe some of the lessons learnt, which were then applied on a more general level to the SDTM mapping.

DISCLAIMER

All views expressed in the paper are those of the author and not necessarily those of Roche Products Ltd.

INTRODUCTION

In an ideal world all of our CRF would be completely CDASH compliant and all our data would map simply into either a standard or sponsor defined SDTM domain. However, as we all know data collection does not always live up to these utopian ideals. In a fairly routine study we will probably find that about 80% of all the data points will map simply to SDTM. The remaining 20% will need differing levels of consideration before they can be successfully mapped into SDTM. Five such issues that we encountered when mapping a legacy Phase III Rheumatoid Arthritis study in SDTM (using SDTM v1.2 SDTM Implementation Guide v3.1.2) were:

- Is death an event or an outcome?
- Where do I map tender and swollen joint counts and how do I deal with different granularity for joint locations?
- Where should multiple interventions to a single study medication record be mapped?
- How should multiple symptoms of a single adverse be tabulated?
- How should the smoking history be tabulated as opposed to the patient's actual substance use?

In most of these examples the best solution would have been to collect the data in a slightly different way which would have allowed for a more natural tabulation of the data. Unfortunately, this option is not always open to us due multiple factors. For example, the existence of standard company CRF pages mapping to an internal database means that changes can have a large impact and cannot be updated overnight. Therefore, the best solution to some of these points may be to re-design the CRF pages but I will also look to suggest a mapping which is appropriate for the current pages.

BACKGROUND

A decision to file a Rheumatoid Arthritis project using CDISC standards was taken in late 2007 with the filing planned for fourth quarter 2010. However, at this stage, the CRFs had been finalized and the operational databases had designed and created for a majority of the planned studies. Therefore, we were in a legacy mapping scenario with filing 3 years away and no work had yet been started on the analysis. This allowed the project team to take a sequential approach to the required tasks, thus, preventing the need for transformations to the analysis datasets to allow for traceability between these and the data tabulations.

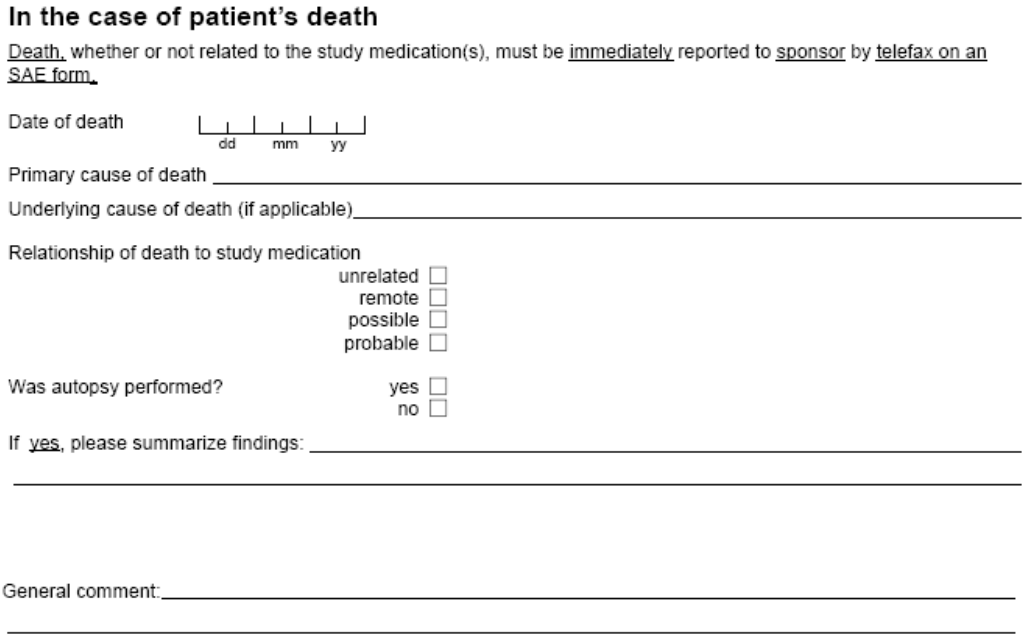
As with most legacy conversions, this created a number of difficulties as certain implemented CRF design concepts did not fit well with the SDTM model. At the time of the initial mapping from the operational database to SDTM, the SDTM v1.2 and SDTM Implementation Guide v3.1.2 were available as a draft version and a number of concepts were not available (for example the Findings About (FA) Domain). The final version of the SDTM and the Implementation Guide became available in early 2009 and greatly helped the mapping exercise, even though it meant updating a number of analysis dataset programs which had been written in the intervening period. Despite this extra effort, it was felt this was a worthwhile update due to the improvement in the tabulation of the raw data.

DEATHS - EVENT OR OUTCOME?

WHY IS THERE A MAPPING QUESTION?

In the previous data model used internally, a separate domain existed in which to store data collected about a patient's death from its own separate CRF page. Further to the cause of death, additional information was collected about any underlying causes and data related to autopsy, if one had been performed. While this was on a separate page, information about death was also collected on the Adverse Events page as a possible outcome and on the disposition page as a possible reason for discontinuation with no explicit links between the three pages. Necessarily these pages were all reconciled by data management to ensure that data was consistent for all patients.

FIGURE 1: BLANK DEATH CRF PAGE



In the case of patient's death

Death, whether or not related to the study medication(s), must be immediately reported to sponsor by telefax on an SAE form.

Date of death ⌊__⌋⌊__⌋ ⌊__⌋⌊__⌋ ⌊__⌋⌊__⌋
dd mm yy

Primary cause of death _____

Underlying cause of death (if applicable) _____

Relationship of death to study medication

unrelated ☐
remote ☐
possible ☐
probable ☐

Was autopsy performed? yes ☐ no ☐

If yes, please summarize findings: _____

_____

General comment: _____

_____

The SDTM model does not consider death to be an event in itself, but rather as an outcome of an adverse event. Consequentially, it does not provide for a separate domain solely reserved for death information. This has no theoretical impact on the data being collected, we can still collect as much additional information about a death as we may have done before. However, the impact here is more psychological, by not considering death as an event in itself with its own CRF page and tabulation domain it is sometimes felt that we are downgrading the importance of the event.

WHAT ARE THE OPTIONS?

This is obviously an example where an update to the CRF design would be the ideal solution. All the information could be collected but an explicit link between this data and the Adverse Event page would allow the data to be mapped into the AE SDTM domain with the cause of death as the AE and the additional information around this mapped to either the FA domain or into the Supplemental Qualifier (investigator comments would obviously would be mapped to the Comments (CO) domain in all scenarios).

With the current CRF design and collection rules we needed to make a decision as to which domain could we map the data. Due to the CRF design we were unable to relate this data back to either an adverse event or disposition event due to the lack of an explicit link between the pages. In this collection model it is undoubtedly an event which gives us the choice of the domains Adverse Events (AE), Disposition (DS), Medical History (MH), Protocol Deviations (DV), Clinical Events (CE) or a sponsor defined domain. We can immediately dismiss the MH and DV domains as inappropriate as neither is a good fit for the data. DS was discounted as in terms of categorizing the data the regular collection of a patient's death could not be described in itself as either a "disposition event" or a "protocol milestone". It could, potentially, be modeled as an "other event" but it was felt that this was not entirely appropriate given the definition of a disposition event.

AE was also dismissed as an option as it would mean collecting the same information twice within the same domain, which could lead to confusion for the users or reviewers of the data. The cause of death could have been mapped to the reported term (AETERM) and the outcome (AEOUT) would have been defaulted to "DEATH" the additional

information would have been mapped within the domain or the supplemental qualifier depending on the field. This theoretically should be an exact match (on the overlapping fields) to the adverse event page data and an additional qualifier would be needed to indentify the records. We could have used category (AECAT) for this, but we took the decision that it was best to leave this open for different ways of categorizing the data.

WHAT WAS THE DECISION?

The final decision taken was to use the Clinical Events (CE) domain which was defined in the draft version of the implementation guide and refined for the final version. The purpose of the CE domain is defined as "to capture clinical events of interest that would not be classified as adverse events." In some clinical studies, for example, where survival is being evaluated this would be a very natural mapping. To ensure a consistent approach to the collection of this data, it was decided to also map deaths in studies where it was not related to an endpoint to the CE domain.

FIGURE 2: ANNOTATED DEATH CRF PAGE

In the case of patient's death CE.CECAT = 'DEATH'

Death, whether or not related to the study medication(s), must be immediately reported to sponsor by telefax on an SAE form.

Date of death# CE.CESTDTC
dd mm yy

Primary cause of death CE.CETERM

Underlying cause of death (if applicable) SUPPCE.QVAL
SUPPCE.QNAM = 'CEDDUTXT'
SUPPCE.QLABEL = 'UNDERLYING CAUSE OF DEATH'

Relationship of death to study medication
unrelated ☐
remote ☐
possible ☐
probable ☐
CE.CEREL

Was autopsy performed?
yes ☐
no ☐
SUPPCE.QVAL
SUPPCE.QNAM = 'CEDDAUYN'
SUPPCE.QLABEL = 'AUTOPSY PERFORMED'

If yes, please summarize findings: _____

SUPPCE.QVAL
SUPPCE.QNAM = 'CEDDAUR1' - 'CEDDAURn'
SUPPCE.QLABEL = 'SUMMARY OF AUTOPSY FINDINGS 1' - 'SUMMARY OF AUTOPSY FINDINGS n'

General comment: _____

CO.COVAL [CO.COVALn - CO.COVALn]
CO.RDOMAIN = 'CE'
CO.IDVAR = 'CESEQ'

TENDER AND SWOLLEN JOINT COUNTS

WHY IS THERE A MAPPING QUESTION?

Tender and Swollen Joint Counts were a key part of the primary endpoint for this Rheumatoid Arthritis study. For this data the first question we need to consider is, which domain we should map this data into, as it is not a natural fit into any of the currently defined SDTM domains. Secondly we need to consider the actual mapping itself. The data collected has two levels of granularity for the location of the tenderness and swelling of the joints. Firstly, the side of the body is collected and secondly, the name of the joint is also collected separately.

FIGURE 3: BLANK JOINT COUNT CRF PAGE

RIGHT				JOINT	LEFT			
Tenderness present?	Swelling present?	Not done	Reason if not done		Tenderness present?	Swelling present?	Not done	Reason if not done
☐	☐	☐	____	Shoulder	☐	☐	☐	____
☐	☐	☐	____	Elbow	☐	☐	☐	____
☐	☐	☐	____	Wrist	☐	☐	☐	____
☐	☐	☐	____	MCP1	☐	☐	☐	____
☐	☐	☐	____	MCP2	☐	☐	☐	____
☐	☐	☐	____	MCP3	☐	☐	☐	____
☐	☐	☐	____	MCP4	☐	☐	☐	____
☐	☐	☐	____	MCP5	☐	☐	☐	____
☐	☐	☐	____	IP Thumb	☐	☐	☐	____
☐	☐	☐	____	PIP2	☐	☐	☐	____
☐	☐	☐	____	PIP3	☐	☐	☐	____
☐	☐	☐	____	PIP4	☐	☐	☐	____
☐	☐	☐	____	PIP5	☐	☐	☐	____
☐	☐	☐	____	Knee	☐	☐	☐	____

Assessor's initials ____ Tender joint count ____ Swollen joint count ____

WHAT ARE THE OPTIONS?

With regard to the target domain for tender and swollen joint count data one option would be to include the data in a already defined SDTM domain, which may not be a perfect fit but could be considered 'good enough'. The one domain which could be considered here would be the Physical Examination (PE) domain. The definition of this domain provided is "Data that captures findings about physical exams. This could be information about which body systems were examined and specific abnormalities." While this may sound appropriate for an evaluation of joints, the definition of the result variable would lead to mapping issues, as if no abnormal findings are found then the result should be recorded as normal. While this would be feasible if the body system examined was set to "SKELETAL" and the name of the joint and side was defined in other variables, we would need to work with controlled terminology of "NORMAL", "SWOLLEN", "TENDER" and "SWOLLEN AND TENDER". While this mapping fulfils all the conditions it would require a relatively complicated mapping programming to derive the results based on the way the data is collected. It would also require subsequent additional effort on the analysis side in order to transform the data to make it suitable for analysis.

For the second issue, as we have discussed, the exact location is collected as two separate fields on the CRF, the side of the body and the name of the joint. One mapping solution to be considered is to used the location variable (--LOC) for the name of the joint and then used the subcategory variable (--SCAT) for the side of the body. This approach would uniquely identify the location in combination with the evaluation name (--TEST/--TESTCD). However, this is not a completely appropriate use of the categorization of the category variables as these should be used to define a category of topic variables, so the topic variable (--TESTCD) should ideally uniquely appear in only one category which would not be the case.

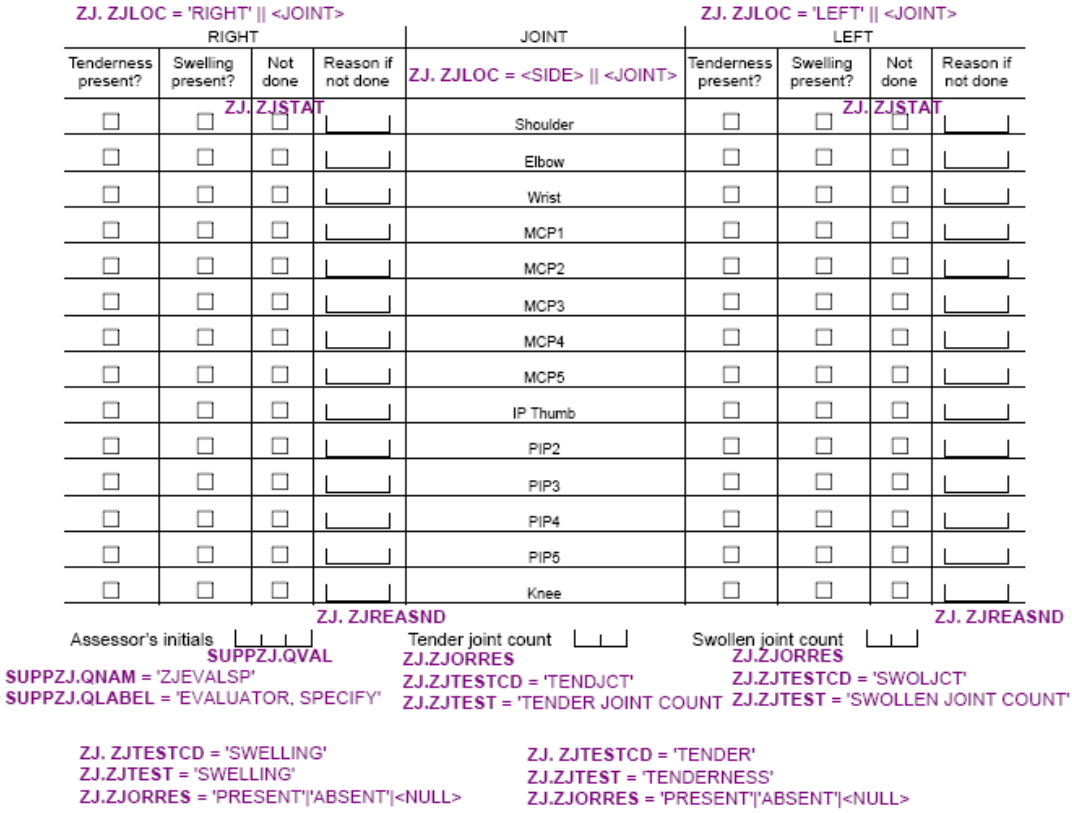
A second option also considered was to create a new variable, which would be stored as a supplemental qualifier, which would provide us with the level of granularity required and could be related to the correct record using the sequence number (--SEQ). This is, again, not an ideal solution as we are adding lots of additional data to a supplemental qualifier as a record qualifier variable which given the FDA's reported lack of enthusiasm for supplemental qualifier datasets does not seem to be a sensible approach.

WHAT WAS THE DECISION?

It was decided to map the tender and swollen joint count data into its own sponsor defined findings domain (in this case it was named ZJ). This decision was taken based on both how the data was collected and how it was going to be used as part of the analysis. It is obviously a findings domain as it is capturing observations resulting from planned evaluations to address a specific test, specifically whether a joint is either tender or swollen. We can then defined the name of the measurement, test or examination as "TENDER" or "SWELLING" and the result will be either "PRESENT", "ABSENT" or in the case where the evaluation was not performed null. In this final instance if Not Done was selected, the status (ZJSTAT) and the reason not done (ZJREASND) variables would both be populated.

With regard to the location of the evaluation, the solution implemented was to concatenate the side of the body with the joint location in a single variable (ZJLOC). This enables us to uniquely define each joint without the need of a supplemental qualifier. As the side of the body or joint name are not of primary importance for any analysis this does not have an adverse impact downstream, but, if this information is ever required it should be a simple task to programmatically split this variable in an analysis dataset.

FIGURE 4: ANNOTATED JOINT COUNT CRF PAGE

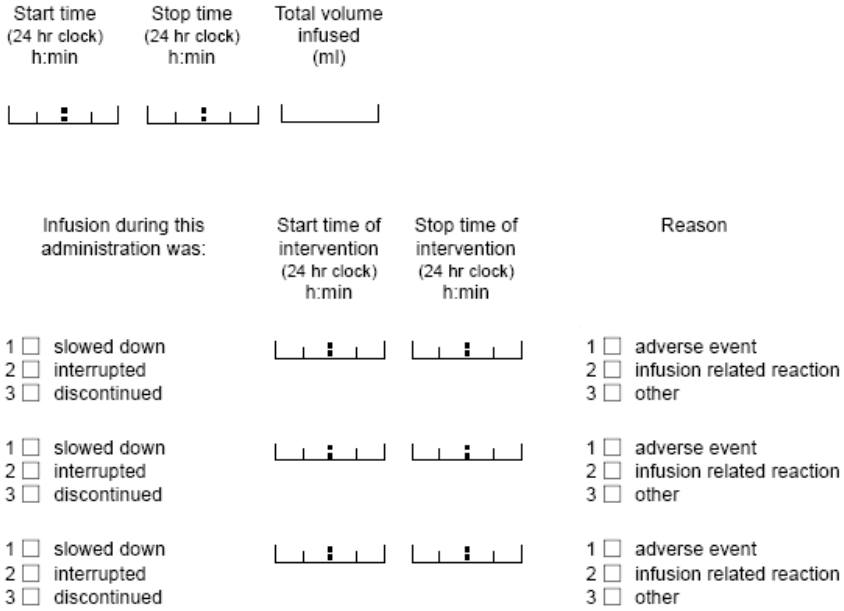


INTERVENTIONS TO INFUSIONS

WHY IS THERE A MAPPING QUESTION?

For the infusions of study drug, at first glance this would be expected to be a relatively trivial exercise and for the data relating specifically to the infusion it does fit very easily in the Exposure (EX) domain. The start and stop times map to their respective variables (EXSTDTC and EXENDTC), the total volume infused maps to the vehicle amount (EXVAMT) and the dosage information maps to the treatment (EXTRT), dosage (EXDOSE) and dosage unit (EXDOSU) at unbinding. Where a question arises is the recording of the interruptions to the infusion. These are collected as multiple fields relating to each intervention, multiple interventions can then be explicitly linked to a single infusion.

FIGURE 5: BLANK INTERVENTION TO INFUSIONS CRF PAGE



WHAT ARE THE OPTIONS?

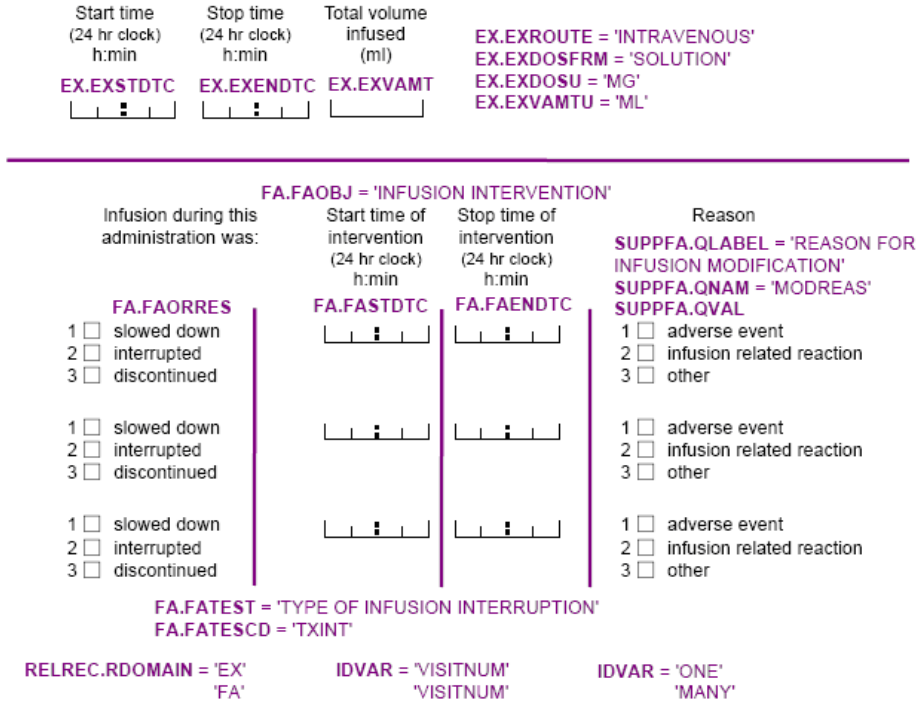
At the time of the initial mapping of this page only the draft version of the SDTM Implementation Guide v3.1.2 was available and there were no available solutions for linking the multiple fields relating to each intervention and then relating multiple interventions to a single infusion. The initial solution proposed to this problem was to create the supplemental qualifier so that the QNAM variable was restricted to 6 characters and the remaining two characters were then available to using as a numeric grouping identifier for the intervention number. This meant that each intervention led to the creation of four supplemental qualifier observations - type of intervention, start and stop times of intervention and reason for intervention. This is an inelegant solution both from a mapping perspective and with a view to using the data in any analyses as it is combining what would more naturally be two variables into one, which needs to be separated before data can be transposed and added to the original record.

An improved solution to this based on the final version of the SDTM Implementation Guide v3.1.2 was based on the original solution, but made use of the new Findings About (FA) Domain. This solution mapped the object of the finding (FAOBJ) to "INFUSION INTERVENTION" with four test codes (FATESTCD) and descriptions (FATEST). Each finding for an individual interventions were then grouped together using the group identifier (FAGRPID) variable. This followed very much the same approach as the Supplemental Qualifier solution. While it no longer has the drawback of the intervention identifier being attached as part of the filed identifier, it still stores related data vertically which would more naturally fit a horizontal structure.

WHAT WAS THE DECISION?

The approach which was subsequently followed was similar to the second rejected options based on using the FA Domain. The object of the finding (FAOBJ) is again the "INFUSION INTERVENTION" the difference with this solution is that there is only one test code (FATESTCD) and description (FATEST) for each intervention. The reason for the intervention is mapped to the result (FAORRES) variable, the start and stop times are mapped to the start and end date time variables (FASTDTC and FAENDTC) in conjunction with the date of the infusion. The reason for modification is then mapped to a supplemental qualifier for the FA domain. The final part of the mapping is to create an observation in the relationship domain (RELREC) to relate the observation in the EX domain to the observations in the FA domain using the visit number (VISIT) variable as the indentifying (IDVAR) variable.

FIGURE 6: ANNOTATED INTERVENTION TO INFUSIONS CRF PAGE

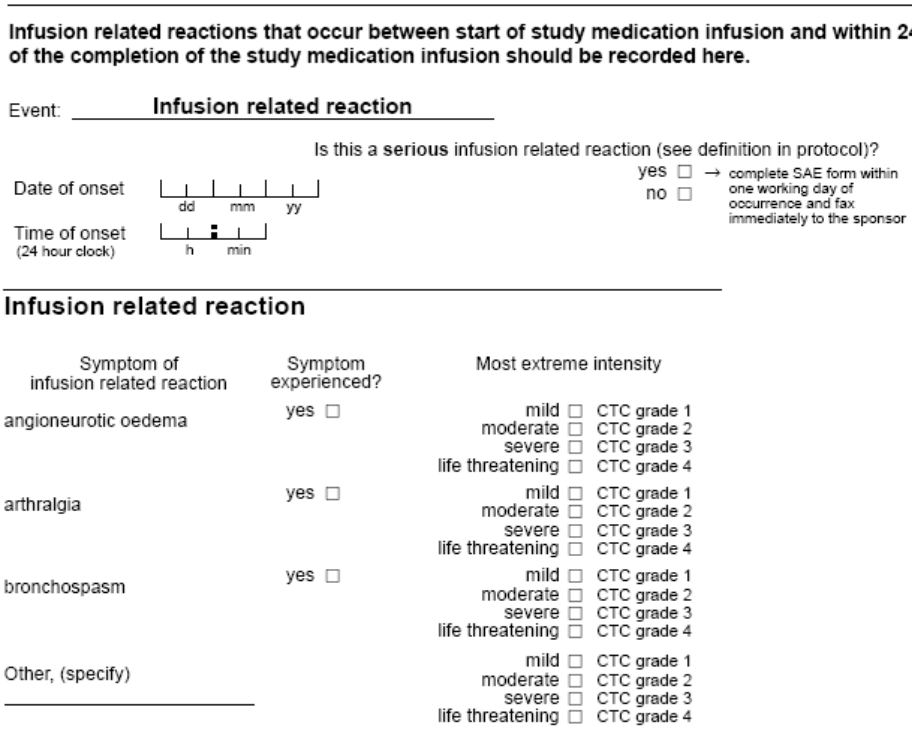


SYMPTOMS OF INFUSION RELATED REACTIONS

WHY IS THERE A MAPPING QUESTION?

In many Rheumatoid Arthritis studies the study drug is administered by infusion, consequently a choice has been made to group a number of adverse events into a single event - an Infusion Related Reaction. These are collected from a pre-specified list as the individual symptoms associated with the infusion related reaction. They are then reported as a single event within the analysis. The mapping of the actual adverse event of Infusion Related Reaction is trivial as it maps into the Adverse Event (AE) Domain. The mapping question that needs to be answered is how to map the Symptoms of the Infusion Related Reaction especially with the concept of pre-specified terms.

FIGURE 7: BLANK INFUSION RELATED REACTION SYMPTOMS CRF PAGE



WHAT ARE THE OPTIONS?

As stated previously the mapping of the adverse event term of "Infusion Relation Reaction", the start date, the start time and whether the event was consider serious or not is trivial. All fields map to the AE domain; the text "Infusion Related Reaction" maps to the reported term for the adverse event (AETERM), the start date and time are combined to map to the start date/time of the adverse event (AESTDTC) and the serious flag maps to the serious event flag (AESER). The other SDTM variable that need to be populated is the pre-specified adverse event flag (AEPRESP) which indicates that this was not a spontaneously reported event.

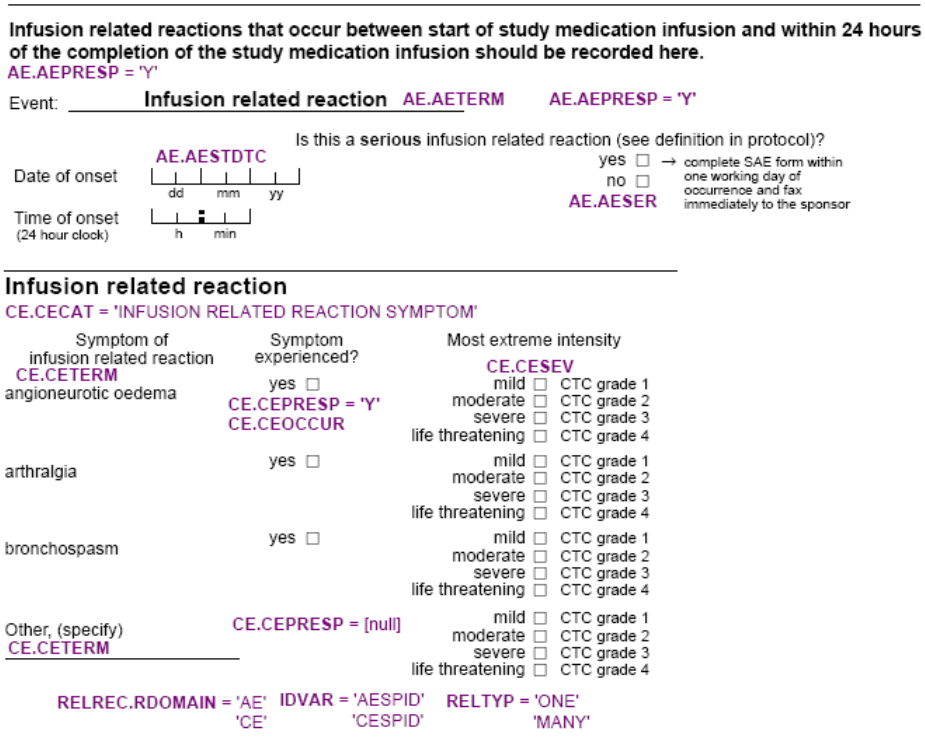
Concentrating of the second part of this adverse event CRF page where the symptoms of the Infusion Related Reaction are captured, we first considered implementing a mapping where each symptom was an individual adverse event (AETERM). In this mapping only symptoms that occurred were mapped to the SDTM domain following the recommendation in SDTM Implementation Guide v3.1.2. Consequently, the pre-specified flag (AEPRESP) was populated for all symptom records and the severity of the symptom was mapped to the Severity/Intensity (AESEV) variable in SDTM. As no date or time information was collected for an individual symptom these variables would not be populated. In order to maintain the link back to the master AE record of "Infusion Related Reaction" the Group ID (AEGRPID) variable was populated with a identifier to link all symptoms associated with a single Infusion Related Reaction back to that master record.

Although this is a reasonable and SDTM compliant solution it is possibly not the most appropriate, especially considering the usage of this data for analysis. It was planned for any adverse event analysis to count and record only the adverse event of infusion related reaction. The symptoms would then be analysed separately and would never be combined with the master record of infusion related reaction. Considering this approach to analysis, it could be considered misleading to include this data in the AE domain, as it could give the impression that adverse events were being under-reported.

WHAT WAS THE DECISION?

The solution that was implemented was to map the first part of the CRF page as described above and to map the second part of the CRF page to the Clinical Events (CE) domain. This solution fits the intent of the CE domain in SDTM Implementation Guide v3.1.2 perfectly as it states "The intent of the domain model is to capture clinical events of interest which would not be classified as adverse events." As CE is an Events domain, it has the same inherent structure as the AE domain, so we can follow the same mappings as were previously considered when mapping the AE domain. The only difference is the additional mapping of the occurrence (CEOCCUR) variable to allow for the collection of data related to the non-occurrence of a symptom. The final part of the mapping was to define an explicit relationship between the record in the AE domain and the records in the CE domain using the relationship (RELREC) domain.

FIGURE 8: ANNOTATED INFUSION RELATED REACTION SYMPTOMS CRF PAGE

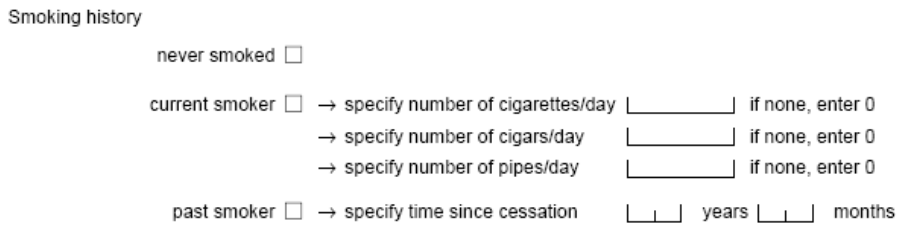


SMOKING HISTORY

WHY IS THERE A MAPPING QUESTION?

Data regarding smoking was collected once during the study at screening as part of the baseline demographics. The first part of the information collected related to the actual smoking status of the patient, i.e. was the patient a non-smoker, current smoker or past smoker. For non-smokers no additional data was collected, if the patient was a current smoker, additional questions were asked regarding the number of cigarettes, cigars or pipes smoked per day, and if the patient was a past smoker then the time since cessation was additionally collected. Historically, the data was mapped to the demographic domain in accordance with the usage of the data in analysis. Within SDTM this would not be considered an appropriate solution considering the availability of other domains and the restricted content of the Demographics (DM) domain.

FIGURE 9: BLANK SMOKING HISTORY CRF PAGE



WHAT ARE THE OPTIONS?

As previously observed the historical solution of adding this data to the demographics, in this case as supplemental qualifier observations in SUPPDM, is not an appropriate solution as this is data not strictly related to a patient's demographics.

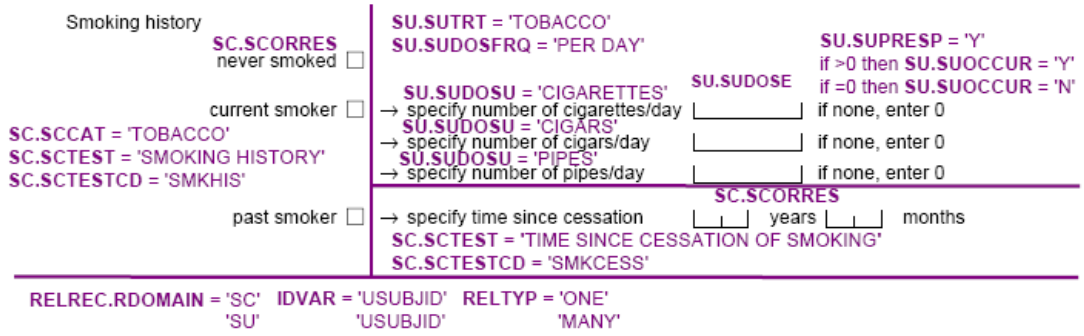
A more plausible solution is to map all the data on the page to the Substance Use (SU) domain. A possible compliant mapping would be to create three records per patient, on for cigarettes, cigars and pipes, this would be captured in the dosage unit (SUDOSU). The reported name of substance (SUTRT) would be "TOBACCO" and for non-smokers and past smokers the dosage (SUDOSE) would 0 and for current smokers would be the number populated in the CRF. The actual smoking status variable could be captured in the category variable (SUCAT). The time to cessation would then be captured as Supplemental Qualifier observations for past smokers.

This suggested mapping leads to a number of concerns. Firstly this is not a completely appropriate use of the category variable as this should be used to define a category of topic variables, the topic variable (--TRT) should ideally uniquely appear in only one category which would not be the case. Secondly we would be artificially creating records where in some case it is not possible to collect this data (i.e. non-smokers). There would also be a large amount of data redundancy, as the information relating to the smoking status is being stored in triplicate where there is only one question which hinders the usability and understanding of the data. As a consequence this mapping was not considered appropriate.

WHAT WAS THE DECISION?

The mapping that was chosen was to split the three questions (Smoking Status, Number of Cigarettes/Cigars/Pipes per day and time since cessation) into separate domains. The smoking status was mapped to the Subject Characteristic (SC) domain as this is data which was collected once per patient and can be considered as an extension to the demographic data, which fits perfectly the purpose of the SC domain. The Subject Characteristic (SCTEST) would be assigned as "SMOKING HISTORY" and the result (SCORRES) would be the status. Additionally mapped to the SC domain for past smokers as a separate Subject Characteristic (SCTEST) would be the time since cessation, the result (SCORRES) in this case would be recorded in an ISO 8601 format. These two data points are then linked using the category variable (SCCAT) which would be assigned to "TOBACCO". The only data that would be mapped to the SU domain would be the actual substance use for current smokers which would be mapped as described above but the category (SUCAT) variable would not be populated. The final part of the mapping was to define an explicit relationship between the smoking status record in the SC domain and the records in the SU domain using the relationship (RELREC) domain.

FIGURE 10: ANNOTATED SMOKING HISTORY CRF PAGE



CONCLUSION

The mapping of legacy studies throws up a number of challenges when it come to mapping collected data to SDTM. In many case there is no definite single "correct" approach to a mapping and several solutions could be considered valid. In these cases we need to take into account a number of different factors when deciding which mapping approach to take. We need to consider both how the data is collected and linked together on a CRF and how the data will be analysed. A compliant mapping is of no use if it cannot be analysed in a simple and understandable way.

Moving forward it should be considered essential to adjust currently existing standard collection models (or develop new collection models) so that they can be easily mapped into SDTM and subsequently be used for the required analyses in a seamless process. As such it is essential that those individuals who will map the data into SDTM and those who use the data for analysis are involved from the outset in the design of the data collection model. The collection of data is meaningless if it cannot be analysed appropriately due to a faulty collection model.

REFERENCES

Study Data Tabulation Model v1.2
Study Data Tabulation Model Implementation Guide: Human Clinical Trials v3.1.2
(www.cdisc.org/sdtm)

ACKNOWLEDGMENTS

I would like to thank:

- Karen Rowe, Olivier Leconte, Patricia Gerend (all Roche) for reviewing this paper
- John Franchino, Frederik Malfait and Jonathan Chainey (all Roche) for both providing input, advice and for their review of this paper
- Peter Van Reusel (Business and Decision Life Sciences) for his advice regarding the Interventions to Infusions example

CONTACT INFORMATION

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

Chris Price
Roche Products Ltd.,
6 Falcon Way
Welwyn Garden City
AL8 7SN
Work Phone: +44 (0)1707 365801
Fax: + 44 (0)1707 383145
Email: chris.price.cp1@roche.com