PhUSE 2018

Paper DS04

The interpretation and development of SDTM standards for non-submission purposes

Paul Vervuren, Nutricia Research, Utrecht, the Netherlands
Lieke Gijsbers, OCS Life Sciences, 's-Hertogenbosch, the Netherlands

ABSTRACT

Recognizing the value of its data, Nutricia Research performed an SDTM conversion project for twenty-one infant nutrition studies. Since 2015, Nutricia Research and OCS Life Sciences have been working together on this project, building expertise in the SDTM conversion of early life nutrition data, and jointlydefining some specific SDTM domains. This paper presents concrete examples and implementation solutions from this project, and reflects on possibilities and limitations of using SDTM as a standard for non-submission purposes.

INTRODUCTION

While clinical studies serve to proof the safety and efficacy of medical products, its data, because of the high intrinsic quality, is of value beyond the lifetime of an individual study or the collection of studies. Applying standards and implementing them in the right way, is key to establish and preserve that value.

Anyone who has worked with the SDTM standard, will probably acknowledge that the intellectual challenge in its application is not so much to be conformant (i.e. 'technically compliant'), but that it lies in making the right implementation decisions. 'Right', meaning first of all the ensurance that the data when mapped to SDTM keep their original meaning. It also means making proper use of the flexibility of the standard to organize the data in a way that is logical, userfriendly and in agreement with the specific needs of the research area. The latter need explains the rise of Therapeutic Area standards. This is not to say that ensuring conformance to the standard is less relevant or that it is without issues or challenges; clearly, resolving and documenting SDTM non-conformities can be quite a task.

Nutricia Research has been working on an SDTM conversion project involving twenty-one infant nutrition studies. Compared to a submission trajectory this project was different because:
- Substantial tailorization was needed in overall design, data collection modules and CRF systems. Arguably this was significantly greater than you would find in a typical submission.
- Studies and domains were added along the way. As the integrated dataset grew, design solutions and definitions needed to be updated.

This project has been a collaborative effort between Nutricia Research (as sponsor) and OCS consulting (as service provider). Next to the right technical expertise and approach, a succesful project approach and collaboration model is key to project success. In this paper we will highlight some typical SDTM domains and the mapping issues encountered, and present our key learnings and recommendations. We hope and expect that broader lessons can be drawn from our work, helping data warehouse projects (e.g. SDTM implementations across compounds or companies) as they are likely to face similar issues.

We will start this paper with an overview of the project and we will reflect on differences with an SDTM implementation for a submission. Next, we will discuss the evolution of the project, addressing in particular adaptions in the approach and collaboration. Then we'll go into the quality control and validation approach. Subsequently, we will dive into several typical domains and present solutions for the SDTM mapping challenges we faced.

SCOPE, REQUIREMENTS AND APPROACH OF OUR PROJECT

This conversion project involved twenty Nutricia Research infant studies. The overall goal of the project was to standardize and pool the data of these studies to enable cross-study analyses for exploratory, scientific research purposes in the area of infant nutrition. Selection of the studies was based on business relevance, design characteristics (e.g. randomized controlled, presence of a breastfed reference arm), and size of the study (number of subjects). Infant formula was the treatment intervention in all studies but there was variety across studies in the formula(s) used.

Our way of dealing with standards differed from regulatory (e.g. FDA) submissions in the sense there was no external requirement to adhere to industry guidance. However, it was thought that conformance to industry

standards would bring two main benefits: 1. tools and software that work with the standards will generally only work if (submission) standards are followed; being just inspired by a standard or loosely following a standard is generally insufficient and may lead to failure to load data or in many conformance errors; 2. Exchange of the data at a possible later stage, e.g. for submission or other purposes like internal or external sharing, would be greatly facilitated. The latter advocates the application of documentation standards like define.xml in addition to the data content standards.

There are many detailed aspects to consider when preparing a submission to a regulatory agency like FDA. These are related to dataset structure and content, (eCTD) folder structure, documentation, and handling of data validation issues (e.g. see Tinazzi & Marchand, 2017). It is outside the scope of this paper to make a full and detailed comparison between the standards applied in our project with those of a regulatory (IND, NDA) submission. But let's briefly look at how typical submission standards were applied in our project.

Component	Regulatory submission (FDA)[1,2]	Nutricia project
SDTM datasets	Required for all submitted clinical data; must be SDTM conformant and able to fix/explain all validation issues to ensure acceptance by agency	Desired; SDTM conformant to facilitate use of tools and associated standards.
ADaM datasets	"Should be used to create and to support the results in clinical study reports, Integrated Summaries of Safety (ISS), and Integrated Summaries of Efficacy (ISE), as well as other analyses required for a thorough regulatory review."	Not part of the conversion project, but the intended standard for analyses based on the integrated data.
Trial Design tables	"All TD datasets should be included, as appropriate for the specific clinical trial, in SDTM submissions as a way to describe the planned conduct of a clinical trial. Specifically, the Trial Summary (TS) dataset will be used to determine the time of study start"	Nice-to-have; not included
Define.xml	"The data definition file describes the metadata of the submitted electronic datasets, and is considered arguably the most important part of the electronic dataset submission for regulatory review."	Desired
Annotated CRF	Required, proving a link between the data collection fields and the SDTM variables and terms.	Desired
Reviewer guides (RG)	Expected with SDTM and ADaM datasets	Nice-to-have, in particular to document data conformance issues and their handling
Study Data Standardization Plan (SDPS)	Expected; mechanism to assists FDA in identifying potential data standardization issues early in the development program	Not used and not considered relevant at this point
Legacy Data Conversion Plan and Report	Expected as part of the RG when legacy data was converted to SDTM and ADaM, in order to provide an explanation of the process, record traceability, issues, adjudications.	Available in mapping sheets, design documentation and meeting minutes.
SAS transport files	Required	Not required (and not used)

With the exception of transport files and SDPS, the various submission standards were either used or considered useful in the context of our project. How exactly standards like Reviewer's Guides for SDTM and ADaM, and other documentation (e.g. legacy data conversion plan and report) will be used in the context of the Nutricia will require further thought.

EVOLUTION OF PROJECT APPROACH AND COLLABORATION

The arrangement between Nutricia Research and OCS Life Sciences for this project was a service agreement where OCS would conduct the SDTM legacy conversion and generate the following deliverables:

- Designs of the SDTM datasets
- Mapping sheets
- Annotated CRFs
- SAS programs performing the conversion
- SDTM data sets
- Validation documentation
- Data exception report

[1] See Study Standards resources on the FDA website: www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm, and, specifically, 'Providing Regulatory Submissions In Electronic Format — Standardized Study Data, Guidance for Industry' www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM292334.pdf, and the related 'Study Data Technical conformance Guide'. www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf.

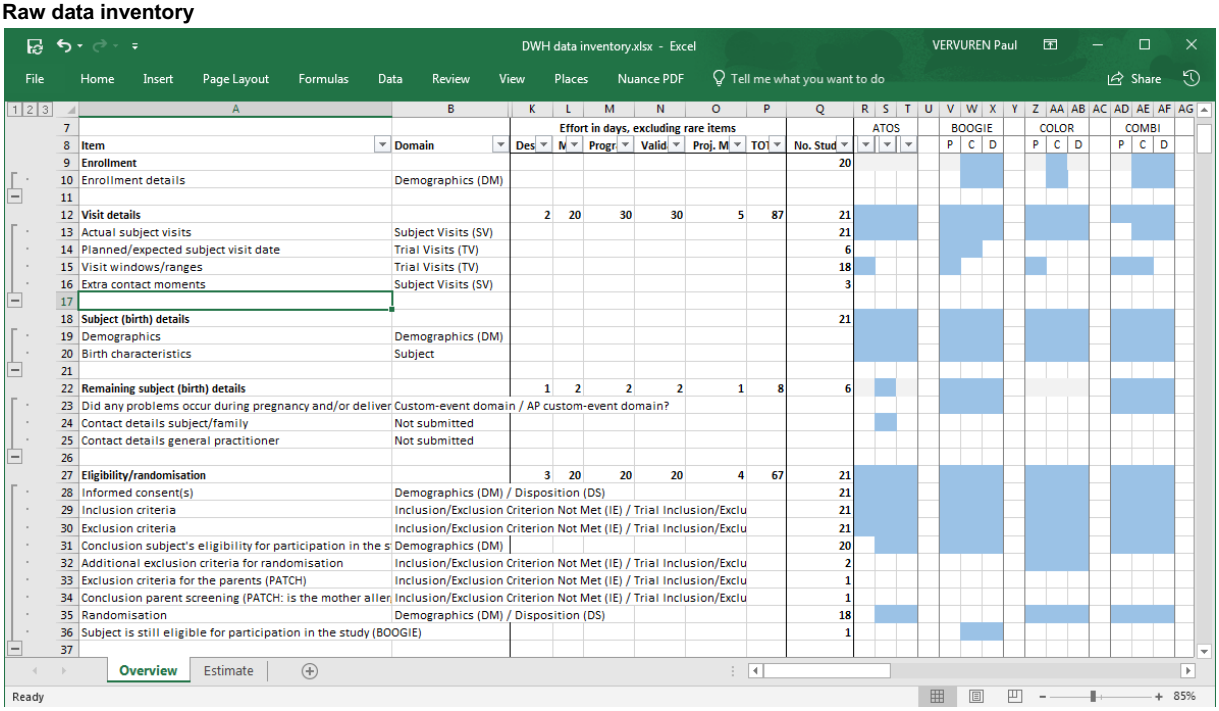
[2] See PhUSE CSS work on regulatory deliverables, specifically on reviewer guides at https://www.phuse.eu/css-deliverables

The project encompassed twenty-one studies. In the first phase of the project the following modules were selected: demographics and subject characteristics, vital signs, adverse events, gastrointestinal tolerance, concomitant medication.

The main assumption underlying this arrangement was that the OCS project team would be able to perform the design and mapping independently and that questions and issues could be mostly handled via email. However, at an early state we discovered that there were implementation choices, data issues and exceptions whose resolution were a matter of preference. In most cases, OCS could define the options but final decisions needed to be taken by Nutricia. Having had several 'implementation deep-dives' triggered by mapping questions, we decided to organize these meetings between the OCS project team and Nutricia domain experts from Data Management and Statistical Programming on a regular basis. Thus we changed the approach from a delivery model to a delivery-and-collaboration model. An additional advantage of this adjusted approach, and the fact that meetings happened at Nutricia offices, was that specialists for specific areas (e.g. for microbiology data) could be quickly and easily involved in the process. In support of this collaboration we established a secure shared drive for the exchange of (draft) deliverables. This greatly helped te combined teams to efficiently exchange information and work in progress. This exchange may seem a trivial matter, and it's not so much the establishent of the technical provision, but the open exchange of draft work and proposals, that made the difference.

In a second phase, the project was extended to include new domains. The idea in principle was to convert all data to SDTM. At the same time there was a need to be cost-efficient and make decisions based on the potential value of individual variables. Because of the great variability in general design, CRF design, dataset structure and naming conventions across studies, it was difficult to make a quick assessment of the number of possible target domains and data that studies had in common at a more granular level (variables, lab tests, etc.). To enable such an assessment, a raw data inventory was made and the results were put in a spreadsheet. The spreadsheet held an overview of all studies, source datasets/variables and their preliminary mapping. This was established based on the review of the CRFs, as well as the protocol (in particular for non-CRF data), and was verified against the datasets (which also acted as a check on the completeness of the transfer of raw data from Nutricia to OCS). This resulted in a full picture of the available raw data, which allowed us to make informed decisions on the scope and priorities, and it made estimations of work breakdown and required effort more transparent. Later on, this also became a useful search tool for users of the data.

Raw data inventory



QUALITY CONTROL

Quality control procedures were put in place to verify the following requirements on the SDTM datasets:

- Traceability: the SDTM datasets and their component parts (variables, value level metadata) can be traced back to the raw source data and mapping decisions
- Reproducability: the SDTM datasets can be reproduced on Nutricia systems using the conversions programs developed by OCS

- Conformance: the SDTM datasets are SDTM conformant, i.e. they meet the specific design specifications as well as the general SDTM model specifications
- Completeness: the SDTM datasets are complete, i.e. all raw data items that were designed to be mapped are indeed present in the SDTM datasets
- Integrity: data integrity is preserved, i.e. data points are not inadvertently affected (e.g. by truncation), no loss of records.

To verify these requirements a risk-based approach was taken based on the probability (P), impact (I) and non-detectability (ND) of issues/errors, each with a five-level scale (see appendix for definitions). The risk score is defined as P x I x ND, and can range from 1 to 125.

Risk score ranges and testing requirements based on Probability, Impact and Non-detectability of issues and errors

Risk Score Range	Risk Score Category	Testing requirement
>65 – 125	Non-tolerable	Mitigation actions required to bring the risk below 65.
>40-65	High	Extensive testing
>20-40	Medium	Single testing
1 - 20	Low	No testing

Risk scores for issues/errors per requirement of project deliverables

Requirement	P	I	ND	Risk Score	Comment
Traceability	2	2	4	16	Process is traceable by design (mapping sheet, program log, etc.); no direct impact on analysis results. Any issues will only pop-up later (thus quite undetectable). No testing is performed.
Reproducability	4	3	2	24	When running programs on a different system issues can easily occur; impact intermediate (also considering that analyses will not be on a critical path), but issues are likely to emerge immediately when running the program. Basic testing performed is to run conversion programs on Nutricia systems, and programs are checked to ensure they meet demands of 'Good Programming Practice'.
Conformance	4	3	4	48	There are many conformance aspects, which increase the likelihood of issues; impact not as high as incompleteness or integrity issues (also because not intended for submission). Generally not directly detectable. Extensive testing required, via automated (metadata) checks and independent review of design versus datasets.
Completeness	2	4	5	40	Checks are built into the process to verify that all datasets and variables are mapped. Impact of incompleteness in the SDTMs could be substantial, and may go unnoticed. Checking required of input versus output in particular on record loss (e.g. when raw data is split across multiple domains .
Integrity	3	4	5	60	Integrity issues like truncation of text strings, formatting/rounding problems; but also mistakes in derivations. Impact can be high and can go unnoticed in certain cases. Extensive checking required.

Based on this risk assessment, extensive validation testing was performed on the SDTM datasets to ensure SDTM conformance, as well as the completeness and integrity of the data. Basic testing was considered adequate for reproducibility and no testing was done for traceability because of the associated low risk.

Following this overall assessment, a validation procedure was designed to verify conformity, completeness and integrity. This procedure was based on Nutricia's instructions and templates for QC of Statistical Programming activities, but tailored to the specific needs and setting of this project. Since there were specifics and complexities to take into account per study (like differences in data structure, CRF design), QC was performed and documented by study. The overall flow is as follows:

1. Developer sets up a QC plan per study in a spreadsheet. This plan details the specific areas (variables, derivations) that require verification. The plan has one record per program (usually one or two domain datasets are produced per program). In case of complex mappings that involved a special subject matter expert, this expert is planned to perform independent review of the data and the mapping specifications.
2. The QC plan is reviewed by the project lead. Updates are made when needed, and, after approval, execution can start.
3. A QC tracking sheet based on the QC plan is kept to monitor and record progress.
4. Independent tester checks the specific areas in the QC plan as well as the default checks on reproducability, conformance and integrity (as listed in the QC report template). When applicable the subject matter expert performs the 'expert check'.
5. Any rework is flagged, performed and verified, and documented by the tester in the QC report.
6. After completion, the project lead checks compliance with the QC plan and signs the QC Summary Form, which documents the compliance check.

Checks on conformance, completeness and integrity are integrated in this procedure in the following manner:

- Conformance: Pinnacle21 checks are performed by the program developer; exception reports are checked during validation by the independent tester. When applicable an expert review is performed.
- Completeness: the independent tester verifies that 'all required or expected variables are present', and that the 'number of variables in output is as expected' (i.e. matching the input 1:1, or logically increased or reduced due to SDTM-based transformations). (Automated processes and checks in the development phase ensure that all variables are in the mapping sheet.)
- Integrity: independent tester is performing a close visual inspection on a selection of records in the SDTM datasets versus its raw data, checking the data variable by variable, looking at possible truncation and other possible discrepancies. It may be assumed that some potential integrity issues will be triggered by conformance checks (e.g. checks on missing values of required variables, conformance to controlled terminology).

MAPPING OF FEEDING DATA

The mapping of feeding data of early life nutrition studies can be rather complex because of the high variety in the study designs, CRFs and diaries. In early life nutrition studies, designs often include a breastfed reference group (breastfeeding being the gold standard for infant nutrition). Some studies allow parents to breastfeed their child besides the usage of the study product (infant formula). Other studies, which are more complex in terms of design, allow participants to enter the study (i.e. start using study formula) at any time point without inclusion requirements on infant nutrition. Also, depending on age, weaning foods (e.g. fruits, vegetables, juices, water and tea) can be part of food intake. Hence to obtain insight in the feeding characteristics of the subjects the following feeding data are generally collected in early life nutrition studies:

- Feeding history data (e.g. whether breastfed and/or formula-fed, for how long, and introduction of weaning foods)
- Study product intake (e.g. number of feedings, amount of intake)
- Intake of other feedings than the study product during the study (e.g. breastfeeding, weaning foods)

In Figure 1 an example study design is shown: infants that are exclusively breastfed form the breastfed reference group and infants that are fully formula fed at time of inclusion are randomised to one of the two study products. Parents can introduce weaning foods during the study period. At baseline (visit 1) it has been assessed what type of feeding was received from birth to present (Figure 2 depicts example CRFs) and during the intervention period the intake of the study product, and the intake of other feedings including breastfeeding has been assessed by means of a daily diary (Figure 3 and 4 depict example diaries).

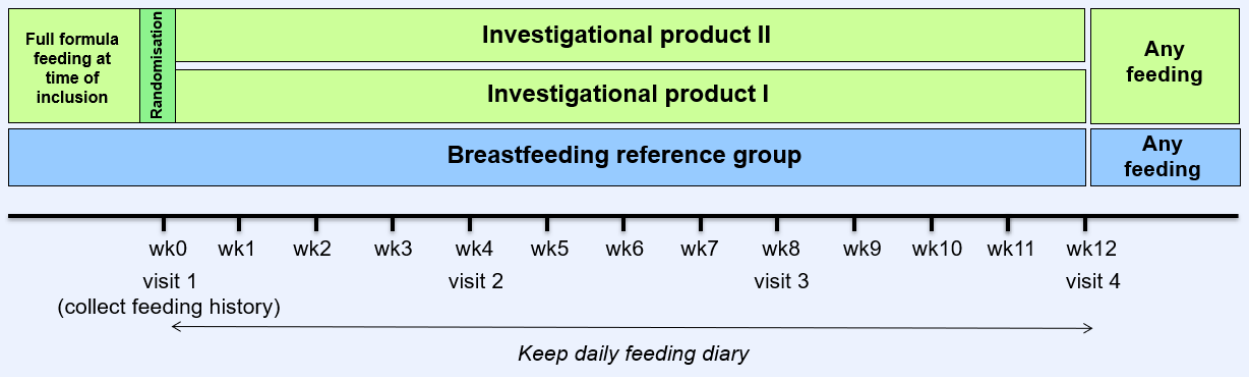
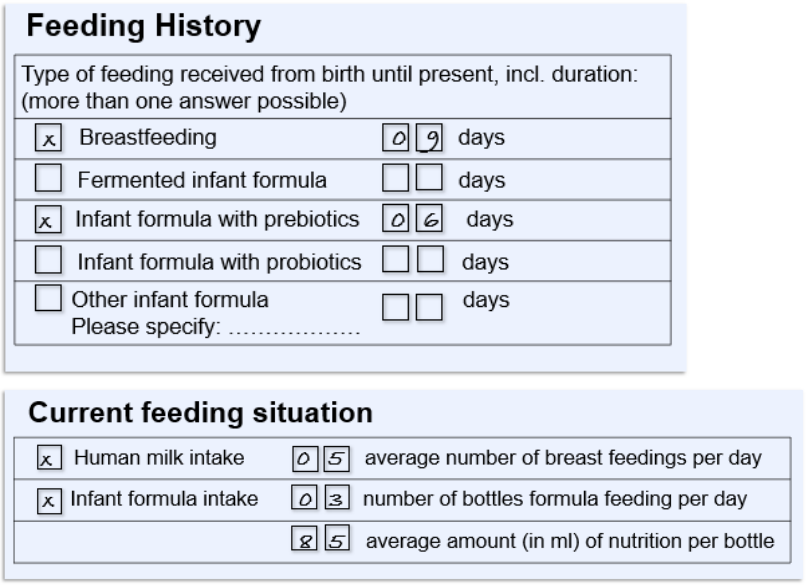


Figure 1. Example of study design for an infant nutrition study



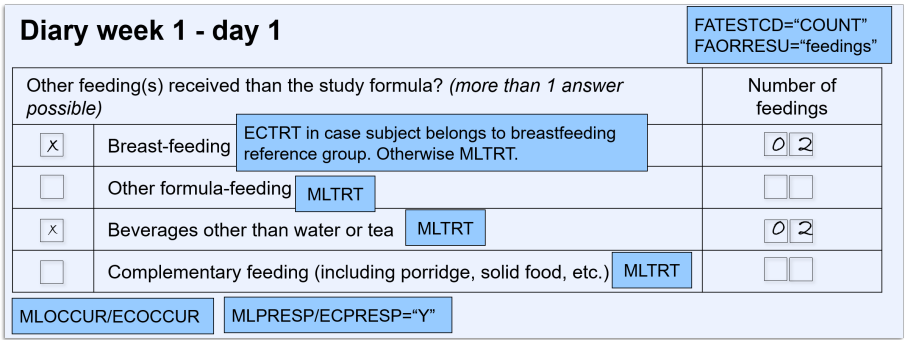
Feeding History

Type of feeding received from birth until present, incl. duration: (more than one answer possible)		
[x] Breastfeeding	09	days
[ ] Fermented infant formula		days
[x] Infant formula with prebiotics	06	days
[ ] Infant formula with probiotics		days
[ ] Other infant formula Please specify:		days

Current feeding situation

[x] Human milk intake	05	average number of breast feedings per day
[x] Infant formula intake	03	number of bottles formula feeding per day
	85	average amount (in ml) of nutrition per bottle

Figure 2. Examples of CRF sections collecting feeding (history) data



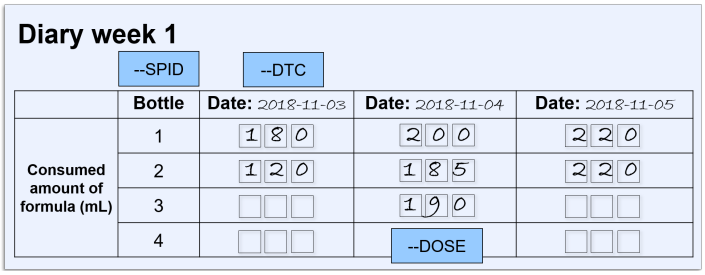
Diary week 1 - day 1

FATESTCD="COUNT"
FAORRESU="feedings"

Other feeding(s) received than the study formula? *(more than 1 answer possible)*		Number of feedings
[X]	Breast-feeding (ECTRT in case subject belongs to breastfeeding reference group. Otherwise MLTRT.)	02
[ ]	Other formula-feeding (MLTRT)	
[X]	Beverages other than water or tea (MLTRT)	02
[ ]	Complementary feeding (including porridge, solid food, etc.) (MLTRT)	

MLOCCUR/ECOCCUR
MLPRESP/ECPRESP="Y"

Figure 3. Example diaries on formula



Diary week 1

--SPID --DTC

	Bottle	Date: 2018-11-03	Date: 2018-11-04	Date: 2018-11-05
Consumed amount of formula (mL)	1	180	200	220
	2	120	185	220
	3		190	
	4		--DOSE	

Figure 4. Example diaries on formula

The first step when mapping data to SDTM is to determine what 'kind' of data is present (i.e. check the Case Report Forms and the collected data) and to check the SDTM Implementation Guide (SDTM IG) for a modeled domain which fits the data. The SDTM IG v3.2 does not contain a standard domain for feeding data. However, one of the new domains available in the draft release of SDTM IG v3.3 is the domain Meal Data (ML). This domain reflects collected details describing a subject's food product consumption. It belongs to the general observation class 'interventions' and has a structure of one record per food product occurrence or constant intake interval per subject. This domain is thus suitable for the representation of feeding data. Although study formula is fundamentally feeding data, it is above all the study product and exposure of interest. Therefore, the study formula data belong to domain Exposure as Collected (EC), which reflects protocol-specified study treatment administrations, as collected. Less obvious was the mapping of the breastfeeding data collected during the intervention period: does the data belong to domain Exposure as Collected, or does it belong to Meal? Together with subject matter experts at Nutricia it was decided that:

- When breastfeeding subjects were included as a study reference group (i.e these subjects do not receive

study formula), breastfeeding could be considered a protocol-specified study treatment and therefore the breastfeeding data collected during the intervention period belongs to domain EC.

- When subjects received breastfeeding in addition to the study formula (i.e. mixed feeding), study formula was considered as the protocol-specified study treatment and breastfeeding was considered as complementary feeding which belongs to domain ML.

All other collected feeding data, i.e. feeding history data and the intake of weaning foods were mapped to domain ML, because the collected data describes a subject's food product consumption, not the exposure to the study product.

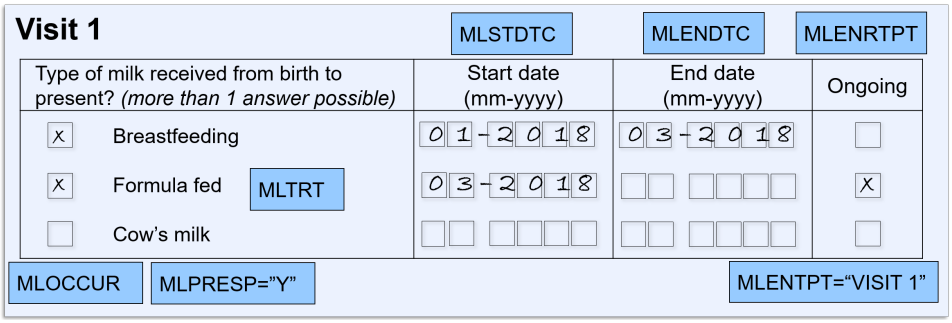


Figure 5. CRF and diary examples of the collection of feeding history data.

After defining the domain(s), the next step is to determine the SDTM variables needed to represent the data. In Figure 5 an example CRF with annotations of the SDTM variables are shown. In general, the following SDTM variables are used in both EC and ML - data collected for both domains are rather similar and therefore the same variables are used - to map the following data:

- --SPID: bottle number in case data has been collected per bottle
- --TRT: name of the feeding
- --PRESP: indicating whether the feeding type was pre-specified
- --OCCUR: occurrence of the pre-specified feeding types
- --DOSE: the (average) amount/volume of feeding consumed
- --DOSU: the unit associated with --DOSE
- --DUR: duration of the feeding
- --DTC: date/time of feeding assessment (i.e visit or diary date)
- --STDTC: start date of the feeding type
- --ENDTC: end date of the feeding type
- --EVINTX: evaluation interval associated with the feeding assessment
- --ENRTPT: indicates whether the feeding was still ongoing at timepoint defined in -ENTPT
- --ENTPT: description of timepoint (e.g. "VISIT 1").

The mapping of (average) number of bottles, the number of breastfeedings and the number of meals was an issue. According to the SDTM IG for domain Substance Use (general observation class Interventions) this data (e.g. comparable to number of cups of coffee) should be stored in the variable --DOSE. This variable was already used to capture the (average) amount/volume of feeding consumed (as described above), so this was no longer possible. The variable –DOSFRQ was also considered, but since for breastfeeding and also in some studies for formula feeding only the number of feedings was assessed and no dosing information (--DOSE) was available this variable did not seem appropriate. Although it also was considered storing these data in a supplemental qualifier variable, the data was ultimately mapped to Findings About (FA), because these data meet one of the criteria when data should be stored in domain FA: data or observations about an event or intervention which have qualifiers of their own that can be represented in findings variables (e.g., units, method). These data do have a unit (e.g. 'bottles/day', 'feedings/day) and therefore fit within domain FA.
To summarize, the amount/volume of feeding was stored in –DOSE and the number of feedings was stored in FA. Since the feeding data were now divided over two domains (Findings About with number of feeding data and Exposure as Collected/Meal Data capturing the other data), identifier variables FALNKID and ECLNKID/MLLNKID were introduced to the domains Findings About and Exposure as Collected/Meal Data respectively in order to link the related data via domain Related Records (RELREC).

In the example diary displayed in Figure 6 the actual consumed amount per bottle was recorded. In some studies, however, this was not explicitly asked. Instead it was assessed per bottle: i.e. how many spoons were used to prepare the product, what was the volume of the prepared product and what was the volume left over (see the example diary in Figure 4). These items did not fit into any of the Intervention class variables, and because the values came with qualifiers (i.e. units), it was decided to use the Findings About domain. Again, FALNKID and

MLLNKID/ECLNKID were used to link the related data via domain RELREC.

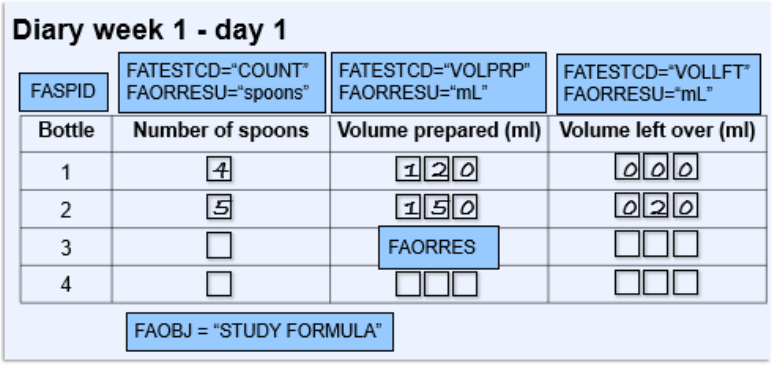


Figure 6. Simplified example diary on study formula with annotations

LABORATORY FAECAL PARAMETERS

In nutrition studies the analysis of faecal parameters (stool physiology and microbiology) is very common. Although the SDTM IG describes a Laboratory Test Results domain (LB) and a Microbiology domain (MB), neither was considered suitable for the type of analyses performed in nutrition studies. The LB domain is tailored to the requirements of mapping parameters with a direct relation to health safety, and emphasises that microbiology data should not be included in the LB domain. Hence, expected variables in LB such as LBORNRLO (Reference Range Lower Limit in Original Unit), LBORNRHI (Reference Range Upper Limit in Original Unit), LBNRIND (Normal Range Indicator) do not make any sense for the nutritional studies laboratory data, because the faecal parameters are explorative and highly variable depending on factors such as dietary intake and natural gut flora. Although MB is a findings domain just like LB, its use is quite different. The MB domain normally describes microbiology specimen findings, including gram stain results and organisms found (with for example MBTEST = "Organism Present", and MBORRES including the name of the organism). This is quite different than the nutrition study parameters which are targeted microorganisms or markers (e.g. --TESTCD should refer to the parameter, for example a microorganism) with quantified results (should be stored in --ORRES). Furthermore, similar to LB, MB is centred around reporting safety parameters of microbiology. Altogether, neither of the existing domains provide a solution that fits with all the nutrition studies laboratory data.

Since existing domains do not fit the data, it was decided to set up a custom domain: "XL - Non-Safety Laboratory Results", storing stool physiology data as well as microbiology data. The custom domain is a findings domain, and its design is very similar to the LB domain, with amongst others, the variables:

- XLTESTCD: Lab Test or Examination Short Name
- XLTEST: Lab Test or Examination Name
- XLORRES: Result or Finding in Original Units
- XLORRESU: Original Unit
- XLCAT: Category for Lab Test
- XLSCAT: Subcategory for Lab Test
- XLMETHOD: Method of Test or Examination

In order to store all relevant collected laboratory data, several supplemental qualifiers were added to the design:

- XLPROBE: Probe Name
- XLORLQLO: Lower Limit of Quantification in Orig Unit
- XLORLQHI: Upper Limit of Quantification in Orig Unit
- XLORLDLO: Lower Limit of Detection in Orig Unit

As the variable name suggests, XLPROBE stores the probe name when the analysis method involved using probes. The Lower/Upper Limits of Quantification/Detection store the limits that are specific for the vendor, batch, method and test.

The biggest challenge has been to introduce test codes (--TESTCD) and names (--TEST) for the faecal parameters. Of the 66 faecal parameters that were analysed, controlled terminology was only available for four lab tests in the NCI Controlled Terminology release of 2018-06-29. Therefore, new laboratory test codes and names were developed using the CDISC guidelines. The guidelines describe several rules for creating test names (--TEST), for example:

- Test names should not exceed 40 characters.
- Test names do not contain the specimen type.
- Test names do not contain units of measure.
- Test names do not contain methods of measurement.

- Test names do not contain collection timing information.
- For all differential test names, the numerator and denominator are spelled out as fully as possible (given the 40-character limitation) and separated by a forward slash. Do not use the words 'ratio' or 'percentage', as that is made clear in the definition.

For creating test codes (--TESTCD) the guidelines describe the following rules:

- Test codes should not exceed 8 characters.
- Make use of existing 'Naming Fragments' (e.g. using suffix 'AC' for a parameter containing the word 'acid').
- For all differential test codes, the absolute count is a short-defined term and the ratio/percentage contains the same short mnemonic for the numerator followed by a second short mnemonic for the denominator. There is no forward slash in the test code.
- For microbiology tests for targeted microorganisms, the first letter of the test code should be the first letter of the genus name.

For several parameters, the guidelines could be applied, for example when using suffix 'AC' for acids, and when dealing with relative measurements (Table 1).

Table 1. Example of mapping --TEST and --TEST.

Parameter	--TESTCD	--TEST
Acetic Acid	ACAC	Acetic Acid
Total Short Chain Fatty Acids	SCFA	Short Chain Fatty Acids
%Acetic Acid (relative to Total Short Chain Fatty Acids)	ACACSCFA	Acetic Acid / Short Chain Fatty Acids

It was concluded that it was not possible to comply to all guidelines for every parameter. First of all, the extensive microorganism names cannot be captured within 8 characters (--TESTCD) using the first letter of the genus name as mentioned in the guidelines, without resulting in duplicate codes. Neither could the description of the microorganism be captured in less than 40 characters (--TEST). For example, parameter 'Clostridium Histolyticum Group and Clostridium Lituseburense Group' already contains well over 40 characters, but it is the minimum description of the parameter in order to be specific. Following the guidelines, the test name for the relative measurement of this parameter should also include a slash ("/") and the base to which the relative measurement was calculated (e.g. 'Clostridium Histolyticum Group and Clostridium Lituseburense Group / Total Bacteria'), resulting in an even longer description. Together with subject matter experts, multiple approaches for establishing terms in a systematic way were discussed. As a solution, but knowing that method information should not be included in the test codes or names, for the microbiology parameters analysed involving the use of probes, the probe names were used to create terms (e.g. CHIS150CLIT135). This was decided because nothing described the parameter test more specifically than the probe name. For other parameters where probe names were not applicable due to other use of methods, a systematic coding structure was introduced considering the genus and species names (but not limited to using the first letter) and whether the parameter targets a group, cluster, species, superspecies etc. (e.g. CAJEJUSP for Campylobacter Jejuni Sp. or BIBIFI for Bifidobacterium Bifidum). Knowing that the converted laboratory data would only be used for internal purposes (i.e. not for submission purposes), it was agreed to discard the rule for maximum length for test codes and names when necessary.

All decisions made together with subject matter experts on creating the custom design and new terminology have been incorporated into a laboratory template at Nutricia. Hence, data from new studies will be collected following the template, which makes future mappings to the custom domain XL smooth and easy.

STOOL CHARACTERISTICS AND GASTROINTESTINAL SYMPTOMS

Stool characteristics and gastrointestinal symptoms, are commonly assessed in nutrition clinical trials to evaluate product tolerability, see an example CRF in Figure 5. One of the challenges in the mapping of these data is that there is no single modelled domain available in CDISC SDTM to represent these data. We decided to map these data to domain Findings About (FA) and Clinical Events (CE). The mapping of these data has been extensively explained by means of example CRFs and diaries in a paper of Lieke Gijsbers [1]. In short, the following approach was used in the mapping of these data:

- The stool characteristics data, such as the stool frequency, consistency, colour and volume, can be interpreted as data about the event 'having a stool' for which no event record is created, and therefore these data fit within domain Findings About. To distinguish between number of stools counted on a single day, mostly assessed via a diary, and the average daily number of stools, mostly assessed via a CRF, different terms of FATESTCD ('COUNT', 'A_COUNT') and FATEST ('Count', 'Average Count') were used. The same approach was used for characteristics per stool (e.g. FATESTCD = 'CONSIS', 'COLOR', 'VOLUME') versus general/average stool characteristics (e.g. FATESTCD = 'A_CONSIS', 'A_COLOR', 'A_VOLUME').

- The gastrointestinal symptoms data were usually collected as pre-specified events, combining the collection of the absence/presence of the event and the severity of the event. These data perfectly fit within domain Clinical Events. However, in some studies also the frequency of the gastrointestinal symptoms was recorded. These data were mapped to domain Findings About. In this case, the data originating from one CRF (and one raw dataset) was split in two domains.

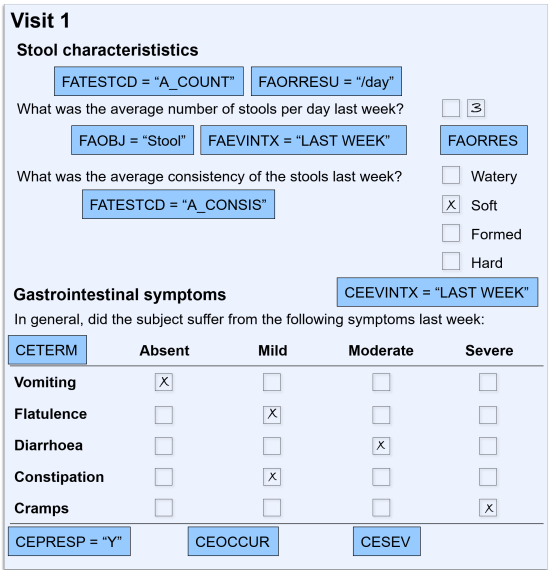


Figure 5. A Case Report Form on stool characteristics and gastrointestinal symptoms with annotations

CONCLUSION

This paper gave an overview of a SDTM conversion project for non-submission purposes involving twenty-one nutrition clinical studies. We described our project approach, the tools and solutions that were developed, and we presented several worked cases to exemplify the challenges and solutions in scope of this project.

ACKNOWLEDGMENTS

We thank Roel Hobo, Elsbeth Verdonk, Bas van Bakel, Leah van der Meer and Jules van der Zalm for their contribution to the project. We are grateful for Leah's support in writing this paper.

REFERENCES

1. Lieke Gijsbers. Mapping of Gastrointestinal Tolerance Data to SDTM, PhUSE 2017, Paper PP25 ("www.phusewiki.org/docs/Conference 2017 PP Papers/PP25.pdf")
2. Angelo Tinazzi and Cedric Marchand. An FDA Submission Experience Using the CDISC Standards. Paper RG04, PhUSE 2017. ("www.phusewiki.org/docs/Conference 2017 RG Papers/RG04 Paper NEW.pdf")

CONTACT INFORMATION

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

Author Name	Paul Vervuren
Company	Nutricia Research
Address	Uppsalalaan 12
City / Postcode	3584CT Utrecht
Work Phone:	+31 30 2095 000
Email:	paul.vervuren@danone.com
Web:	www.nutriciaresearch.com

APPENDIX

Definition of Probability, Impact and Non-detectability for risk assessment

Probability

1	Issues cannot happen and did not happen in the past
2	Issue are known to happen in exceptional circumstances
3	Issue are known to happen with reasonable frequency
4	Issues happens frequently and re-occurs often
5	Issue are happening frequently or are reported often

Possible Impact

1	No possible impact can be identified neither on the subject/patient, user or on the data in whatever form
2	Slight delay in the production of results, results are correct
3	No results can be produced, or there is a loss of time for having correct results
4	Important delays in producing results that might endanger the subject/patient or slightly incorrect results are produced (e.g. way they are presented might be misinterpreted)
5	There is a direct safety hazard for the subject/patient; this can be due to corrupted results that are displayed

Non-detectability

1	Any issue or malfunction is immediately detected and reported; any possible malfunction is stopped automatically or eliminated
2	The issue is reported and requires intervention to allow continued functioning
3	The issue generates an alarm or signal but the intervention needed is not clear and short-cuts are possible or the user does not see any results appear while he is waiting for the result
4	The issue generates an alarm or signal but the issue continues anyway and no intervention is requested; in other cases the issue is only detectable a posteriori
5	An issue or malfunction can go completely unnoticed