PharmaSUG China 2023 - Paper DS-174

Implementation and challenge of submission package translation from Chinese to English

Linqing Wang, HUTCHMED Limited

ABSTRACT

With increasing study submission across contries, high-quality translated data become more and more important. During the translation process, we usually wonder how to streamline the translation process so that we can lower the risk of re-work, if any methods to ensure translation accuracy and consistency, how to review translated data package and any evidence to prove the translation well performed.

This paper will introduce our experience in the implementation of submission package translation from chinese to english, and our efforts to explore a standard translation process.

INTRODUCTION

In 2022, we were preparing the FDA submission. Several Chinese study packages were about to be tranlated from Chinese into English. A tranlstion vendor involved in to tranaslate dataset and files. A key question emerged. That is how to plan the translation process in order to ensure quality?

So before translation started, we had several rounds of discussion to:

- gather all the datasets and documents to be translated
- had an alignment for translation requirement
- distinguish all the potential issues that may encountered during the translation and set resolutions
- draft a translation plan and set role of responsibility
- set a list of all the output for translation

We had a picture of a draft translation plan. Still a lot of issues emerged. Let's start to introduce!

TRANSLATION PLAN

The purpose is to 'translate' Chinese define packages into English define packages to supoort FDA submission. Programming held internal discussion, a brief plan as below:

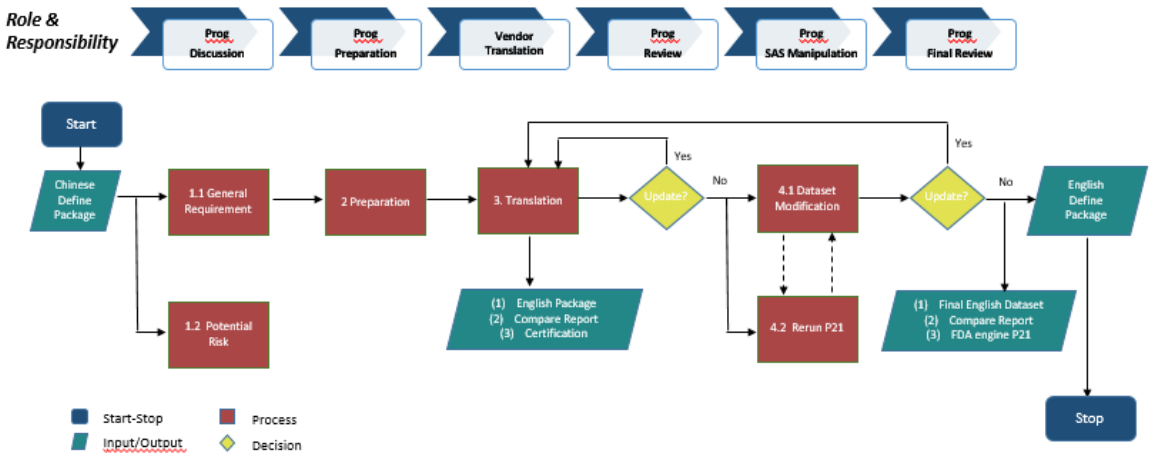

Figure 1. Translation Plan

GENERAL REQUIREMENT

Translation of clinical study data is not only language transformation in data, it also a challenge for programmers to ensure it meets submission guidance and CDISC standard. Beside that, consistency with dataset and between datasets and documents is also a key aspect to evaluate the quality of translation in which brings difficulties in reviewing whole package.

POTENTIAL RISK

For a translator that have no certain experience in SAS and understandidng of CDISC standard, a huge gap exists between programmer and translator.

So first of all, the potential risk in data translation should be identified in advance. Such as:

- Length:

  Long text value may come from multiple variables and should be translated as a whole sentence

  Length of –TEST, QLABEL should within 40

  Length changed after translation

  Re-manipulation length of variables may lead to record/variable number change
- Controlled Terminology should be followed.
- Label:

  Standard label and naming convention should be followed and assigned.

  If translator capable of assigning label?
- Dictionary:

  Translated drug name or event name should be consistent with dictionary
- P21 report:

  Issues vary in different engine, more issues if using FDA engine
- Consistency

  Within SDTM datasets: same variables in multiple dataset

  Within SDTM and ADaM: ADaM variable and its predecessor

  Within ADaM and External datasets

  Within datasets and documents

Corresponding resolution were pre-defined:

- Length:

  long text value translation: SDTM parent domain incorporate with SUPP can be provided and a reference should be provided

  Length of --TEST, QLABEL: pre-defined in CT or modified after translation

  Length changed: programming manipulation and change in variable/record number should be

explained.

- Controlled Terminology:

  CT reference should be followed and guide translator of the usage

- Label:

  Do not translate label and assign label by programming

- Dictionary

  Re-map to En value by programming

- P21 Report:

  For previous issue: remain the explain

  For each new issue: it should be evaluated whether perform modication to reduce P21 issue.

- For long text value, a reference should be provide to reveal related variables. Expecially for SDTM a SDTM dataset that already incoorparate SUPP in can be provided to furcilitate translation.

- Consistency

  Demonstrate on the relationship of datasets and documents

  Programmatic review and peer review

Considerring the above risk (also other risk that not identified) and pre-defined resolutions. The preparation work is very import.

PREPARATION

As listed below, the material are all required for translator:

Type	Content	
Datasets	SDTM, ADaM	In sas7bdat
Datasets	External file to derive ADaM[(a)]	In sas7bdat or pdf
Documents	cSDRG, ADRG	In editable doc
Documents	aCRF	In pdf
Documents	SDTM/ADaM specification[(b)]	In editable xlsx
Reference	including long_string_text[(c)], codelist[(d)], dataset label[(e)], variable label[(f)]	In editable xlsx
Reference	Translated protocol and SAP	

(a) per FDA data standards catlog, if external files are source files to derive ADaM datasets, they should be converted into pdf files if SAS program read the external file directrly or submit .xpt files if they were created into customized dataset before SAS program read.

(b) Only part of Tabs need translation. Such as Methods

(c)since long string value may exist so that its value must be stored in mutilple variables. The list is to show the relationship of these related variables to support translation

(d) As standard controlled terminology value should be used and sponsor defined codelist may exist. A reference to reflect the original value and translated value are desired.

(e) (f) not a necessary part in this step. This in only depends on the capability of translation vendor. Whether label is assigned by translation venfor or not, Programmer can prepare label translation by referring to CDISC.

TRANSLATION

After translation, vendor will send back all translated material:

Type	Content	
Datasets	SDTM, ADaM	In sas7bdat
Datasets	External file to derive ADaM	In sas7bdat or pdf
Documents	cSDRG, ADRG	in editable doc and pdf
Documents	aCRF	in pdf
Documents	SDTM/ADaM specification	In editable xlsx
Documents	Compare Report: English vs Chinese	In pdf
Documents	Certification	

TRANSLATION REVIEW

Translation reviewing is a very detailed work.

Compare report should be reviewed first to see:

- If all datasets were compared
- If only chinese variable were translated and compare different
- If un-translated variables compare different.
- If variable numbers and variable types were the same.

Then start dataset reviewing by programming and by peer review. It may encounter many unexpected issues during the reviewing:

- Dataset size were extremely large and hard to open or manipulate.
- Type of un-translated variable were changed. Such as ‘xxID’,’xxDTC’,’IDVARVAL’,’xxENTPT’
- Time is truncated in datetime variable
- Case Sensitive: in most case, only first character in first word is capitalized
- QLABEL length and xxTEST are not all limited within 40.
- QLABEL translation didn’t follow naming convention
- SUPP variables value translation is not appropriate
- Datasets inconsistency: such as AECONTRT in SUPPAE and ADAE, --STRESC in SDTM and AVALC in ADaM, External datasets and ADaM
- Datasets and Documents inconsistency: such as TI/TS and protocol, SDTM and aCRF
- Trial Design: such as IETEST length exceed 200
- External file translate is not consistent with ADaM
- Chinese text still exist
- Some general practice may be hard to explain: such as ‘既往’ may be translated into ‘Past’, but general practice is ‘history’ in MH and ‘Prior’ in CM/PR. ‘治疗’ can be translated as ‘Treatment’,

but it is ‘Therapy’ in PR.

Response to above issues are necessary, but it also lead us to pay attention:

- Most CT were case sensitive value, it must be demonstrated before translation
- Trial design domain is not necessary for translation. And for TI, a full version and a short version can be prepared.
- Better to provide all SDTM parent domain incorporate with SUPP
- Reference should also be applied in aCRF
- External dataset is not necessary for translation, especially when they are source from other SDTM/ADaM or large amount of record which means hard to review.
- Better to develop a checklist for Consistency

DATASET MANIPULATION

After receive updated data package and no more issue found, Programmer can started data manipulaton. Data manipulaton aim to solve below issue:

- Assign label to datasets and variables
- Split variables for long text. SUPP dataset may be re-generated.
- Re-map to English Dictionary
- Other necessary manipulation per P21 and translation
- Compare Report to compare translated En datasets with Manipulation En datasets

Programmer will re-run a Pinnacle 21 report using FDA engine. New message may emerged and certain update must be made to reduce P21 issues:

- Inconsistent value for AEDECOD with AETERM

  The case happen to ‘心率异常’ and ‘心率失常’ were translated into same english value, while their AEDECOD are different. Another case was ‘血小板计数降低’ and ‘血小板减少症’
- QLABEL length must be restricted to 40
- DSDECOD and DSTERM unconsistency when DSCAT = ‘PROTOCOL MILESTONE’;
- EPOCH should be set to null when DSCAT = ‘PROTOCOL MILESTONE’;
- EPOCH value inconsistency
- Chinese punctuation can not be compressed.

If the issues were caused by translation, then report the issue to translation update. Other issue which need programming manipulation must be recorded. After data manipuation, datasets number should be compared first and then compare English datasets with manipulated Englisth dataset. Any compare difference in variables and variable numbers should be checked and explained that it caused by data manipulation. A modification report must be generated to summarized general modification and details by domain.

SUMMARY

From Chinese define package to English define package, you may found it hard to expected and take a lot work in reviewing and manipulation. Going back to the translation process, let’s make a summary:

For preparation:

Not all materials require translation. Be patient in identifying the part need translation or manipulation, and a reference is helpful:

Type	Index		
SDTM	1	Format	• in sas7bdat • Parent domian incorporate with SUPP is more friendly in translation and translation consistency
	2	Trial Design	• Do not translate. • Create according to translated protocol and IG • Trial Inclusion/Exclusion Criteria: Create a full text version for cSDRG and a short text version for TI.
	3	Dictionary	• Do not translate. • Map to English MedDRA/WhoDrug by programming
	4	Label	• Do not translation. • Prepare labels according to IG and assigned by programming, including datasets, variables and QLABEL
ADaM	1	Format	• in sas7bdat
	2	Dictionary	• Do not translate. • Map to English MedDRA/WhoDrug by programming
	3	Label	• Do not translation. • Prepare labels according to IG and assigned by programming, including datasets and variables
External Datasets	1	Format	• In sas7bdat if source file is read in and stored as sas dataset before use. Convert to xpt in define package. • In pdf if source file is read in SAS program directly. • Name in lower case only, space and undersore are not allowed
cSDRG	1		• In editable doc • Do not translate 'Data Conformance Summary' and 'Appendix of IE Criteria' • Highlight the importance of referring to translated protocol
ADRG	1		• in editable doc • Do not translate 'Data Conformance Summary'
aCRF	1		• In pdf
Define Specfication	1		• In editable xlsx • Review in advance to find out which tabs need translation, such as Method Tab
Reference	1	Long_text	• Include the dataset name and variables name that should be translated as a whole sentence

	2	Controlled Terminology	• Based on original CT version. • Suggest to cover all all –TEST, --CAT, --SCAT. • Certain codelist for SUPP variables if there are desired value. • Filter out all the dataset and variables applying codelist
Other	1	Translated Protocol	
	2	Translated SAP	

For Translation:

Some guidance before translation is preferable:

Type	Index	
CT Usuage	1	• CT is case sensitive
	2	• CT is also applicable for aCRF translation (by domain)
Consistency	1	Within SDTM and Within ADaM • same name, same value
	2	Within SDTM and ADaM • Same value for ADaM variable and its predecessor. Such as –STRESC and AVALC, --TRT, --TERM
	3	Within ADaM and External datasets • same name, same value
	4	Within Dataset and Documents: • SDTM and aCRF
	5	Within Documents • Protocol and cSDRG • Protocol and ADaM
Variable Type	1	• Do not change variables type, especially for Date and Numeric variable
Datetime	1	• Check if any truncation
Punctuation	1	• Chinese Punctuation should be replaced.

Translation Review:

Type	Index		
Datasets	1	SDTM & ADaM & External Datasets	• Dataset Size • Variable Number and Variable Type • Truncation • If CT are applied and case sensitive. • Free-Text or variable without CT: by program or peer review • A checklist can be prepared for programmatic review and consistency review

Documents	1	cSDRG & ADRG	• Check against protocol • Modified directly • Finalized until P21 re-run
	2	aCRF	• Check against SDTM • Modified directly
	3	Specification	• Peer review • Modified directly
	4	Compare Report	• If all datasets compared • If only Chinese variable were translated and compare different • If un-translated variables compare different, such as datetime • If variable numbers and variable types were the same.

Datasets Manipulation:

Type	Index		
Datasets	1	SDTM & ADaM & External Datasets	• Assign Label to datasets and variables • Split variables for long text. SUPP dataset may be re-generated. • Re-map to English Dictionary • Other necessary manipulation per P21 and translation • Compare Report to compare translated En datasets with Manipulation En datasets • Any chinese character or punctuation
Documents	1	SDTM & ADaM P21	• Re-run using FDA Engine • Check each issue and judge if any modification to reduce issue. • Check if issue caused by translation
	2	Compare Report	• Compare Modified En dataset with Translated En dataset
	3	Modification Report	• Consider all the update by program pre-planed and issues required modification in P21 • A general description and details by domain

CONCLUSION

Datasets translation is a complex and detail work especially for chinese translation. Since translator usually has no experience in SAS programming and no idea on CIDSC requirement or files. Programmer must be proactive in moving on the corporation of translator and programmer. A well-prepared and systermic plan plays a key role in ensuring high-quality and high-efficient translation. Besides, a pre-defined chekclist for dataset and consistency is also helpful. So be prepared, it will be smoothly.

REFERENCES

FDA “STUDY DATA TECHNICAL CONFORMANCE GUIDE”

https://www.fda.gov/regulatory-information/search-fda-guidance-documents/study-data-technical-conformance-guide-technical-specifications-document

CONTACT INFORMATION

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

Linqing Wang
HUTCHMED Limited
13162570753
Linqingw@hutch-med.com