PharmaSUG China 2023 - Paper AP-191

Subject eCRF (Casebook) Generation and Translation

Yiding Qu, Jiangsu Hengrui Pharmaceuticals Company Ltd., China

ABSTRACT

In recent years, Chinese pharmaceutical companies have increasingly ventured into the international market. However, they encounter various challenges when submitting their new drug applications (NDAs) to the U.S. Food and Drug Administration (FDA). One major hurdle is the language barrier, as the FDA mandates that all submitted documents be in English. Additionally, the FDA may request additional pooled analyses, such as the Integrated Summary of Safety (ISS) or Integrated Summary of Efficacy (ISE), along with the relevant documents from each individual study included in the analysis. Translating all these documents in-house poses a significant challenge for Chinese pharmaceutical companies, while hiring qualified translators can be both time-consuming and expensive.

To address this issue, we have developed a tool that streamlines the translation of Subject eCRF (Casebook) with minimal reliance on human translators and financial resources. By leveraging the original Chinese blank eCRF and EDC datasets, our tool can generate comprehensive Subject eCRF documents in English within seconds, amounting to thousands of pages. While the tool is EDC-dependent, the underlying concept and logic are straightforward and uncomplicated. We are excited to share our approach with you and provide insights on automating the document translation process.

INTRODUCTION

The subject eCRF, also known as a casebook, is an electronic form utilized to capture and record data for each individual participant enrolled in a clinical trial. It serves as a comprehensive repository of all information collected during the trial concerning that subject. The subject eCRF is essentially a CRF (Case Report Form) populated with data specific to a single participant, organized by various visits and sections. In many instances, regulatory authorities such as the CDE (Center for Drug Evaluation) or FDA (Food and Drug Administration) require the submission of the subject eCRF as part of the NDA (New Drug Application) process.

Typically, the generation of casebooks is handled by the EDC (Electronic Data Capture) system itself. However, in situations where the study is conducted in a country different from the regulatory agency to which the application is being submitted, translation becomes necessary. In our specific case, the studies were conducted in China, and we intend to submit our application to the FDA, which requires all documents to be in English. This includes the casebook, as the data collected in the EDC system is in Chinese and needs to be translated into English. This task goes beyond the scope of the casebook generation program within the EDC system.

At this point, we have several options to consider. Firstly, we can opt to translate the blank CRF, code list, and all raw data into English. By doing so, we would rebuild the entire database and utilize the casebook generation program within the EDC system to generate the casebook in English. Alternatively, we could export all the casebook materials in Chinese and then either handle the translation work in-house or outsource it to professional vendors. However, both approaches involve a significant amount of translation work, which incurs substantial costs in terms of both money and time. Additionally, ensuring consistency between the translated documents and the original Chinese CRF and raw data can be challenging to maintain.

We uncovered a different solution, rather than translating hundreds of thousands of pages of documents or the entire database, our solution requires minimal financial cost and can easily maintain the consistency between the translated documents and the original database.

RAW DATA TRANSLATION

Nevertheless, we need to translate the raw data. As SAS programmers, we receive the raw data in the form of SAS datasets, with one dataset per domain, and possibly a format catalog file, depending on the

type of EDC used. It is reasonable to assume that we might be able to transform those datasets from Chinese to English through SAS programs, with assistance from the professional translators. Thus, we developed three SAS macros.

The first macro is designed to identify the portions of the data that require translation, specifically the Chinese characters. This helps pinpoint the areas that need to be translated.

The second macro serves the purpose of cross validating the translated text with the original file. This validation step ensures the accuracy and consistency of the translation.

Finally, the third macro enables the recreation of the datasets with the translated text. It effectively generates new datasets that incorporate the translated information, allowing for seamless integration into the English casebook generation process.

IDENTIFY THE CHINESE FROM RAW DATASET

Our first SAS macro is called “u_ds_dbcs_chk_raw”, which is used to identify all Chinese characters within the raw datasets. The input is the libname of the raw datasets and output an Excel file. Each spreadsheet represents a raw dataset, and each column represents a variable with Chinese characters. An additional column with post fix “_E” is added beside each variable, representing the English translation of this specific Chinese text.

In order to minimize the translation work, only unique Chinese text for this variable is added into the Excel file. This file would then be sent to the translators and their job would be to fill in the column with post fix “_E” with the corresponding English translation.

Figure 1

NOBS	PUB_TNAME	PUB_TNAME_E	AETERM	AETERM_E
1	不良事件		ALT升高	
2			AST升高	
3			B型钠尿肽前体升高	
4			C反应蛋白升高	
5			D-二聚体升高	
6			D-二聚体增高	
7			D二聚体升高	
8			II III AVF V5 V6 ST段呈水平型压低（0.05mv）	
9			I度房室传导阻滞	
10			PICC左臂置管处感染	
11			QT间期延长	
12			ST-T改变	
13			ST段呈水平型压低（V4-V6）0.05-0.15MV	
14			ST段轻度改变	

Figure 1 Output Excel File of *u_ds_dbcs_chk_raw*

VALIDATE THE TRANSLATION

Once the translation is completed, the translated file would be sent back to the SAS programs for further processing. However, when it involves file transferring and editing, unexpected events could happen. For example, accidental removal of spreadsheets, columns, or data, unintentional changes to cell formats or content, encoding problems, and more.

To address these risks and ensure the accuracy of the translated file, it is essential to perform cross-validation by comparing it with the original file. To achieve that, we developed a second SAS macro called “u_dbcsfile_chk”. This macro takes the input of the original file sent to the translator and the file sent back, spot any difference between the two files and ensure all Chinese text is translated. This validation step adds an extra layer of assurance to guarantee the quality and accuracy of the translated file before proceeding.

Figure 2

NOBS	PUB_TNAME	PUB_TNAME_E	AETERM	AETERM_E
1	不良事件	Adverse Event	ALT升高	ALT increased
2			AST升高	AST increased
3			B型钠尿肽前体升高	Pro-B-type natriuretic peptide increased
4			C反应蛋白升高	C-reactive protein increased
5			D-二聚体升高	D-dimer increased
6			D-二聚体增高	D-dimer increased
7			D二聚体升高	D-dimer increased
8			II III AVF V5 V6 ST段呈水平型压低（0.05mv）	Horizontal ST segment depression (0.05 mv) in leads II, III, AVF, V5 and V6
9			I度房室传导阻滞	First degree atrioventricular block
10			PICC左臂置管处感染	Infection at the site of PICC line insertion of left arm

Figure 2 File with Translated Text

RECREATE RAW DATASETS

We are now ready to translate our Chinese raw datasets to English with the third SAS macro “u_ds_dbcs2sbcs_raw”. This macro would take the input of the original SAS datasets in Chinese and the translation file, combine them together and recreate the raw datasets in English. During this process, the dataset and variable names would remain unchanged, the values of each variable would be translated in English, and the format, label, and length of the variables would also be adjusted accordingly.

SUMMARY

The entire raw data translation process can be summarized as in the following diagram. First all Chinese text with the raw datasets were identified and put in a blank translation file which would be translated by the professional translators. The translated file received would then be validated, any inconsistencies would cause the validation process to fail and sent back to the translators. After the validation, the translation would be incorporated into the original Chinese raw datasets to generate the fully translated English raw datasets.

Figure 3

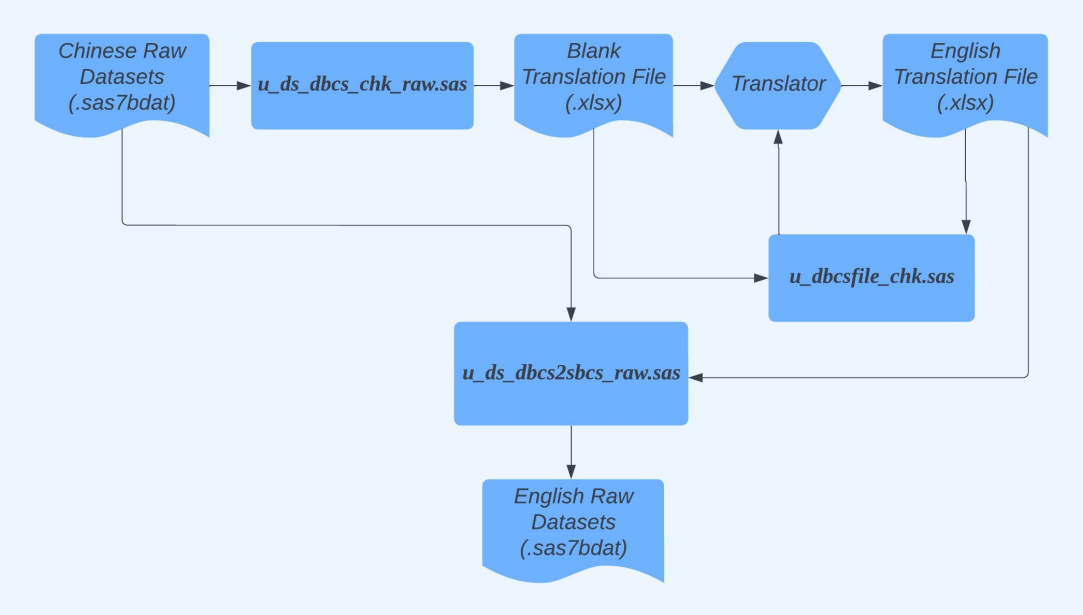


Figure 3 Raw data translation process

CRF TRANSLATION

Besides the raw data, the structure and appearance of the casebook depends mostly on the blank eCRF, hence, the blank eCRF must be translated to English. However, most of the time, the eCRF is a MS Word or PDF file, which is not standardized or structured, extracting information from these files would be a huge challenge.

In addition, the full eCRF contains a lot of repeated information, the exact same forms may appear in different visits, translating the entire document would cause further waste of time and budget, on top of that, maintaining consistency of translation among those same forms would be another issue.

Therefore, we chose to translate the database definition file of the CRF (eSDM file), rather than the CRF itself. This file fully describes all the details of the CRF, including the label of each form and field; length, code list, special attributes of each field; the corresponding SAS dataset name or variable name of each form and field; name of each visit and the full list of forms included in this visit; and many more. Basically, it contains all the information you need to recreate the entire CRF with the exact same format generated by the EDC system itself. The only information within this file would be the label of each form, field and visit, which is far more efficient than translating the full eCRF.

Note that depending on the types of EDC used, the visit information might not be included in the database definition file, if that is the case, you might need to gather that information in a separate file, i.e., an excel spreadsheet with the relationship of visits and forms; and translate it.

Figure 4

表	表名称	表名称英文	变量名	变量名英文	变量	变量类型	变量长度	编码名称	是否只读
subject	受试者信息	Subject Information	受试者代码	Subject ID	subjid	字符	255		
subject	受试者信息	Subject Information	项目代码	Protocol ID	studyid	字符	255		是
subject	受试者信息	Subject Information	研究中心名称	Name of study site	sitename	字符	255		
subject	受试者信息	Subject Information	研究中心编号	Study site No.	siteid	字符	255		是
subject	受试者信息	Subject Information	研究者姓名	Name of investigator	invname	字符	255		
subject	受试者信息	Subject Information	研究者代码	Investigator ID	invid	字符	255		是
subject	受试者信息	Subject Information	CRF状态	CRF status	lockstat	单选	20		是
subject	受试者信息	Subject Information	项目名称	Project name	studyname	长字符			是
subject	受试者信息	Subject Information	姓名缩写	Initials	subjinit	字符	255		
subject	受试者信息	Subject Information	受试者状态	Subject status	status	单选	20	SUBJECT_STATUS	是

Figure 4 Part of the eSDM File

CASEBOOK GENERATION

At this point, we have prepared the English SAS raw datasets, English eSDM file and possibly a translated visit table file (representing the relationship of visits and forms), and we can proceed to generate the casebook for each subject. The logic is straightforward, utilize the eSDM file and visit table file to generate a blank template full CRF, for each subject, find the corresponding data within the raw datasets and fill it in the template, and finally output the filled CRF into a pdf file, one for each subject, as show below. Based on the above workflow, we developed a tool Casebook Generator, to automatically generate thousands of pages of casebook per subject.

Our tool has been developed using the C# programming language and the .NET Framework. The user interface, input parameters and files of this tool are EDC-dependent. In our case, we need the header information to generate the cover page, the path of three inputs, raw datasets in SAS format, eSDM file and Visit Table file; and the path of the output casebook documents, language settings to control the style and font family of the output PDF file, and a special setting for splitting dataset, since our EDC might combine data from different forms into one SAS dataset.

While the specific settings may vary between different EDC systems, the fundamental principle remains the same. The casebook is generated by populating the translated data into a CRF template, which is created based on the database definition file of the EDC system. This process ensures that the translated data is accurately inserted into the appropriate fields and sections of the casebook, following the structure defined by the EDC system's database definition.

Figure 5

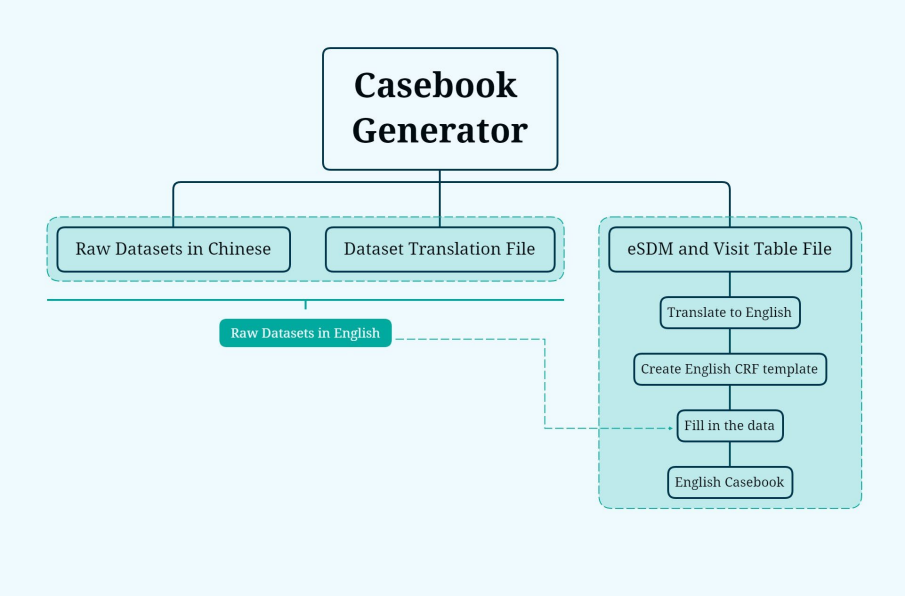


Figure 5 Workflow of Casebook Generator

Figure 6

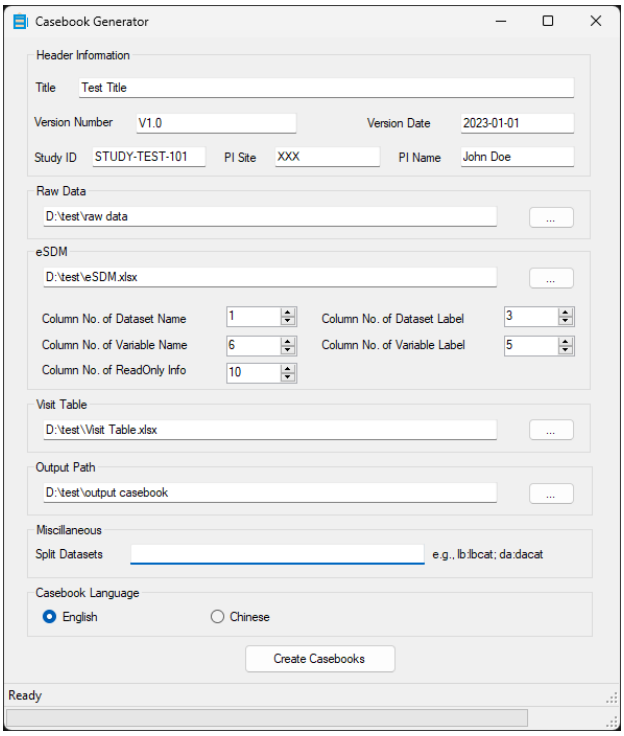


Figure 6 User Interface of Casebook Generator

Apparently, building this tool is not an easy task. The raw SAS datasets need to be organized in a structured way which corresponds to the forms and fields in the CRF. A thorough understanding of all the details in the CRF, and the internal logic of EDC system is required. Benefiting from the Object-Oriented feature of modern programming languages, the structure of the Casebook Generator program could be designed to mimic the structure as in the CRF page, as shown in following diagram.

Figure 7

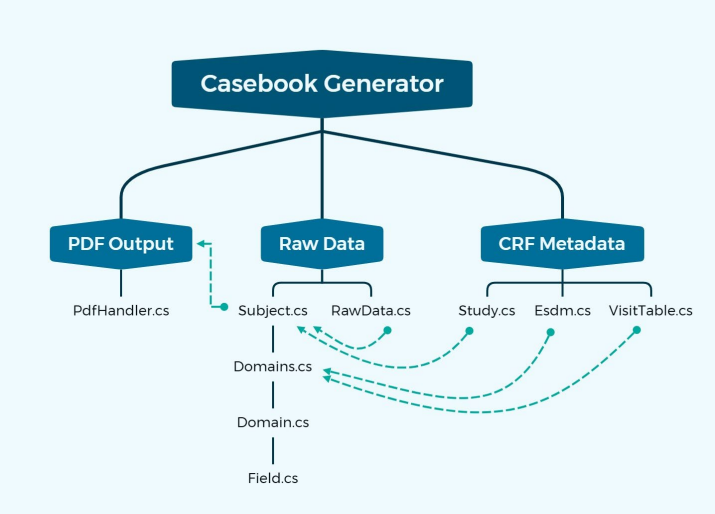
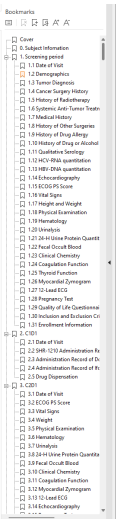


Figure 7 Structure of Casebook Generator

The data structure of the CRF is described by four objects, Subject, Domains, Domain and Field. Data from each subject could be broken down into different domains, representing different forms in CRF, domains repeated within different visits is generalized to a same class of domain, and each domain is further broken down into fields, representing a single data point in the CRF, each field would have its own attributes, such as label, length, format, code list and other special characteristics. The study metadata, eSDM file and Visit Table file are first loaded into the program to determine the structure and attributes of domains and fields of each subject, building a full tree of objects with missing data. After that, raw data is split by subjects, data from each subject is then filled into each Field object to fulfill the tree. The subject object would then be fed into a PdfHandler object to output the casebook document of this subject in PDF format.

This model could be generalized to any EDC system; however, details of the program could vary considerably. Unraveling the relationship between the structure of CRF and the raw datasets; and maintaining consistency between the original CRF and the appearance of the output PDF document was the toughest part. Consulting with colleagues from the data management department or even the designer of the EDC might be helpful.

Figure 8



1.2 Demographics

Subject ID: 01001	CRF status: Signed by PI
Name of study Site: [illegible]	Study site No.: 01
Name of investigator: [illegible]	Investigator ID: [illegible]
Visit name: Screening period	Visit No.: 1
Demographics	
Date of signing informed consent form: 2018-11-22	
Date of birth: 1963-07-08	
Age: 55	
Age unit: Years	
Gender: Male	

Figure 8 Snapshot of Casebook Generator Output

Finally, the output of Casebook Generator reflects the entire database and the amount of data included is enormous. For a regular phase III clinical trial, the casebook of each subject could easily reach thousands of pages, complete validation of these documents is nearly impossible. Therefore, keeping track of every step within the program and maintaining a detailed log is essential. Identify any suspicious inconsistency during the mapping process within the log, for example, an input SAS raw dataset has no corresponding CRF page, a variable has no corresponding field in the CRF, or the other way around. Most of the time, these issues were caused by some special settings of the CRF, however, keeping detailed log would be a great help for debugging and investigating if something goes south, especially when the output has such a massive scale and requires pinpoint accuracy.

Figure 9

```
        domain.
113228  2022-12-01 17:43:36.0396 Info|Finished filling data of domain End of study page-Date of Visit of this subject.
113229  2022-12-01 17:43:36.0396 Info|Start filling data of domain End of study page-End of Study Page of this subject from raw dataset ds...
113230  2022-12-01 17:43:36.0555 Info|Filling data for domain End of study page-End of Study Page, in total 1 rows in raw dataset ds need to be filled
        in domain.
113231  2022-12-01 17:43:36.0555 Info|Finished filling data of domain End of study page-End of Study Page of this subject.
113232  2022-12-01 17:43:36.0555 Info|Start filling data of domain Pregnancy report-Pregnancy Reporting/Follow-Up of this subject from raw dataset
        pfu...
113233  2022-12-01 17:43:36.0555 Warn|Trying to fill data of domain Pregnancy report-Pregnancy Reporting/Follow-Up for this subject, but the input
        parameter does not have the corresponding dataset pfu
113234  2022-12-01 17:43:36.0555 Info|Empty value is set for domain Pregnancy report-Pregnancy Reporting/Follow-Up.
113235  2022-12-01 17:43:36.0555 Info|Start filling data of domain C29D1-Date of Visit of this subject from raw dataset sv...
113236  2022-12-01 17:43:36.0555 Info|No data exist for the domain C29D1-Date of Visit.
113237  2022-12-01 17:43:36.0555 Info|Empty value is set for domain C29D1-Date of Visit.
```

Figure 9 Snapshot of Casebook Generator Log

CONCLUSION

Our approach of generating translated casebook involves two steps. Firstly, we translate the raw data and relevant EDC metadata file. Then, we developed an automation tool to generate the casebook using the translated datasets and files. This approach can be applied not only to casebook generation but also to the translation of any EDC-related documents, for example, eCRF, which is just a blank template of casebook requires no input of raw datasets. For optimal results, it is recommended to thoroughly explore your EDC system and develop your program accordingly. While developing the SAS macros and the tool may be time-consuming, once they are created, they will significantly improve efficiency and minimize the financial costs associated with translating and generating casebooks.

REFERENCES

None

CONTACT INFORMATION

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

Yiding Qu
Jiangsu Hengrui Pharmaceuticals Company Ltd.
+86-19962179092
yiding.qu@hengrui.com