

Streamlining Clinical Data Translation from Chinese to English: Ensuring Confidentiality and Efficiency in Pharmaceutical Partnerships

Zoe Chen, MSD China, Shanghai
Jin Wang, MSD China, Shanghai

ABSTRACT

In recent years, collaboration between global and Chinese pharmaceutical companies has intensified. This led to a heightened demand for specialized translation services, particularly for clinical trial data. As statistical programming professionals, we handle a significant volume of translation work that goes beyond simple text translation. To meet this demand, we have streamlined our processes and resources to ensure the accurate translation of clinical data from Chinese to English, which is essential for supporting critical processes like the study Analysis and Reporting (A&R) and the Investigator's Brochure (IB).

Our workflow involves extracting data text elements, translating them, and merging them back, followed by final validations to ensure compliance with data confidentiality requirements. We utilize both SAS and R in this process. Additionally, we have developed a Shiny app to automate parts of the translation process and assist with validation, which enhances efficiency and maintains high-quality standards in our translations.

INTRODUCTION

BACKGROUND

In recent years, the pharmaceutical industry has witnessed a significant increase in collaboration between global and Chinese companies. This trend has been driven by the growing importance of the Chinese market and the need for innovative therapies that cater to diverse populations. As these collaborations expand, the demand for accurate and specialized translation services has intensified, particularly concerning clinical trial data. Clinical trials are critical for the development of new drugs, and the integrity of the data collected during these trials is paramount. Accurate translation of clinical data from Chinese to English is essential for ensuring that A&R meet the required standards and that findings are communicated effectively to stakeholders worldwide.

PURPOSE OF THE PAPER

This paper aims to address the challenges associated with the translation of clinical trial data, emphasizing the role of statistical programming professionals in this process. Unlike simple text translation, the translation of clinical data involves a complex workflow that requires a deep understanding of both the source and target languages, as well as the specific terminology used in clinical research. To meet the heightened demand for accurate translations, we have streamlined our processes and resources, ensuring compliance with data security and confidentiality requirements. This paper will outline our workflow, the tools and technologies we employ, and the impact of these innovations on the efficiency and quality of our translation services.

CHALLENGES IN TRANSLATION OF CLINICAL TRIAL DATA

VOLUME AND COMPLEXITY OF DATA

Translating clinical trial data presents unique challenges due to the volume and complexity of the information involved. Clinical trials generate extensive datasets, including patient information, treatment protocols, adverse event reports, and statistical analyses. Each element contains specialized terminology that must be accurately translated to preserve data integrity. Additionally, consistency of translations across multiple documents, for instance, those parts of A&R and the IB, is crucial. Inconsistent

translations can lead to misunderstandings and regulatory issues, jeopardizing the success of a clinical trial.

COMPLIANCE AND CONFIDENTIALITY

In the pharmaceutical industry, compliance with regulatory standards is non-negotiable. Translating clinical trial data involves navigating a landscape of stringent regulations that govern data security and confidentiality. Translators must ensure that sensitive information is protected throughout the translation process, which requires robust data management practices. Failure to comply with these regulations can result in severe penalties, including fines and damage to the company's reputation.

RESOURCE LIMITATIONS

Finally, resource limitations can pose a challenge in the translation of clinical trial data. Many pharmaceutical companies may not have dedicated teams of translators with the necessary expertise in both clinical research and statistical programming. This can lead to bottlenecks in the translation process, particularly when deadlines are tight. To address this issue, companies must invest in training and development for their staff or consider outsourcing translation tasks to specialized service providers. However, outsourcing can introduce additional challenges related to quality control and consistency, making it essential to establish clear guidelines and communication channels.

WORKFLOW FOR TRANSLATION

TRANSLATION WORKFLOW

A well-defined workflow is crucial for ensuring the efficiency and accuracy of clinical data translation from Chinese to English.

The translation process consists of three primary steps, as illustrated in the workflow on Figure 1 hereafter: extraction of the translation sheet, translation of terms, and merging the translated terms back. This approach effectively transforms dataset translation into document translation, thereby significantly reducing challenges and enhancing data confidentiality. Detailed descriptions of each step are provided in the table below.

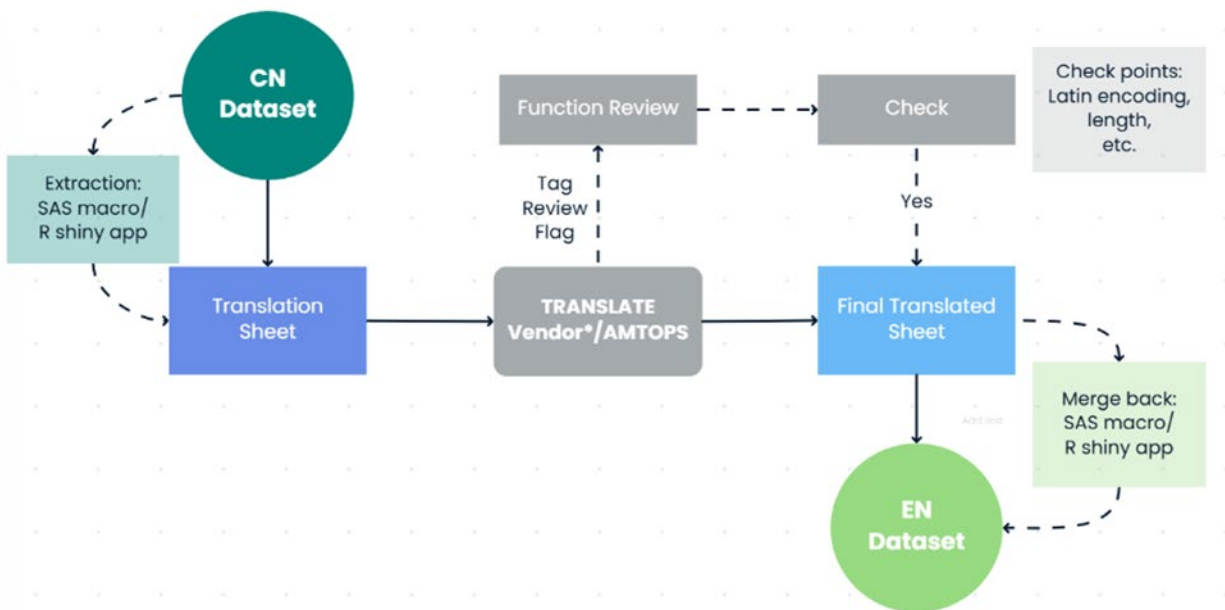


Figure 1. Chinese to English Dataset Translation Workflow

STEP	DETAILS	METHODS/TOOLS
Extraction	- Extract all the elements to translate including dataset labels, variable labels, and unique terms across the datasets - Save in a translation sheet (Figure 2)	- SAS macro: %data0trans0excel - R Shiny app
Translation	- Vendor or internal machine translation solution translate the excel sheet - Each function of study team review the translated sheet based on review flag - Programmer performs the check	- Vendor - Machine Translation Solution
Reconciliation	- Merge terms back to get the final English datasets - Programmer performs the check and comparison before delivery	- SAS macro: %merge0terms0back - R Shiny app

A	B	C	D	E	F	G
Study	DataSet	Variable	Values_CN	Values_EN	Coded_Va	Code
ABC1234-123	ae	AEACN	剂量不变			
ABC1234-123	ae	AEACN	不适用			
ABC1234-123	ae	AEACN	暂停用药			
ABC1234-123	ae	AEACNOTH	无			
ABC1234-123	ae	AEACNOTH	合并用药			
ABC1234-123	ae	AEACNOTH	合并非药物治疗			
ABC1234-123	ae	AEACNOTH	合并用药\$合并非药物治疗			
ABC1234-123	ae	AEBODSYS	各类检查		AEBDSYCD	10022891
ABC1234-123	ae	AEBODSYS	代谢及营养类疾病		AEBDSYCD	10027433
ABC1234-123	ae	AEBODSYS	呼吸系统、胸及纵隔疾病		AEBDSYCD	10038738
ABC1234-123	ae	AEBODSYS	胃肠系统疾病		AEBDSYCD	10017947
ABC1234-123	ae	AEBODSYS	皮肤及皮下组织类疾病		AEBDSYCD	10040785
ABC1234-123	ae	AEBODSYS	血液及淋巴系统疾病		AEBDSYCD	10005329
ABC1234-123	ae	AEBODSYS	肾脏及泌尿系统疾病		AEBDSYCD	10038359
ABC1234-123	ae	AEBODSYS	全身性疾病及给药部位各种反应		AEBDSYCD	10018065
ABC1234-123	ae	AEBODSYS	肝胆系统疾病		AEBDSYCD	10019805
ABC1234-123	ae	AEBODSYS	感染及感染类疾病		AEBDSYCD	10021881

Figure 2. Example of Extracted Translation Sheet

REVIEW AND VALIDATION

High-quality dataset translation is essential to support the A&R work. Consequently, thorough review and validation processes are critical to ensure the quality. When vendor collaboration is necessary, an issue tracking log will be shared to document all the translation issues. After extracted translation sheets are available, statisticians and programmers utilize aligned standards to filter and categorize records, assigning them to the appropriate functions. Additional columns are added in the sheets for reviewer function, flag, and comments. Programmers will conduct checks to the translated sheets. Check points include length, encoding, especially if there are characters not following Latin encoding e.g. special characters, Roman numbers, etc. The extracted translation sheet and the final English datasets can also be validated using an R Shiny app, which will be further discussed in the next section.

TOOLS AND TECHNOLOGIES

The integration of advanced tools and technologies can significantly enhance the translation process. This section will discuss the tools currently implemented in the workflow and GenAI we plan to endorse in the next step.

SAS

We developed a SAS macro to extract translation terms from all datasets into a single Excel file. The translation terms are categorized into three levels: 1) Dataset Labels, 2) Variable Labels, and 3) Variable Values (Figure 2). These unique terms can be translated either by external vendors or through internal translation tools.

Upon receiving the translated Excel file, we utilize SAS code to perform a thorough validation; for instance, to check that the translated observations are consistent with the originals, to verify that the translated columns contain only Latin-encoded characters and to confirm that the maximum length for each translated term does not exceed 200 characters.

Once the translated values pass the SAS checks and receive approval from relevant functions, they are merged back into the datasets, replacing the original Chinese terms using the SAS macro.

We chose SAS to develop the transferred English dataset for several key reasons. It ensures that the data format remains consistent with the original SAS format, simplifying the verification of encoding and length constraints, which in turn enhances translation accuracy. Additionally, SAS efficiently generates datasets from the validated translations, facilitating subsequent analysis and reporting.

R SHINY APP

We developed an R Shiny app to enhance the efficiency and accuracy of the dataset translation process. This shiny app facilitates the generation of translation sheets, compares with the sheets produced by SAS macro, merges the term back based on the translated sheets, and verifies the final translated English datasets (Figure 3). By doing so, it validates the two most critical steps in dataset translation: term extraction and data merging back, allowing for a comparison of results from the two programming languages, SAS and R. Additionally, the tool features a user-friendly interface, enabling users to easily input information such as protocol number, data type, and CPI path to execute the processes, thereby eliminating the need to read and update the underlying programs.

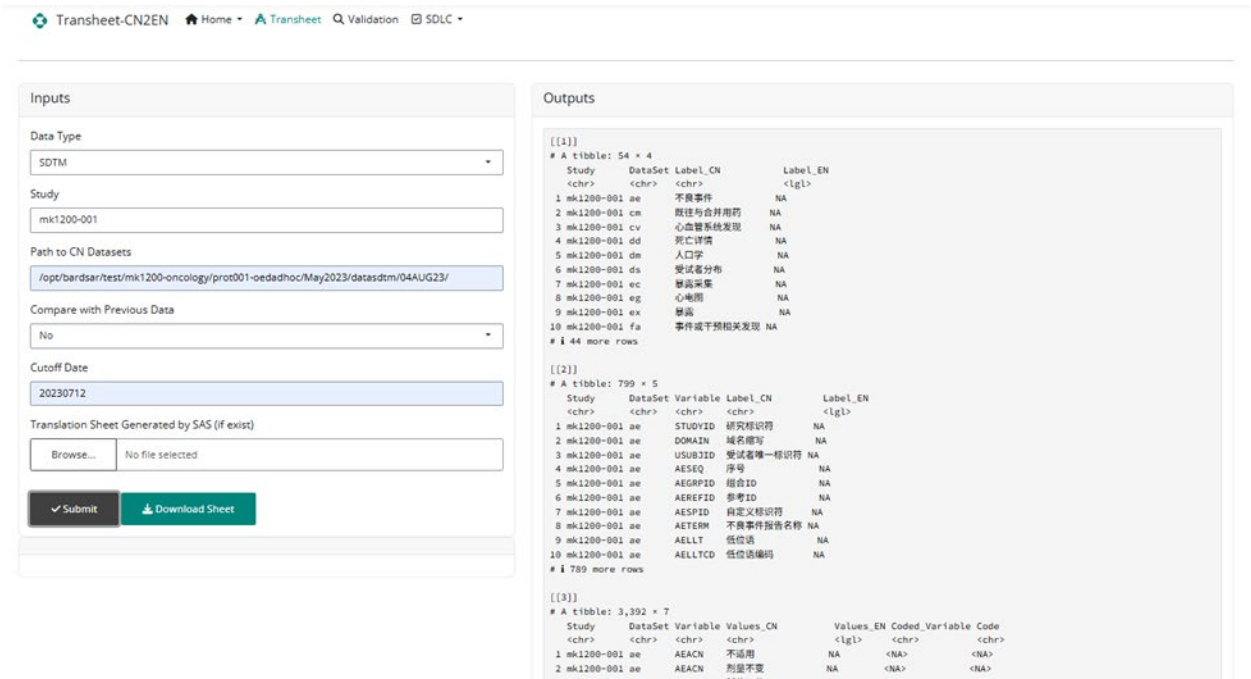


Figure 3. A Screenshot of R Shiny App

TRANSLATION TOOLS

A platform designed to host the medical machine translation solution in MSD managed environment is available for internal translation purposes. It provides secure and effective machine translation technology to automate the translation process of eSUB packages from English to Chinese and Chinese to English for BARDS to meet regulatory and clinical development requirements. The features include document translation, translation memory and term base management, etc.

Generative AI has revolutionized the field of translation by offering a level of contextual understanding and fluency that traditional translation software often lacks. Next step, the team plans to leverage GenAI to enhance the translation quality, particularly for free text.

RESULTS AND IMPACT

EFFICIENCY IMPROVEMENTS

The implementation of our streamlined translation workflow has led to significant improvements in efficiency. By automating key components of the translation process with our Shiny app, we have reduced the time required for data extraction, translation, and validation from 4 weeks to 2 weeks - an approximate 50% reduction. This reduction in turnaround time has enabled our team to handle a higher volume of clinical trial data without compromising quality. For instance, in a recent project involving a multi-center clinical trial, we successfully translated over 100 datasets within a compressed timeline.

QUALITY ASSURANCE

In addition to efficiency gains, our enhanced processes have also improved the quality of our translations. By implementing rigorous validation steps and utilizing both SAS and R, we have minimized errors—such as verifying the length of split variables (e.g., CMTERM, AETERM, XXDES) and checking for special characters in advance—and ensured that all translated materials meet the highest standards of accuracy. Furthermore, feedback from study team and stakeholder has been overwhelmingly positive, with reviewers noting the clarity and precision of the translated documents, which has facilitated smoother review processes.

CONCLUSION

In conclusion, the collaboration between global and Chinese pharmaceutical companies has created a pressing need for specialized translation services in clinical trial data. Our research highlights the multifaceted challenges associated with this task, including the volume and complexity of data, compliance with regulatory standards, and resource limitations. By streamlining our translation processes and leveraging technology, we have successfully addressed these challenges, resulting in significant improvements in both efficiency and quality.

The development of our Shiny app has been instrumental in automating parts of the translation workflow, allowing us to maintain high standards while meeting the increasing demands of the industry. As the pharmaceutical landscape continues to evolve, the importance of accurate and timely translation services will only grow. Our findings emphasize the need for ongoing investment in training, technology, and quality assurance to ensure that clinical trial data is communicated effectively across linguistic.

Future research should explore further innovations in translation technology and methodologies, as well as the potential for integrating artificial intelligence and machine learning to enhance the translation process. By continuing to adapt and improve our practices, we can better support the critical work of clinical research and contribute to the advancement of global healthcare.

CONTACT INFORMATION

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

Zoe Chen
MSD China

zhang.yu.chen1@merck.com

Jin Wang
MSD China

jin.wang9@merck.com