PharmaSUG China 2022 - Paper SR-118

Dictionary Preparation for RTF Table Translation

Junfan Zhang, MSD China, Beijing;
Luwei Pang, MSD China, Shanghai;

ABSTRACT

Today, global-multi-center enrollment is a norm in clinical trials to evaluate the safety and efficacy of new treatments on human health outcomes. For submission of drugs and biological products in non-English speaking countries, most local regulatory agencies mandate the result to be translated into local language. A global automated translation solution is in demand to ensure high quality, full traceability and re-producibility when the trial result tables in English are translated into other languages. An essential component of this solution is a translation dictionary. Creation of a translation dictionary from scratch is a huge effort. Such knowledge however is readily available in prior submissions from the English reports and their translated versions. This paper describes automated preparation of an English-Chinese dictionary as input for an automated translation application based on English and its translated Chinese clinical study reports (CSR). The dictionary creation requires accurate identification, as well as precise extraction and mapping of table titles, body contents and footnotes from both documents. Details are provided to illustrate why and how this can be achieved using VBA and SAS implementations.

INTRODUCTION

With the increasing number of clinical trials conducted across multiple countries and the consummation of submission guidelines in different countries, the need to obtain translated outcomes is becoming more important. Among all the required documents, the translated RTF tables are especially significant.

Currently, one of the most efficient methods to obtain the translated RTF tables is to generate them in English and then collaborate with a vendor to perform the translation. This approach needs time and precision. Since consistency is not guaranteed, the translated tables require a careful review.

An automated translation tool is in demand to provide precise, traceable, and re-producible translated tables. In this process, the preparation of a translation dictionary is an essential part of the work. Because of the similarity for the outcome tables in different trials, most information needed in the translation dictionary can be found in prior submitted the English CSR files and their translated versions.

This paper introduces the process on how the dictionary is obtained from the English and translated Chinese reports, and how the dictionary is modified to the final applicable and concise translation dictionary. All the dictionary preparation process are achieved with VBA and SAS implementations.

DICTIONARY PREPARATION PROCESS

Published English CSR files, relative translated Chinese CSR files, and English result tables generated by programmers, are the materials required to obtain the dictionary.

In Figure 1, the English and Chinese CSR files are used to create Chinese benchmark tables. After matching the Chinese benchmark tables with the English result tables, the dictionary terms are extracted and modified as needed. Results from previous steps, produce the dictionary version 1 which contains terms that have multiple translations. This requires output to be shared with medical writers who have to select the correct translated term. Then the dictionary is used to translate the English result tables. After translation, the translated tables are compared with the Chinese benchmark tables, differences or issues found through the comparison process are used to update the dictionary.

Details about creating and modifying the dictionary will be explained in the following sections.

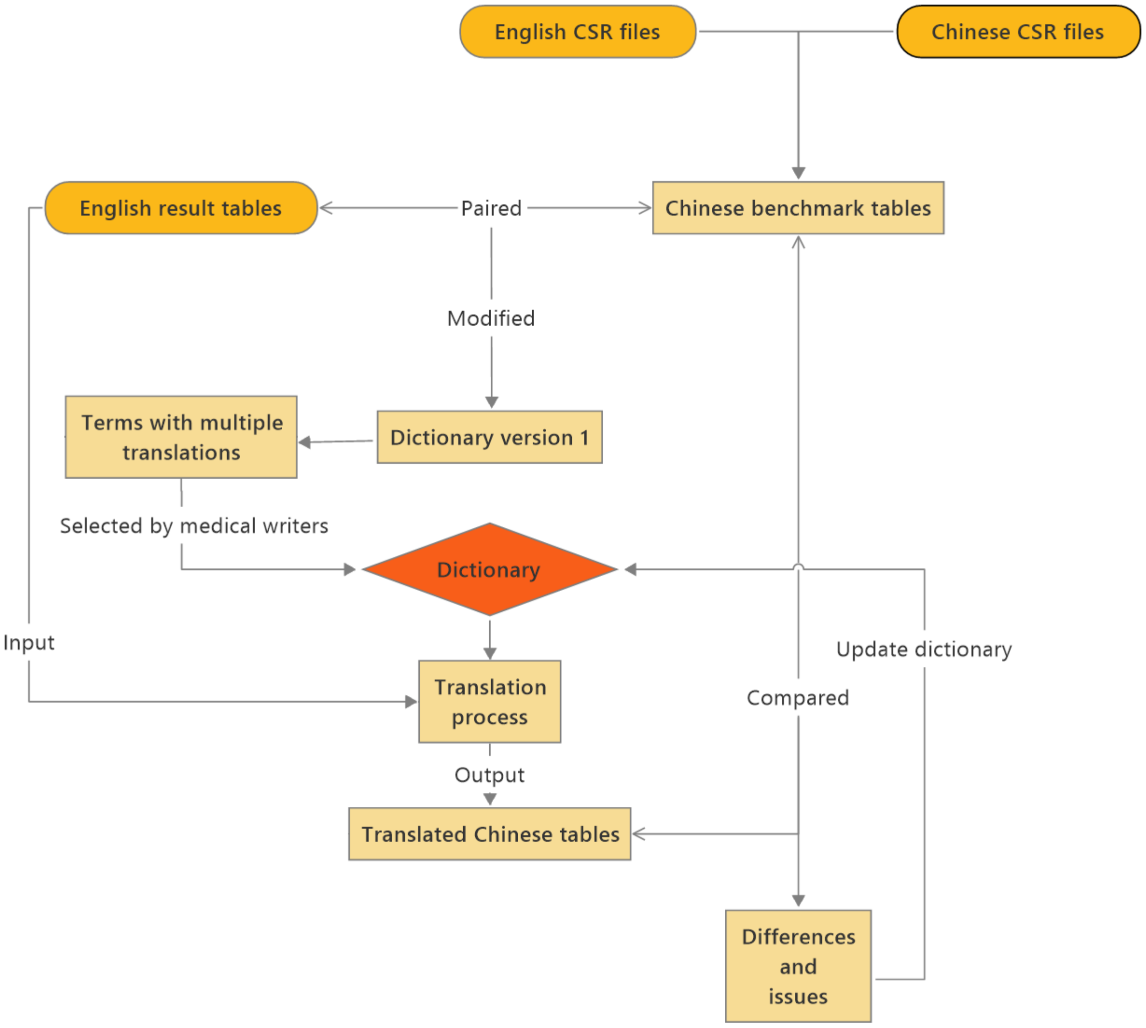


Figure 1. Dictionary preparation flow

DICTIONARY CREATION

As illustrated in the process section, English CSR files, their translated Chinese versions, and English result tables are implemented as inputs for dictionary creation. Since a typical CSR file contains hundreds of analysis result tables, manually extraction and mapping of table contents are time-consuming and error prone. The VBA and SAS implementations greatly enhance the working efficiency and ensure the accurate outputs.

The VBA application facilitates the accurate identification and precise extraction of individual result tables from the Chinese CSR files while the SAS tool enables the automatic creation of Chinese benchmark tables. Additionally, the VBA application promotes the extraction and mapping of table titles, body contents and footnotes from the paired Chinese benchmark tables and English result tables, which achieves the creation of draft dictionary. In the next step, the VBA application provides the flexibility of organizing and formatting the dictionary by removing the cells, which are not in the translation scope, removing the repeated records and separating the multiple-line headers and footnotes.

STEP 1: IDENTIFY AND EXTRACT STUDY RESULT TABLES

Given the standardized result table format required by the CSR guidance (Figure 2), the VBA Select.Find application facilitates the automatic and accurate identification of each result table from the Chinese CSR files. This step matches the content using the customized wildcards selection criteria. The wildcards criteria promotes the precise positioning of contents, which start with the Chinese character “ChrW(34920)” and end with specific table source footnotes. Each result table located in the Chinese CSR files is automatically copied and saved as individual RTF file with the table section number as the filename.

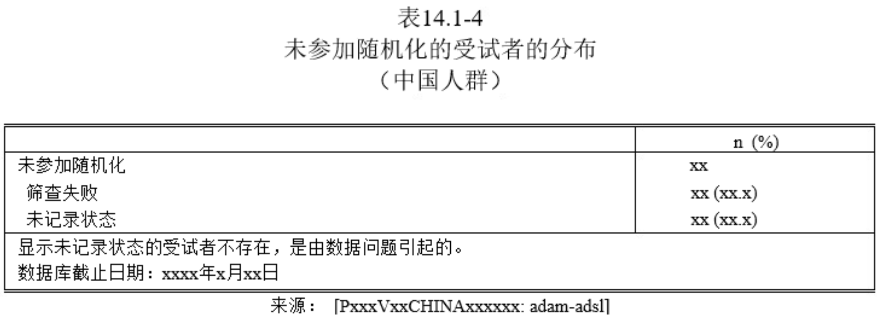


表14.1-4
未参加随机化的受试者的分布
（中国人群）

	n (%)
未参加随机化	xx
筛查失败	xx (xx.x)
未记录状态	xx (xx.x)
显示未记录状态的受试者不存在，是由数据问题引起的。 数据库截止日期：xxxx年x月xx日	

来源： [PxxxVxxCHINAxxxxxx: adam-adsl]

Figure 2. Study result table from Chinese CSR files

The VBA code shown below demonstrates the process to locate, copy and save each study result tables from Chinese CSR files.

```
Do
  With Selection.Find
       .Text = ChrW(34920) & "[ \. 0-9-]@" & "[^13]" & "*\[P*\]"
       .MatchWildcards = True
       .Forward = True
       .Execute
   End With

   If Selection.Find.Found Then
          Selection.Copy
          ……
          Set newdoc = Documents.Add
          With newdoc
              .Content.Paste
              .SaveAs2 FileName:="……", FileFormat:=wdFormatRTF
              .Close
          End With
   End If
Loop
```

STEP 2: CREATE CHINESE BENCHMARK TABLES

The SAS application performs the extraction of table filenames and table titles from English result tables and the similar extraction of table section numbers and titles from English CSR files. For each study result table, the table section number is successfully linked with the original filename of the English result tables through the matching of corresponding English titles. This linking information supports the creation of the Chinese benchmark tables. The extracted Chinese result tables, as mentioned in step 1, are renamed with the original filenames of the English result tables via the table section number and saved as benchmark tables.

There are advantages for creating the Chinese benchmark tables.

(1) The benchmark tables, along with the English result tables are implemented as inputs to create dictionary, which enhance the content mapping efficiency and precision.

(2) The benchmark tables facilitate the comparison with translated RTF files generated by the automatic translation tool.

(3) The benchmark tables promote the validation and improvement of dictionary and automatic translation tool.

STEP 3: CREATE DICTIONARY

When both the Chinese benchmark tables and the English result tables are ready, the VBA application provides the flexibility to extract and map the paired contents including table titles, body cells and footnotes (Table 1). The outputs are automatically saved in the excel worksheets.

	n (%)
未参加随机化	xx
筛查失败	xx (xx.x)
未记录状态	xx (xx.x)
显示为记录状态的受试者不存在，是由数据引起的。 数据库截至日期：xxxx 年 xx 月 xx 日	
	n (%)
Not Randomized	xx
Screen Failure	xx (xx.x)
Status Not Recorded	xx (xx.x)
The subject showing status not recorded does not exist and was caused by a data issue. Database Cutoff Date: xxxxxxxxx	

Table 1. Extraction output of table contents from Chinese benchmark tables and English result tables

The VBA code shown below demonstrates the process to extract table contents:

```
For t_index = 1 To table_n
        If t_index = 1 Then start_r = 0
        Else start_r = xlwbk.Worksheets(1).UsedRange.Rows.Count
        End If
        col_n = doc.Tables(t_index).Columns.Count
        row_n = doc.Tables(t_index).Rows.Count
        For Col = 1 To col_n
            strcelltext = ""
            For Row = 1 To row_n + 1
                strcelltext = ""
                strcelltext = doc.Tables(t_index).Cell(Row, Col).Range.Text
                Clean = Trim(strcelltext)
                posFromEnd = InStrRev(Clean, vbCr + Chr(7))
                If (posFromEnd > 0) Then
                    Clean = Trim(Left(Clean, posFromEnd - 1))
                End If
                xlwbk.Worksheets(1).Cells(start_r + Row, Col) = Clean
            Next
          Next
Next
```

In addition, the VBA application promotes the organization and formatting of the draft dictionary, including:

(1) Removing the cells with the specific number patterns, which are not in the scope of CSR table translation.

The regular expression function in VBA application allows the identification of customized number pattern, the example codes shown below:

```
regex.Pattern = "^[\s]*[\(]?[-]?[0-9\.]+(,[ 0-9\.-]+)?[\)]?([ ]?[\(][-]?[0-9\.]+(,[ 0-9\.-]+)?[\)])?$"
```

(2) Removing the repeated translation records.

(3) Separating the multiple-line footnotes into cells.

(4) Highlighting the different translation results for the same English contents.

The different Chinese translated results for the same English contents are highlighted and further confirmed with the Medical Writers to keep the accurate one.

The formatted dictionary (Table 2) is implemented as the fundamental dictionary for the next step.

Not Randomized	未参加随机化
Screen Failure	筛查失败
Status Not Recorded	未记录状态
The subject showing status not recorded does not exist and was caused by a data issue.	显示为记录状态的受试者不存在，是由数据引起的。
Database Cutoff Date: xxxxxxxxx	数据库截至日期：xxxx 年 xx 月 xx 日

Table 2. Formatted dictionary for English result table translation

DICTIONARY MODIFICATION

The fundamental dictionary obtained from previous steps, contains most information required for translation. However, a few steps of modification, including the manipulation of special characters; the application of wildcard; and the processing of multiple lines header, are required to guarantee the applicability and conciseness of the dictionary. All the modification steps are processed with the SAS implementation.

SPECIAL CHARACTERS MANIPULATION

There are two types of special characters that need to be handled: table format related characters and meaningful characters like superscript symbols. The table format related characters need to be removed from the dictionary terms and the meaningful characters need to be transformed into appropriate format.

Remove Table Format Related Characters

Since the fundamental dictionary is obtained directly from the RTF tables with VBA, some characters related to table format are included into the table as well, e.g., “horizontal tab” and “bell”. These characters lead to mismatches when the translation is processed. Therefore, the removal of them is required for the dictionary terms.

Format related characters are hard to type and recognize in SAS. The PUT function is used to identify these format related characters. Here is an example to identify the “horizontal tab” in the English term.

Table 3 displays an example of English term with “horizontal tab” in the fundamental dictionary. The leading blank spaces before 31 is a “horizontal tab”, therefore it cannot be removed with the SAS space-handling functions like strip and trim.

English	**Chinese**
31 to 76	31 至 76

Table 3. Sample of fundamental dictionary term with format related characters

SAS code used to identify the special character is:

```
data _null_;
    set ds1;
    put EN;
    put EN hex16.;
run;
```

The output of the mentioned code for the displayed term is:

```
31 to 76
09333120746F2037
```

According to the ASCII code list, 33 matched to number 3, therefore, 09 is the hexadecimal expression for “horizontal tab” as ASCII code. Then the “horizontal tab” can be removed with SAS compress function:

```
data ds2;
    set ds1;
    english=compress(EN,"09"x);
run;
```

Transform Meaningful Characters

When capturing the information through VBA, it removes the format for corner markers like superscripts, and includes Chinese characters, which are accidentally added into the RTF tables. These characters need to be transformed in the dictionary to guarantee the correct translation.

Format Corner Markers

Table 4 illustrates the reason to add the appropriate format to the corner markers. If the fundamental dictionary term is used directly, the corner symbol “†” will no longer be a superscript resulting in an inconsistency between the English result table and the translated table.

Original English Term	**Fundamental Dictionary**	**Fundamental Translation**	**Correct Translation**
Response Duration† (months)	Response Duration† (months)	缓解持续时间† (月)	缓解持续时间† (月)

Table 4. Example of corner markers in dictionary

Table 5 provides the example of the transformed terms for 3 commonly used superscripts. The format information “\super” is added to make the translated table consistent with the original English table. The symbols are transformed to the ASCII code expressions to match with the RTF table terms that are read into SAS during the translation process.

Original English Term	**Transformed Term**
†	{\super\’86 }
‡	{\super\’87 }
§	{\super\’a7 }

Table 5. Transformed superscripts

Chinese Characters in English Tables

During the translation process, the input table should only have the English letters and punctuation marks used in English environment. Therefore, Chinese characters that are accidentally included in the English tables are unable to be parsed in the English environment resulting in translation issues.

Chinese characters are transformed to the relative ASCII code expressions to allow it in the English environment and match the SAS inputted English terms. Table 6 provides several examples of Chinese characters that might be included in English tables and their ASCII code expressions for RTF tables.

Original Term	Transformed Term
’	{\’92}
“	{\’93}
”	{\’94}
-	{\’96}

Table 6. Transformed Chinese characters

APPLICATION OF WILDCARDS

The information from tables of different studies is similar, however, the words and numbers occurred in different studies are not exactly the same. Therefore, the use of wildcards for some special terms enables the dictionary to become more universal and makes the dictionary more succinct.

Table 7 illustrates several wildcards used in the dictionary. The terms that have only one specified wildcard required situations are transformed into wildcard versions. For example, if the English term include 2 timepoints, “at week 2 day 1”, no wildcard will be included into this term. However, if the English term is “week 2”, or “day 1”, both of them will be transformed into wildcard versions. SAS prxmatch and prxexchange functions are used to create the wildcard terms.

English	Chinese
@N@ Days	@N@天
@N@ Hours	@N@小时
@N@ Minutes	@N@分钟
@N@ Months	@N@个月
@N@ Weeks	@N@周
@N@ Years	@N@年
Week @N@	第@N@周
At Week @N@	第@N@周
Age=@N@ Years	年龄=@N@岁
Beyond @N@ months	@N@个月以后
Crossover Treatment Cycle @N@	交叉治疗周期@N@
Site Number=@N@	研究中心编号 = @N@
Source: @N@	来源：@N@
Subject ID=@N@	受试者编号 = @N@
Total (N=@N@)	总计(N = @N@)
Treatment Cycle @N@	治疗周期@N@
Trial Number=@N@	试验编号 = @N@
Unique Subject ID=@N@	唯一受试者 ID=@N@

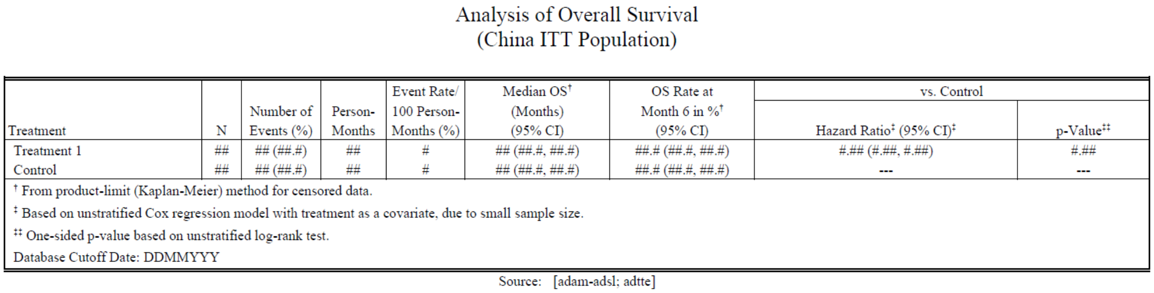
Table 7. Examples of wildcards used in the dictionary

Here is the example of SAS code used to create one type of wildcard term – timepoint:

```
data ds3; set ds2;
    length English_new Chinese_new $5000.;

if prxmatch('m/(week|month|day|hour|year|minute)/i',english) then do;

    if not prxmatch('m/(week|month|day|hour|year|minute)/i',
    prxchange('s/(week|month|day|hour|year|minute)//i',1,english))
    then do;
        if prxmatch('m/(\d+\s*)(week|month|day|hour|year|minute)(\(?s\)?)?/
        i',english)
        then do;
          English_new=prxchange('s/(\d+\.\d+|\d+)(\s*)(week|month|day|hour|
          year|minute)(\(?s\)?)?/\@N\@\2\3\4/i',1,english);
          Chinese_new=prxchange('s/(\d+\.\d+|\d+)(\s*)(周|个月|月|天|小时|
          年|分钟)(\(?s\)?)?/\@N\@\2\3\4/i',1,chinese);
        end;
      else if prxmatch('m/(week|month|day|hour|year|minute)(\(?s\)?)?(\s*)
      (\d+\.\d+|\d+)/i',english)
      then do;
          English_new=prxchange('s/(week|month|day|hour|year|minute)(\(?s\)
          ?)?(\s*)(\d+\.\d+|\d+)/\1\2\3\@N\@/i',1,english);
          Chinese_new=prxchange('s/(\d+\.\d+|\d+)(\s*)(周|个月|月|天|小时|
          年|分钟)(\(?s\)?)?/\@N\@\2\3\4/i',1,chinese);
      end;
    end;
end;

run;
```

MULTIPLE LINES HEADER

During the translation process, the information in the tables is translated cell by cell. However, according to the table structure, a few table headers are separated into different cells, which can also be recognized as multiple lines headers. Figure 3 and Figure 4 provide an example of English and relative Chinese tables with multiple line headers.



Analysis of Overall Survival
(China ITT Population)

Treatment	N	Number of Events (%)	Person-Months	Event Rate/ 100 Person-Months (%)	Median OS† (Months) (95% CI)	OS Rate at Month 6 in %† (95% CI)	vs. Control	
							Hazard Ratio‡ (95% CI)‡	p-Value‡‡
Treatment 1	##	## (##.#)	##	#	## (##.#, ##.#)	##.# (##.#, ##.#)	#.## (#.##, #.##)	#.##
Control	##	## (##.#)	##	#	## (##.#, ##.#)	##.# (##.#, ##.#)	---	---

† From product-limit (Kaplan-Meier) method for censored data.

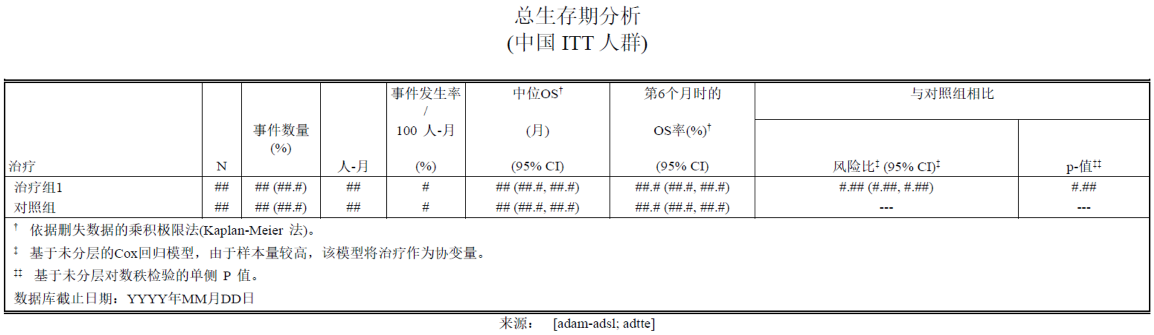
‡ Based on unstratified Cox regression model with treatment as a covariate, due to small sample size.

‡‡ One-sided p-value based on unstratified log-rank test.

Database Cutoff Date: DDMMYYY

Source: [adam-adsl; adtte]

Figure 3. English table with multiple lines headers



总生存期分析
(中国 ITT 人群)

治疗	N	事件数量 (%)	人-月	事件发生率 / 100 人-月 (%)	中位OS† (月) (95% CI)	第6个月时的 OS率(%)† (95% CI)	与对照组相比	
							风险比‡ (95% CI)‡	p-值‡‡
治疗组1	##	## (##.#)	##	#	## (##.#, ##.#)	##.# (##.#, ##.#)	#.## (#.##, #.##)	#.##
对照组	##	## (##.#)	##	#	## (##.#, ##.#)	##.# (##.#, ##.#)	---	---

† 依据删失数据的乘积极限法(Kaplan-Meier 法)。

‡ 基于未分层的Cox回归模型，由于样本量较高，该模型将治疗作为协变量。

‡‡ 基于未分层对数秩检验的单侧 P 值。

数据库截止日期：YYYY年MM月DD日

来源： [adam-adsl; adtte]

Figure 4. Chinese table with multiple lines headers

Due to the different language structures between English and Chinese, as displayed in Table 8, the English and Chinese multiple lines headers are not able to be mapped line by line. These headers need to be merged and translated entirely. Text information for the header, as well as the location information, are required to be included in the dictionary. Table 9 provides the header terms for Figure 3 and Figure 4 in the dictionary.

English	Chinese
OS Rate at	第 6 个月时的
Month 6 in %†	OS 率(%)†
(95% CI)	(95% CI)

Table 8. Example of headers not able to be mapped line by line

<table>
<tr><th>English</th><th>Chinese</th></tr>
<tr><td>@_blank_ | _blank_ | Treatment@01@_blank_</td><td>_blank_</td></tr>
<tr><td>@_blank_ | _blank_ | Treatment@02@_blank_</td><td>_blank_</td></tr>
<tr><td>@_blank_ | _blank_ | Treatment@03@Treatment</td><td>治疗</td></tr>
<tr><td>@_blank_ | Number of | Events (%)@01@_blank_</td><td>_blank_</td></tr>
<tr><td>@_blank_ | Number of | Events (%)@02@Number of</td><td>事件数量 (%)</td></tr>
<tr><td>@_blank_ | Number of | Events (%)@03@Events (%)</td><td>_blank_</td></tr>
<tr><td>@_blank_ | Person- | Months@01@_blank_</td><td>_blank_</td></tr>
<tr><td>@_blank_ | Person- | Months@02@Person-</td><td>_blank_</td></tr>
<tr><td>@_blank_ | Person- | Months@03@Months</td><td>人-月</td></tr>
<tr><td>@Event Rate/ | 100 Person- | Months (%)@01@Event Rate/</td><td>事件发生率/</td></tr>
<tr><td>@Event Rate/ | 100 Person- | Months (%)@02@100 Person-</td><td>100 人-月</td></tr>
<tr><td>@Event Rate/ | 100 Person- | Months (%)@03@Months (%)</td><td>(%)</td></tr>
<tr><td>@Median OS{\super\'86} | (Months) | (95% CI)@01@Median OS{\super\'86}</td><td>中位 OS{\super\'86 }</td></tr>
<tr><td>@Median OS{\super@S@} | (Months) | (95% CI)@02@(Months)</td><td>(月)</td></tr>
<tr><td>@Median OS{\super@S@} | (Months) | (95% CI)@03@(95% CI)</td><td>(95% CI)</td></tr>
<tr><td>@OS Rate at | Month @N@ in %{\super@S@} | (95% CI)@01@OS Rate at</td><td>第 6 个月时的</td></tr>
<tr><td>@OS Rate at | Month 6 in %{\super\'86} | (95% CI)@02@Month 6 in %{\super\'86}</td><td>OS 率(%){\super\'86}</td></tr>
<tr><td>@OS Rate at | Month @N@ in %{\super@S@} | (95% CI)@03@(95% CI)</td><td>(95% CI)</td></tr>
<tr><td>vs. Control</td><td>与对照组相比</td></tr>
<tr><td>@_blank_ | Hazard Ratio{\super\'87} (95% CI){\super\'87}@01@_blank_</td><td>_blank_</td></tr>
<tr><td>@_blank_ | Hazard Ratio{\super\'87} (95% CI){\super\'87}@02@Hazard Ratio{\super\'87} (95% CI){\super\'87}</td><td>风险比{\super\'87} (95% CI){\super\'87}</td></tr>
<tr><td>@_blank_ | p-Value{\super@S@}@01@_blank_</td><td>_blank_</td></tr>
<tr><td>@_blank_ | p-Value{\super\'87\'87}@02@p-Value{\super\'87\'87}</td><td>p-值{\super\'87\'87}</td></tr>
</table>

Table 9. Dictionary terms for headers in figure 2 and figure 3

OTHER MODIFICATIONS AND VALIDATION

Due to the inconsistencies using manual translation, some English terms have multiple translations in the CSR files, thus leads to multiple records in the dictionary. These terms are flagged and sent to medical writers to decide the most appropriate translation term, only the selected best translation term is kept in the dictionary.

With dictionary obtained with the previous steps, the translation process is applied, and the translated Chinese RTF tables are obtained. These translated tables are compared with Chinese benchmark tables. The dictionary is modified again based on the differences and issues found through the comparison process.

CONCLUSION

With the well-prepared dictionary, the users are able to trace how the tables are translated and the translation outcomes are guaranteed to be consistent across all different studies. With more studies using the dictionary for translation, the dictionary will be periodically updated with the new commonly used terms.

ACKNOWLEDGMENTS

The authors would like to thank Qian Wang, Hui Liu, Danfeng Fu, and Biao Chen for their tremendous support, effective advice, and helpful input to this project.

We are also grateful for the support and review by Janet C. Low, Nicole Zhang, Wang Zhang, and our function head, Amy Gillespie.

CONTACT INFORMATION

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

Junfan Zhang
MSD China
jun.fan.zhang@merck.com

Luwei Pang
MSD China
lu.wei.pang@merck.com