基于Python的cSDRG自动化工具开发——实现临床研究文档生成革命

罗义/高级SAS程序员

2025-08-29

基于Python的cSDRG自动化工具开发

Presented by
罗义, 高级SAS程序员, 凯莱英临床(凯诺)

Meet the Speaker

罗义

Title: 高级SAS程序员

Organization: 凯莱英临床(凯诺)

从事统计编程工作4年，曾任职于普瑞盛医药、东阳光药，有丰富的工具开发经验。

2024年7月加入凯莱英临床（凯诺）统计编程部门，目前主要负责统计编程、工具开发等工作。

Disclaimer and Disclosures

- *The views and opinions expressed in this presentation are those of the author(s) and do not necessarily reflect the official policy or position of CDISC.*

Agenda

1. 关于cSDRG
2. 自动化cSDRG的实现
3. 关于引入AI的讨论

关于cSDRG

- cSDRG简介
- cSDRG内容来源拆解

cSDRG简介

目的

(cSDRG: Clinical Study Data Reviewer's Guide)

为FDA审阅者提供SDTM数据集的背景信息，辅助审评。整合分散在方案、临床研究报告等文档中的关键信息，确保数据可溯源性。

1. Clinical Study Data Reviewer's Guide Purpose

The Clinical Study Data Reviewer's Guide (cSDRG) provides FDA Reviewers with additional context for SDTM datasets received as part of a regulatory submission. The cSDRG is intended to describe SDTM data submitted for an individual study in the Module 5 clinical section of the eCTD. The cSDRG purposefully duplicates information found in other submission documentation (e.g. the protocol, clinical study report, define.xml, etc.) in order to provide FDA Reviewers with a single point of orientation to the SDTM datasets.

Package：https://advance.hub.phuse.global/wiki/x/SAeZAQ

cSDRG简介

结构

- 必含部分：介绍、方案描述、受试者数据描述、数据一致性总结
- 可选附录：
  - 附录I：入选/排除标准（当TI数据集无法完整描述时）。
  - 附录II：合规性问题详情（不建议使用，因对审评员价值有限）。

If the inclusion/exclusion criteria cannot be fully documented in the Trial Inclusion/Exclusion Criteria (TI) dataset due to SAS v5 limitations, the criteria can either be provided in Appendix I, or as a hyperlink to the full criteria in an annotated CRF. All significant conformance findings should be documented in the Data Conformance Summary; however, a detailed record-level description of conformance issues may be included in Appendix II. Sponsors are strongly discouraged from including Appendix II due to its limited usefulness for FDA Reviewers.

cSDRG内容来源拆解

介绍

- 目的/缩略语：可统一标准自定义

1.1 Purpose

This document provides context for tabulation datasets and terminology that benefit from additional explanation beyond the Data Definitions document (define.xml). In addition, this document provides a summary of SDTM conformance findings.

1.2 Acronyms

Acronym	Translation
aCRF	Annotated Case Report Form
eCRF	Electronic Case Report Form
eDT	Electronic Data Transfer (e.g. central lab data, ECG vendor data, PK data, etc.)

cSDRG内容来源拆解

介绍

- 研究数据标准和字典版本：define

1.3 Study Data Standards and Dictionary Inventory

Standard or Dictionary	Versions Used
SDTM	SDTM v1.2/SDTM IG v3.1.2 including SDTM Amendment 1
Controlled Terminology	2011-07-22 Added 'WASHOUT PERIOD 1' 'WASHOUT PERIOD 2' to EPOCH extensible codelist as the study design includes two washout periods.
Data Definitions	define.xml v1.0
Medications Dictionary	Proprietary medication dictionary
Medical Events Dictionary	MedDRA v14.1

cSDRG内容来源拆解

方案描述

- 方案编号和标题：方案、sdtm数据集
- 方案设计：方案（方案设计流程图）

2.1 Protocol Number and Title

Protocol Number:	SDRG-001A
Protocol Title:	A Phase II, Randomized Double-Blind Placebo-Controlled Dose Ranging Study to Evaluate SDRG-999 in Adults with Asthma
Protocol Versions:	SDRG-001A

cSDRG内容来源拆解

方案描述

- 试验设计数据集：方案、sdtm数据集

2.3 Trial Design Datasets

Are Trial Design datasets included in the submission? Yes

2.3.1. TI – Trial Inclusion/Exclusion Criteria

The trial inclusion/exclusion criteria are not fully described in the TI domain. Please refer to Appendix I for the full text of the criteria.

2.3.2. TS – Trial Summary

The TS domain includes the deprecated parameter Adverse Events Dictionary (AEDICT) to support internal processes.

cSDRG内容来源拆解

受试者数据描述

- 概述：sdtm数据集

3.1 Overview

Are the submitted data taken from an ongoing study? No

Were the SDTM datasets used as sources for the analysis datasets? Yes

Do the submission datasets include screen failures? No

Were any domains planned, but not submitted because no data were collected? No

Are the submitted data a subset of collected data? No

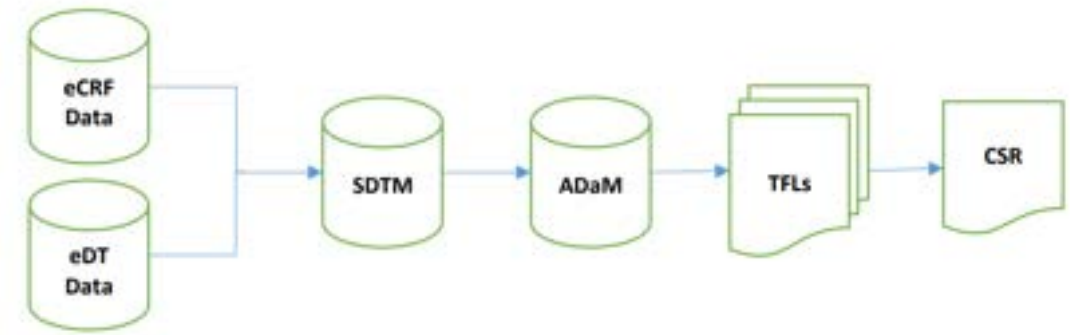
Is adjudication data present? No

cSDRG内容来源拆解

受试者数据描述

- 溯源流程图：根据cSDRG completion-guidelines-v1.4示例

Example:



End of Example

cSDRG内容来源拆解

受试者数据描述

- 注释病例报告表：aCRF

3.3 Annotated CRFs

Collected fields that have not been tabulated have been annotated as "Not Mapped". SDRG Inc. collects certain data elements to facilitate certain operational processes including data cleaning and dynamically creating additional forms in the electronic data capture system. All fields that have been annotated as "Not Mapped" meet these criteria.

Explanation of data fields [Not Submitted]

aCRF page Number(s)	Data Collection Field	Explanation of why [NOT SUBMITTED]
5	Were there any product complaints?	For internal use only.
30	PI Signature Date	Not needed for analysis.

cSDRG内容来源拆解

受试者数据描述

- SDTM受试者域： sdtm数据集、define

3.4 SDTM Subject Domains

Dataset – Dataset Label	Efficacy	Safety	Other	Custom	SUPP-	Related Using RELREC
AE – Adverse Events		X				ZA
CM – Concomitant Medications	X	X				
CO – Comments			X			
DD – Death Details			X			
DM – Demographics			X			
DS – Disposition			X			
EX – Exposure			X			

cSDRG内容来源拆解

数据一致性总结

- 一致性的输入：define

4.1 Conformance Inputs

Was a validator used to evaluate conformance? Yes

If yes, specify the versions of OpenCDISC and the OpenCDISC validation rules:

Pinnacle 21 Enterprise 3.4 (FDA), SDTM v3.1.3 rules

Were sponsor-defined validation rules used to evaluate conformance? Yes

If yes, describe any significant sponsor-defined validation rules:

SDRG Inc. executes a sponsor-defined conformance rule to confirm variable values that are 200 characters have not been truncated.

Were the SDTM datasets evaluated in relation to define.xml? Yes

Was define.xml evaluated? Yes

cSDRG内容来源拆解

数据一致性总结

- 问题总结： sdtm数据集（P21报告结果解释）

Example:

Dataset	Diagnostic Message	Severity	Count	Explanation
LB	Missing Units on Value	Error	22	**Not an error:** Lab results for pH and Specific Gravity have no units

End of Example

cSDRG内容来源拆解

数据一致性总结

- 补充的一致性明细：自定义，结合项目具体情况

4.3. Additional Conformance Details

This section documents summary findings from validation rules other than the ones previously reported, which in the sponsor's opinion merit explanation. Fill in the table in this section as you would the one in section 4.2. Leave columns blank where not applicable.

This section is not intended to contain the full validator details report. Sponsors are discouraged from submitting the full report, but if the sponsor considers it necessary, the full report may be submitted as cSDRG Appendix II.

If there are no additional conformance details to be documented, **do not delete the section**. Add verbiage such as "There are no additional details to be documented."

自动化cSDRG的实现

- 自动化数据提取系统架构
- 自动化文本表格构建算法
- 变量输出机制
- UI及使用介绍

自动化数据提取系统架构

导入数据主要有两种类型

- **xlsx**
  - P21导出的相关文件（包括单个define生成的P21报告）、所有sdtm数据集生成的P21报告、 define生成的spec
  - 项目sdtm spec文件

- **sas7bdat**
  - sdtm数据集
  - 项目sdtm spec导出的sas数据集

自动化数据提取系统架构

用Python提取xlsx数据

- 读取excel文件选择pandas库，因为后续处理均通过DataFrame类型进行。
- 读取过程中进行标准化处理，将所有DataFrame内元素处理为字符串，如果有需要数值计算的地方再分步处理，方便后续步骤中统一格式。
- 读取时一个sheet对应一个DataFrame对应一个变量，不统一读取多个sheet，方便后期溯源。

自动化数据提取系统架构

xlsx数据提取，示例：

```
#### 导入 define spec
df_definespec_define = pd.read_excel(path_definespec,sheet_name='Define',header=0)
df_definespec_valueLevel = pd.read_excel(path_definespec,sheet_name='ValueLevel',header=0)

df_definespec_datasets = pd.read_excel(path_definespec,sheet_name='Datasets',header=0)
# ------判断有哪些T系列数据集
ds_all_sdtm = list(df_definespec_datasets['Dataset'])
list_t_data = ['TA','TE','TD','TI','TM','TS','TV']
list_in_t_data = []
list_in_t_label = []
for i in list_t_data:
    for j in ds_all_sdtm:
        if i==j:
            list_in_t_data.append(i)
            list_in_t_label.append(list(df_definespec_datasets.loc[df_definespec_datasets['Dataset']==i]['Label'])[0])
dict_in_t = dict(zip(list_in_t_data,list_in_t_label))

df_definespec_dictionaries = pd.read_excel(path_definespec,sheet_name='Dictionaries',header=0)
meddra_version = 'MedDRA_'+list(df_definespec_dictionaries.loc[df_definespec_dictionaries['Dictionary']=='MedDRA']['Version'])[0]
whodra_version = 'WhoDrug_'+list(df_definespec_dictionaries.loc[df_definespec_dictionaries['Dictionary']=='WhoDrug']['Version'])[0]

#### 导入p21报告
# 单个define
df_p21report_summary = pd.read_excel(path_sdtm_p21_report_define,sheet_name='Validation Summary',header=0)
sdtmig_version = [i for i in list(df_p21report_summary['Pinnacle 21 Validator Report'].astype(str)) if i[:9]=='Standard:'][0][10:]
p21_version = [i for i in list(df_p21report_summary['Pinnacle 21 Validator Report'].astype(str)) if i[:17]=='Software Version:'][0][18:]
sdtmct_version = [i for i in list(df_p21report_summary['Pinnacle 21 Validator Report'].astype(str)) if i[:22]=='CDISC SDTM CT Version:'][0][23:]
# 数据集和define
df_p21report_issue = pd.read_excel(path_sdtm_p21_report_all,sheet_name='Issue Summary',header=3).astype(str).replace({'nan':''})
```

自动化数据提取系统架构

用Python提取sas7bdat数据

- 用pyreadstat库的read_sas7bdat函数可获取sas数据集元数据，主要可获取sas数据集编码类型。
- 用sas7bdat库的SAS7BDAT函数传入编码类型读取sas数据集，输出为DataFrame类型。
- 动态设置全局变量，将所有sdtm数据集分别存放至单独变量中。

自动化数据提取系统架构

sas7bdat数据提取，示例：

```
#### 读取sdtm sas数据集
#### path_sdtm_xx=r"D:\\ [illegible] \\xx.sas7bdat"
list_var_sdtm = ['=r'.join(item) for item in zip(list_sdtm_name, file_list_rtf_str)]
for i in range(len(list_var_sdtm)):
    exec(list_var_sdtm[i],globals())

#### 导入sas7dbat
#### ds_xx=SAS7BDAT(path_sdtm_xx,encoding="utf-8")
list_var_sdtm_sas = [i.replace('path_sdtm_','ds_')+'=SAS7BDAT('+i+',encoding="'+code_type+'")' for i in list_sdtm_name]
for i in range(len(list_var_sdtm_sas)):
    exec(list_var_sdtm_sas[i],globals())

## sas7dbat 转 dataframe
## df_xx=ds_xx.to_data_frame().astype(str)
list_var_sdtm_sas = [i+'='+i.replace('df_','ds_')+'.to_data_frame().astype(str)' for i in list_sdtm_df_name]
for i in range(len(list_var_sdtm_sas)):
    exec(list_var_sdtm_sas[i],globals())
```

自动化文本表格构建算法

所有输出变量的算法几乎均来自cSDRG completion-guidelines-v1.4定义，小部分需要结合具体项目情况。

文本变量：用于存储模板文件中所有可抓取的文本信息。
（例如，用变量at_studyid存储方案编号）

表格变量：用于存储模板文件中所有可抓取的表格信息。
（例如，用变量att_ta存储sdtm.ta）

自动化文本表格构建算法

文本变量（部分）

来源	说明	定位符	变量
sdtm.ts	TSPARMCD='TITLE'时，取STUDYID	<@方案编号@>	at_studyid
sdtm.ts	TSPARMCD='TITLE'时，取TSVAL	<@研究名称@>	at_study_name
sdtm.ts	TSPARMCD='SPONSOR'时，取TSVAL	<@申办者@>	at_sponser
sdtm.ta	取ARM所有值去重	<@试验分组@>	at_ta_txt
sdtm.te	取ELEMENT所有值去重	<@试验元素@>	at_te_txt
sdtm.ts	TSPARMCD='DCUTDESC'时TSVAL in ['数据库锁定','数据库锁库','锁库']，则为否	<@提交的数据是否来自一项正在进行的研究@>	at_ongoing_yn
除试验设计数据集外的所有数据集	判断所有数据集观测数： 如果存在数据集观测数为0，赋值为"是"，同时输出对应的所有数据集 如果不存在数据集观测数为0，赋值为"否"	<@是否因为没有数据收集导致存在域是计划提交却没有提交@>	at_null_obs_yn
所有数据集	判断所有数据集中是否存在--EVAL变量： 1. 如果存在且不为空，则赋值为"是"，同时输出对应的数据集和对应数据集的--EVAL变量去重后的值 2. 如果不存在，则赋值为"否"	<@是否有判决数据存在@>	at_judgment_yn
赋值：“SDTM 数据集所用编码为 XX。”	编码类型自动生成	<@其他关注的内容@>	at_oth
单个define生成的P21报告	取sheet "Validation Summary"中的“Pinnacle 21 Validator Report”的“Standard:”为sdtm ig版本 取sheet "Validation Summary"中的“Pinnacle 21 Validator Report”的“Software Version:”为P21版本	<@域是否符合CDISC SDTM验证规则@>	at_sdtm_version

自动化文本表格构建算法

文本变量存储为字符串类型变量，示例：

```
#### <@是否有判决数据存在@>
ds_judgment_keys = []
ds_judgment_value = []
for i in list_sdtm_df_name:
    var_eval = i.split('_')[1].upper()+'EVAL'
    # 判断 --EVAL变量是否存在
    if var_eval in eval(i).columns:
        list_temp = [j for j in list(eval(i)[var_eval]) if j]
        u_list_temp = pd.unique(list_temp).tolist()
        u_str_temp = '、'.join(u_list_temp)
        # 判断 --EVAL变量是否有值
        if len(u_list_temp)>0:
            ds_judgment_keys.append(i.replace('df_','').upper())
            ds_judgment_value.append(u_str_temp)

if len(ds_judgment_keys)==0:
    at_judgment_yn = '否'
else:
    at_judgment_yn = '是, '+'; '.join([': '.join(list(i)) for i in zip(ds_judgment_keys,ds_judgment_value)])
```

自动化文本表格构建算法

表格变量（部分）

来源	说明	定位符	变量
sdtm.ta	取ARMCD，ARM去重	<@T_试验分组@>	att_ta
sdtm.te	取ETCD，ELEMENT去重	<@T_试验元素@>	att_te
sdtm.tv	取VISITNUM，VISIT去重	<@T_试验访视@>	att_tv
sdtm.ts	取TSPARMCD，TSPARM，TSVAL去重	<@T_试验概要@>	att_ts
所有sdtm数据集生成的P21报告	取sheet "Issue Summary"中的"Source","Message","Found","Explanation"列，分别对应"数据集","诊断信息","数量","解释" 1. 需要在对P21报告回复的时候将列名给为"Explanation"，如果列名不对应则此列会输出为空 2. 新版本的P21没有"严重程度"列，故未作	<@T_问题总结@>	att_summarize
sdtm.ti	取TIVERS,IECAT,IETESTCD,IETEST去重，分别对应"方案版本","分类","IETESTCD","入选/排除标准的完整描述" 如果未做变量TIVERS，则结果进输出"分类","IETESTCD","入选/排除标准的完整描述"这3列	<@T_入选排除标准@>	att_ti
1. 所有sdtm数据集 2. define生成的spec	"数据集 – 数据集标签","SUPP-","观察类别"可自动判断生成，其他列需要手动填写，因为每个项目不能固定	<@T_SDTM受试者域@>	att_sdtm_sub_domain
1. 所有数据集 2. define生成的spec	define中不会做空的SUPP数据集，所以此处结合所有数据集和define生成的spec综合判断是否存在SUPP数据集： 若存在SUPP数据集，则取对应SUPP数据集的"QNAM"和"QLABEL"列去重	<@T_SUPP_SDTM受试者域_NUM01@>	att_sdtm_sub_supp[0]

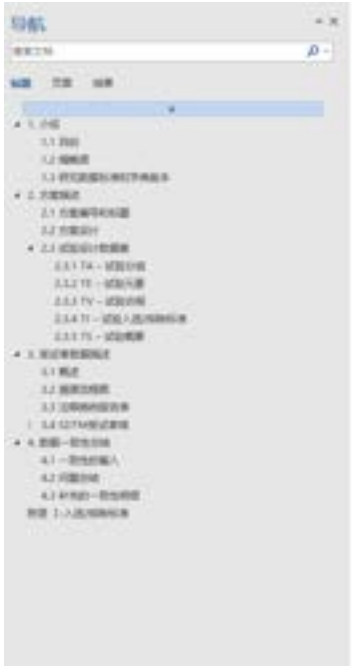
自动化文本表格构建算法

表格变量存储为DataFrame类型变量，示例：

```
#### <@T_研究数据标准和字典版本@>
# SDTM
sdtmig_version_num = re.compile('\d+\.\d+',re.S).findall(sdtmig_version)[0]
if sdtmig_version_num=='3.2':
    sdtm_version_num = '1.4'
if sdtmig_version_num=='3.3':
    sdtm_version_num = '1.7'
if sdtmig_version_num=='3.4':
    sdtm_version_num = '2.0'
sdtm_inc_ig_version = 'SDTM v{}/SDTM IG v{}'.format(sdtmig_version_num,sdtm_version_num)
dic_sdt_dictversion = {
    '标准或字典': ['SDTM'
                ,'Controlled Terminology'
                ,'Data Definitions'
                ,'Medications Dictionary'
                ,'Medical Events Dictionary'],
    '使用的版本': [sdtm_inc_ig_version
                ,sdtmct_version
                ,'define.xml v2.0'
                ,whodra_version
                ,meddra_version]
}
att_sdt_dictversion = pd.DataFrame(dic_sdt_dictversion)
```

变量输出机制

模板文件（标题页）



<@研究名称@>

研究数据审阅者指南

申办者	<@申办者@>
方案编号	<@方案编号@>
方案版本号	
版本号	<@版本号@>
版本日期	<@版本日期@>

变量输出机制

模板文件（方案概述页）

- 2.1 方案编号和标题

  方案编号：<@方案编号@>
  方案标题：<@研究名称@>
  方案版本：

- 2.2 方案设计

  试验流程见下图。

- 2.3 试验设计数据集

- 提交的材料中包括试验设计数据集吗？
  <@提交的材料中包括试验设计数据集吗@>

<@T_提交的材料中包括试验设计数据集吗@>

- 2.3.1 TA – 试验分组

  <@试验分组@>
<@T_试验分组@>

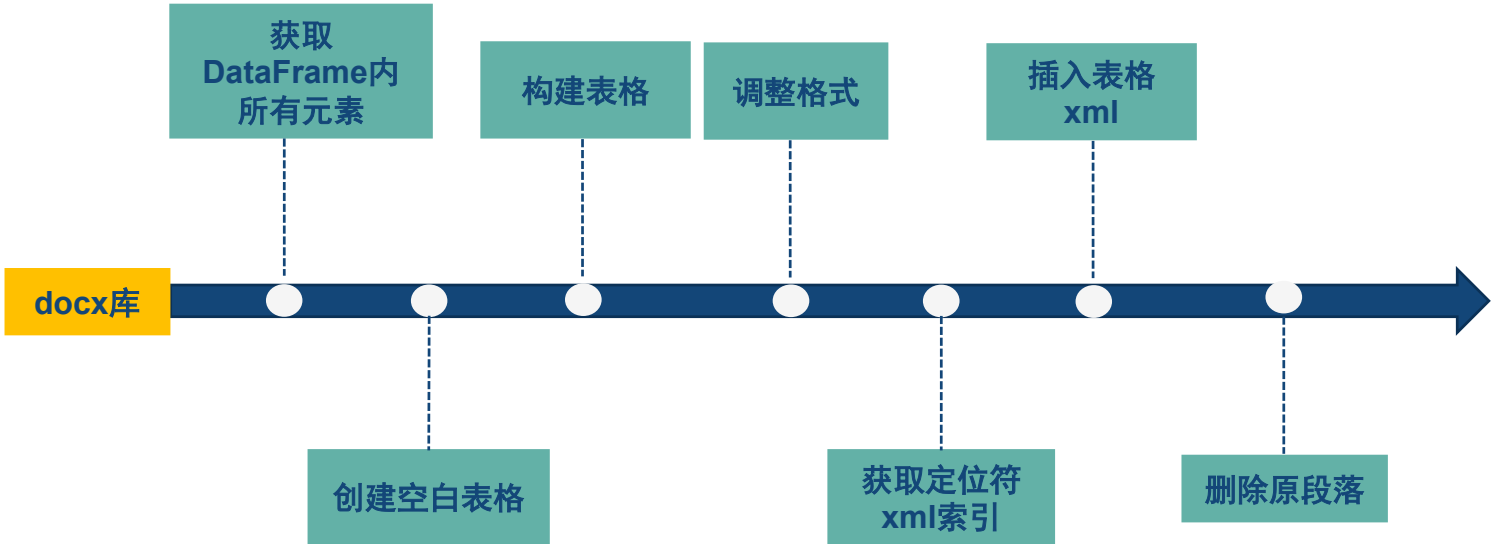
变量输出机制

在模板文件支持下，理想情况下可以标记并唯一定位至准确位置。但输出的文本或表格并非完全与模板内置格式兼容，相同的内容输出至模板顺序不同可能会导致格式完全乱掉。这一问题无法直接实现，需要借助docx库以xml元素形式插入。

构建完所有变量之后，需要对所有定位符进行替换，确保输出形式正常且完全输出之后，删掉未用到的定位符。输出时，除了段落中的表格和文本需要区分之外，页眉页脚中的表格和文本同样需要区分。

变量输出机制

输出过程（表格）



变量输出机制

表格输出内容，示例：

```
#### 替换表格
def replace_text_with_dataframe(input_path, output_path, search_text, df):
    doc = Document(input_path)
    paragraphs = list(doc.paragraphs)  # 获取所有段落
    for paragraph in paragraphs:
        if search_text in paragraph.text:
            # 创建表格，行数为 DataFrame 的行数 + 1（标题行），列数为 DataFrame 的列数
            table = doc.add_table(rows=1, cols=len(df.columns))
            # 填充表头
            for col_idx, col_name in enumerate(df.columns):
                table.cell(0, col_idx).text = str(col_name)
            # 填充表格内容
            for row_idx, row in df.iterrows():
                row_cells = table.add_row().cells
                for col_idx, value in enumerate(row):
                    row_cells[col_idx].text = str(value)
            # 为表格添加边框
            set_table_borders(table)
            # 获取表格的XML元素
            tbl = table._tbl
            # 获取段落的XML元素
            p = paragraph._element
            # 获取段落的父元素（body或tc）
            parent = p.getparent()
            # 获取段落的索引
            idx = parent.index(p)
            # 将表格插入到段落的位置
            parent.insert(idx, tbl)
            # 删除原段落
            parent.remove(p)
```

变量输出机制

最终结果示例：

1. 介绍

1.1 目的

本文档提供列表数据集和术语相关内容，这些内容属于数据定义文档（define.xml）额外的解释。此外，本文档还提供了 SDTM 一致性发现的结果。

1.2 缩略语

缩略语	含义	翻译
aCRF	Annotated Case Report Form	注释病例报告表

1.3 研究数据标准和字典版本

标准或字典	使用的版本
SDTM	SDTM v3.2/SDTM IG v1.4
Controlled Terminology	2022-06-24
Data Definitions	define xml v2.0
Medications Dictionary	WhoDrug_GLOBALB3Sep23
Medical Events Dictionary	MedDRA_26.1

2. 方案描述

2.1 方案编号和标题

方案编号：

方案标题：

方案版本：

3. 受试者数据描述

3.1 概述

- 提交的数据是否来自一项正在进行的研究？
  是。
- SDTM 数据集是否用作分析数据集的来源？
  是。
- 提交的数据集包括筛选失败的受试者吗？
  是，提交的数据集中 CM、DM、DS、DV、IE、PR、SE、SUPPCM、SUPPDM、SUPPDS、SUPPDV、SUPPPR、SUPPSV、SV 中包括筛选失败的受试者。
- 是否因为没有数据收集导致存在域是计划提交却没有提交?
  是，DD 所有受试者均无记录，所以未递交。
- 提交的数据是收集数据的子集吗？
  否，提交的数据与收集的数据范围一致。
- 是否有判决数据存在？
  否。
- 其他关注的内容
  SDTM 数据集所用编码为 UTF-8。

变量输出机制

最终结果示例：

• 3.4 SDTM 受试者域

数据集－数据集标签	疗效性	安全性	其他	自定义域	SUPP	使用 RELREC 相关的域	模型类别
AE－不良事件					X		事件类
CM－既往与合并用药					X		干预类
DD－死亡详情					X		发现类
DM－人口学					X		特殊用途
DS－处置					X		事件类
DV－方案偏离					X		事件类
EC－暴露实施					X		干预类
EG－心电图					X		发现类
EX－暴露							干预类
FA－事件或干预相关发现					X		发现类相关
IE－不满足的入选/排除标准							发现类

• 3.4.9 FA - 事件或干预相关发现

补充术语列表：

QNAM	QLABEL
FAMHNO1	银屑病其他相关疾病 1
FAMHNO2	银屑病其他相关疾病 2
FAMHNO3	银屑病其他相关疾病 3

补充修饰语变量 FAMHNO1、FAMHNO2、FAMHNO3 虽然在 CRF 上进行了标注但因未收集到信息在 SUPPFA 中被移除，所有变量均无记录，所以 SUPPFA 未提交。

• 3.4.10 LB - 实验室检查

补充术语列表：

QNAM	QLABEL
LBCLSIG	临床评估
LBDEC	备注
LBXTORI	来源计划外其他检查表单
LBDESC	异常有临床意义描述

补充修饰语变量 LBXTORI、LBDESC 虽然在 CRF 上进行了标注但因未收集到信息在 SUPPLB 中被移除。

UI及使用介绍

用户界面

UI及使用介绍

运行总结

文件准备时间：约15分钟
运行时间：约1分钟
效率提升：80-90%
完成率：约90%，10%左右的内容需要手动确认填写

关于引入AI的讨论

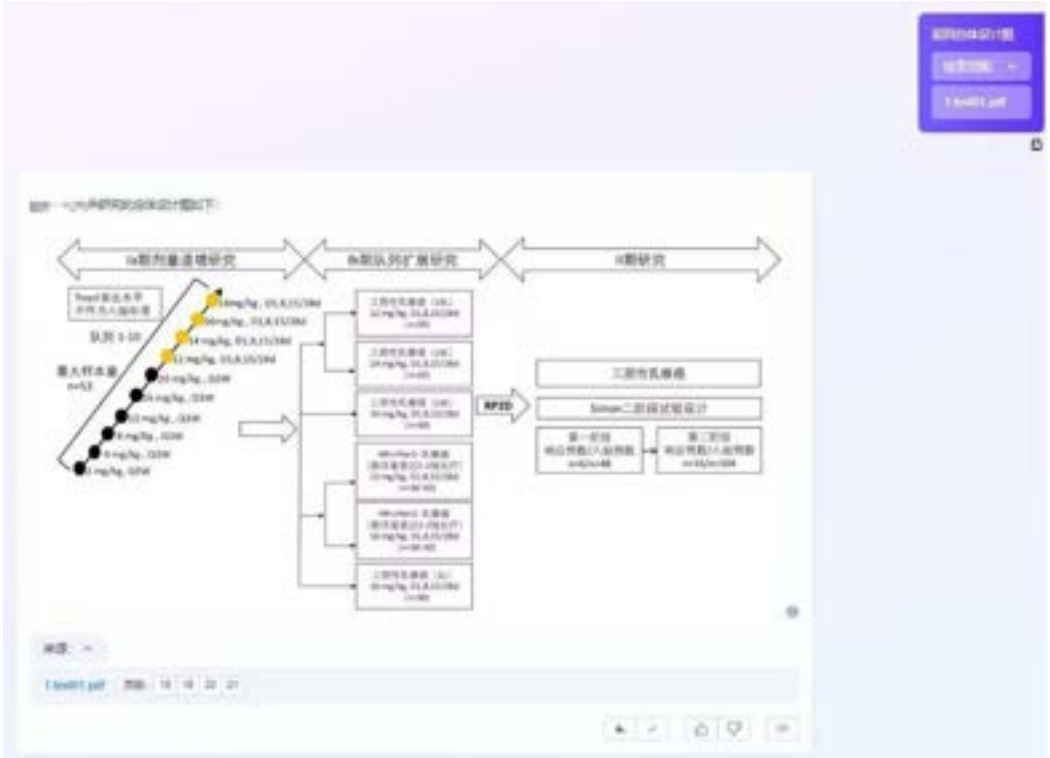
AI应用的可行性

在开发过程中，有一些无法通过代码抓取或处理的内容，例如：aCRF中注释为未提交的内容、方案中流程图、方案设计等。这部分内容都无法精确识别，即便考虑多种解决方案也无法完全覆盖。

经测试将aCRF导入DeepSeek可较为准确地识别出“未提交”字段和对应页码。但由于文本量限制，未能读入完整的文件（一份170页的aCRF能读取前80%左右的内容）。

AI应用的可行性

当前市面上尚无成熟AI能精准提取并还原如临床试验方案这类复杂文档中的流程图。经多轮实测，免费工具中仅腾讯元宝表现尚可，但其稳定性与持续性仍待观察。若对准确率和可控性要求较高，建议采用本地部署方案（参考右图，由恒生电子私有化环境输出）：需集成大模型与高精度OCR，前期投入较大，需结合实际情况使用。



AI应用总结（仅代表个人观点）

优势

- 提升开发者效率与创新： 对工具开发者而言，AI能显著提升工作效率，激发创新思路，并拓展技术视野。
- 优化信息获取与问题解决： 在日常工作中，如文献查阅和问题解答，AI常能提供意想不到的、有价值的解决方案或洞见。
- 加速专业信息检索： 在专业领域（例如基于检索增强生成RAG的文件检索与输出），AI能大幅节省用户的时间成本。

局限

- 可靠性与准确性挑战： AI无法保证100%的稳定性。目前市面上的AI系统普遍尚未完全成熟，存在生成不准确或“幻觉”内容的风险（即看似合理实则错误的信息）。
- 本地部署成本与性能考量： 若需进行本地部署，其所需的硬件投入、算力消耗以及维护成本是需要重点评估的因素。

Thank You!