

AI-Assisted Automated Development of the P21C Validator Report Comparison Tool

Huanhuan Sun, Fosun Pharma

ABSTRACT

During project execution, the Pinnacle 21 Community (P21C) validator report needs to be run multiple times. Traditional manual comparison of old and new reports leads to repetitive work, high error rates, and inefficiency, while making it difficult to effectively track issue closure status. This is especially problematic in later project stages when version differences are minimal, yet every detail still requires line-by-line verification, resulting in resource waste.

To address this business pain point, we have developed an automated solution designed to intelligently manage the lifecycle of P21C report comparisons. This solution leverages Generative AI-assisted development model to accomplish code development, logic debugging, and rapid feature iteration. Its core technical approach is centered on structured field recognition and intelligent difference matching, allowing it to adapt to custom column name variations across different projects. The system achieves automatic discrepancy identification between two report versions, identification of new issues, classifies closed-loop statuses, and intelligently migrates historical comments, enabling a one-click comparison of new and old P21 reports.

This solution significantly reduces the manual verification workload and lowers error rates, proving especially valuable for the rapid confirmation of stable versions in later project phases. Concurrently, it assists Statistical Programmers (SPs) in summarizing data issues for Data Managers (DMs) to perform data cleaning, thereby fostering a more effective and closed-loop workflow between them. This initiative strongly advances the evolution of clinical trial data quality management towards automation, intelligence, and standardization, aligning perfectly with the long-term industry trends of cost reduction, efficiency improvement, and proactive quality management.

INTRODUCTION

During project execution, the P21C Report needs to be generated multiple times to locate, track, and resolve various issues. Previously, after each new version of the Report was produced, the comparison between the new and old versions was performed entirely manually. This manual operation mode has obvious pain points: heavy repetitive workload, low efficiency, and a high susceptibility to human errors such as missed checks and incorrect copying. Not only does it consume significant time and effort, but it also distracts attention from core issues and increases the cost of repeated corrections. This is particularly wasteful in the later stages of the project, when version differences are minimal yet every item still has to be verified line by line. Given these practical pain points, there is an urgent need to develop an automated tool to replace the cumbersome manual comparison process and optimize the management efficiency of P21C Reports.

Generative AI technology is now highly mature and can deeply empower the implementation of various business automation scenarios. This tool leverages AI assistance for code development and rapid iteration, enabling intelligent difference identification between old and new version reports.

P21C VALIDATOR REPORT COMPARISON TOOL INTRODUCTION

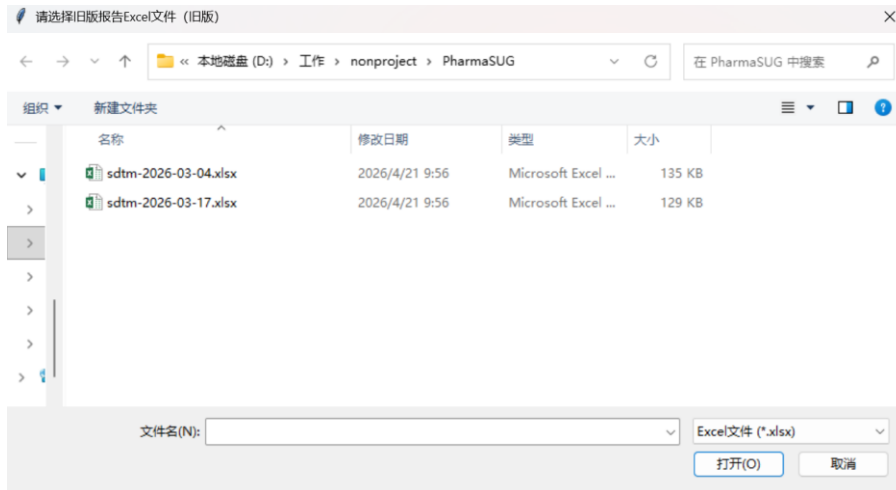
1. CORE CAPABILITIES

The tool features a one-click intelligent difference comparison capability. It can automatically distinguish between historically closed and pending issues, accurately identify newly added issues, and automatically synchronize and migrate historical comment field between old and new version reports. It boasts a high degree of column-name adaptability—regardless of project-specific custom column names, the tool can intelligently recognize the designated column as the historical comment field. For the Issue Summary worksheet, the last column is set as the historical comment by default, for the Details worksheet, the column immediately following the Category field is used as the historical comment field by default. This completely replaces the inefficient manual copy-and-paste operation.

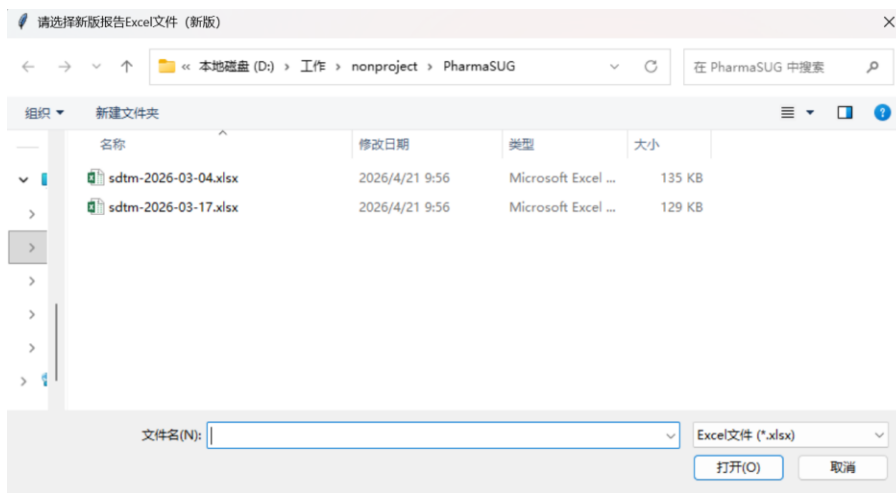
2. DEPLOYMENT FORM

The tool is packaged as a standalone executable (.exe) program that does not require a Python runtime environment. It can be run directly by double-clicking on any Windows device.

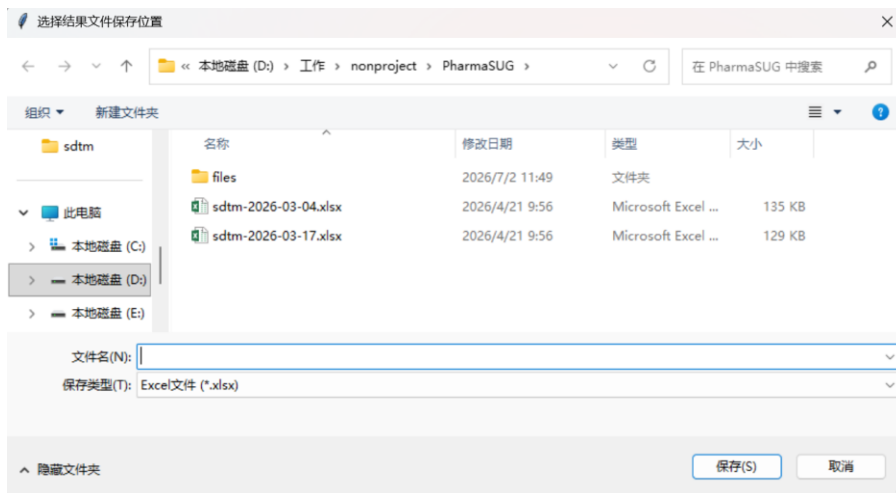
After double-clicking the application, a dialog window will appear for selecting the storage path of the P21C Report, and the tool will remember the path you last chose. You will need to sequentially select the old-version P21C Report file, the new-version P21C Report file, and then specify the name and location for the resulting output file. After clicking save, the comparison process will run, and a pop-up notification will appear upon completion.



Display 1. Pop-up window for selecting the path of old version report



Display 2. Pop-up window for selecting the path of new version report



Display 3. Specify the name and location for the resulting output file



Display 4. The prompt window after the comparison is completed

3. THE OUTPUT FILE AFTER THE TOOL COMPARISON

The Issue Summary worksheet is overwritten and replaced, with three new columns added: Difference Status, Found Change Details, and Historical Comments. A new tab, Historically Fixed Issues, is also added.

	A	B	C	D	E	F	G	H
1	Source	Pinnacle 21 ID	Message	Found	差异状态	Found变动详情	历史备注	
2	GLOBAL							
3	GLOBAL	DD0101	define.xml文件缺失	1	无变化			
4	AE							
5	AE	SD0002	必需AEDECOD变量的值为空值	15	Found条数变更	旧1→新15, 变动14		
6	AE	SD2012	未能在MedDRA词典中找到AEHLGT的值	6	Found条数变更	旧1→新6, 变动5		
7	AE	SD0021	结束时间点缺失	20	Found条数变更	旧9→新20, 变动11		
8	AE	SD1449	MedDRA下有一个或多个变量的值缺失: AELLT, AELLTCD, AEDECOD, AEPTCD, AEHLT, AEHLTCD, AEHLGT, AEHLGTCD, AEBODSYS, AEBODSYCD, AESOC, AESOCCD	15	Found条数变更	旧1→新15, 变动14		
9	EG							
10	EG	SD1023	VISIT/VISITNUM的值域TV域的数据不一致	42	Found条数变更	旧12→新42, 变动30		
11	EX							
12	EX	SD0021	结束时间点缺失	16	Found条数变更	旧5→新16, 变动11		
13	EX	SD1446	暴露开始日期在最近的处理日期之后	1	无变化			
14	EX	SD1078	所有记录中都缺失相关允许变量	1	无变化			
15	IE							
16	IE	SD0065	USUBJID/VISIT/VISITNUM值与SV域的数据不一致	1	本次新增Issue			
17	LB							
18	LB	SD0007	标准单位值不一致	121	本次新增Issue			
19	LB	SD1023	VISIT/VISITNUM的值域TV域的数据不一致	91	Found条数变更	旧58→新91, 变动33		
20	LB	SD0026	有LBORRES值时LBORRESU值缺失	440	Found条数变更	旧114→新440, 变动326		
21	LB	SD0029	有LBSTRES值时LBSTRESU值缺失	557	Found条数变更	旧122→新557, 变动435		
22	LB	SD1470	填写LBORNRO时, LBORRES的值不是数值。	904	Found条数变更	旧276→新904, 变动628		
23	XB							
24	XB	SD1123	当XBSTAT是未查时填写XBORRES值	2	本次新增Issue			
25	XU							
								暂不更新, 目前铁蛋白的单位有

Display 5. Issue Summary worksheet

	A	B	C	D	E	F	G	H	I	J	K
1	Source	Pinnacle 21 ID	Message	Found	Action	Generation	Status	差异状态	Found变动详情	历史备注	
2	DM	SD1358	ACTARMCD或ACTARM为空值时未填写ARMNRS	2				已修复删除 (新版无此报错)			
3	DM	SD1359	未提供ARMCD/ARM时, 也未填写ARMNRS	2				已修复删除 (新版无此报错)			
4	DS							已修复删除 (新版无此报错)			
5	DS	SD1078	所有记录中都缺失相关允许变量	1				已修复删除 (新版无此报错)			
6	EC	SD1149	预期变量缺失所有记录的值	1				已修复删除 (新版无此报错)			
7	EX	SD1149	预期变量缺失所有记录的值	1				已修复删除 (新版无此报错)			
8	LB	SD1078	所有记录中都缺失相关允许变量	1				已修复删除 (新版无此报错)			
9	PC							已修复删除 (新版无此报错)			
10	PC	SD2239	PCTPT值不一致	1				已修复删除 (新版无此报错)			
11	QS							已修复删除 (新版无此报错)			
12	QS	SD0056	未能找到SDTM规定的必需变量QSCAT	1	需更新	SHH/GDT	SHH:已更新	已修复删除 (新版无此报错)		SHH:已更新	
13	SUPPDS							已修复删除 (新版无此报错)			
14	SUPPDS	SD0001	数据源中没有记录	1				已修复删除 (新版无此报错)			
15	SUPPXU							已修复删除 (新版无此报错)			
16	SUPPXU	SD0001	数据源中没有记录	1				已修复删除 (新版无此报错)			
17	VS							已修复删除 (新版无此报错)			
18	VS	SD1078	所有记录中都缺失相关允许变量	1				已修复删除 (新版无此报错)			
19	XU	SD0001	数据源中没有记录	1				已修复删除 (新版无此报错)			
20	YS							已修复删除 (新版无此报错)			
21	YS	SD0031	有YSENDTC、YSENRF、或YSENRTPT值时YSSTDTC、	16	需更新	SHH/GDT	SHH: 增加	已修复删除 (新版无此报错)		SHH: 增加 YSSTRF	
22											
23											
24											
25											
26											
27											
	Validation Summary Dataset Summary Issue Summary 历史已修复Issue Details Rules ...										

Display 6. The new tab Historically Fixed Issues

Detail worksheet: The tool uses five variables—Domain, Variables, Values, Pinnacle 21 ID, and Message—as key fields for conditional matching. The content from the column immediately following the Category field in the old version is then matched as historical comments to the corresponding records in the new-version report.

	A	B	C	D	E	F	G	H	I
1	Domain	Record	Count	Variables	Values	Pinnacle 21 ID	Message	Category	历史备注
2	GLOBAL					DD0101	define.xml文件缺失	Presence	
3	AE	1	1	AEDECOD	null	SD0002	必需AEDECOD变量的值为空值	Presence	
4	AE	2	2	AEDECOD	null	SD0002	必需AEDECOD变量的值为空值	Presence	
5	AE	3	3	AEDECOD	null	SD0002	必需AEDECOD变量的值为空值	Presence	
6	AE	4	4	AEDECOD	null	SD0002	必需AEDECOD变量的值为空值	Presence	
7	AE	9	9	AEDECOD	null	SD0002	必需AEDECOD变量的值为空值	Presence	
8	AE	12	12	AEDECOD	null	SD0002	必需AEDECOD变量的值为空值	Presence	
9	AE	21	21	AEDECOD	null	SD0002	必需AEDECOD变量的值为空值	Presence	
10	AE	22	22	AEDECOD	null	SD0002	必需AEDECOD变量的值为空值	Presence	
11	AE	24	24	AEDECOD	null	SD0002	必需AEDECOD变量的值为空值	Presence	
12	AE	28	28	AEDECOD	null	SD0002	必需AEDECOD变量的值为空值	Presence	
13	AE	32	32	AEDECOD	null	SD0002	必需AEDECOD变量的值为空值	Presence	
14	AE	34	34	AEDECOD	null	SD0002	必需AEDECOD变量的值为空值	Presence	
15	AE	36	36	AEDECOD	null	SD0002	必需AEDECOD变量的值为空值	Presence	
16	AE	37	37	AEDECOD	null	SD0002	必需AEDECOD变量的值为空值	Presence	
17	AE	38	38	AEDECOD	null	SD0002	必需AEDECOD变量的值为空值	Presence	
18	AE	1	1	AEENDTC	null	SD0021	结束时间点缺失	Consistency	缺失
19	AE	2	2	AEENDTC	null	SD0021	结束时间点缺失	Consistency	缺失
20	AE	3	3	AEENDTC	null	SD0021	结束时间点缺失	Consistency	缺失
21	AE	5	5	AEENDTC	null	SD0021	结束时间点缺失	Consistency	缺失
22	AE	6	6	AEENDTC	null	SD0021	结束时间点缺失	Consistency	缺失
23	AE	9	9	AEENDTC	null	SD0021	结束时间点缺失	Consistency	缺失
24	AE	10	10	AEENDTC	null	SD0021	结束时间点缺失	Consistency	缺失
25	AE	16	16	AEENDTC	null	SD0021	结束时间点缺失	Consistency	缺失
26	AE	21	21	AEENDTC	null	SD0021	结束时间点缺失	Consistency	缺失
27	AE	22	22	AEENDTC	null	SD0021	结束时间点缺失	Consistency	缺失
	Validation Summary Dataset Summary Issue Summary 历史已修复Issue Details Rules ...								

Display 7. Details worksheet

4. APPLICATION VALUE

By automating repetitive cross-verification, historical comment migration, and other routine tasks, the tool significantly reduces the probability of human errors and decreases the daily operational workload—especially in terms of rapid confirmation of Issue Summary for stable versions during the later stages of the project. This enables the project team to concentrate on analyzing and implementing corrective actions for critical outstanding issues, substantially improves the overall efficiency of report reviews, and enhances the stability and reliability of quality management throughout the entire project lifecycle.

KEY PRACTICAL POINTS FOR AI-ASSISTED BUSINESS PERSONNEL TO RAPIDLY IMPLEMENT TOOL DEPLOYMENT

When business personnel encounter repetitive needs such as report comparison, they do not need to be proficient in programming; they can directly leverage AI to complete tool development. The following are key practical takeaways:

1. Have the AI assist in determining the technical solution.

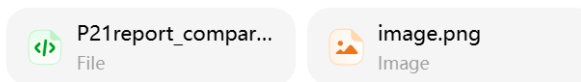
Simply describe the business requirements clearly to the AI (e.g., "need to compare two Excel reports, identify new issues, and inherit historical comments"), and the AI will recommend the most suitable implementation approach. For example, the AI will point out that for Excel data processing, primary key matching, and difference marking, using Python's pandas library is the most efficient, stable, with minimal code and high fault tolerance.

2. Provide complete business context and data structure.

When communicating with the AI for the first time, be sure to provide detailed information about the overall framework of the report, worksheet structure, meaning of key columns, and comparison logic (e.g., what the primary key is, which fields need to be inherited). This prevents the AI from generating incorrect code without understanding the business context (e.g., misidentifying columns, omitting important information).

3. When adding new requirements iteratively, provide the existing code and restate the premises.

When the tool already has a basic version and additional features are needed later, do not assume that the AI can fully recall previous chat history and accurately determine the requirements. Instead, restate the business scenario and the detailed new requirements, and provide the current fully working code. Let the AI first understand the existing code logic before developing the new features. This avoids logical conflicts or redundant functionality, ensuring smooth iteration.



请先了解上述代码的逻辑并在上述代码中新增需求：说明一下，p21 report中有多个sheet，上述截图是p21 report的所有sheet及Details sheet的信息，我们之前在issue summary中做了比对处理，也有一些格式的设置，现需要不改变原逻辑的情况下添加需求：对Details sheet做处理，上述图片是Details 的信息，需要用Domain, Variables, Values, Pinnacle 21 ID, Message做条件匹配上旧版本report中Category列后一列的内容，不管Category列后一列列明是不是空，并且不改变Details sheet的其他格式及超链接功能

Display 8. Example of Prompt Words

With the above methods, business personnel—without needing to learn a programming language—can, with AI assistance, quickly complete the entire development of an automation tool, from requirements to deployment and from prototype to iterative updates, truly achieving an efficient working model that is "business-driven and AI-empowered."

CONCLUSION

In response to the practical pain points in our project—where the P21C Report needs to be run repeatedly, and comparing old and new versions relies entirely on manual operations, resulting in low efficiency, high error rates, and the need to verify every single entry even when version differences are minimal in later project stages, we leveraged Generative AI-assisted development model to develop a standalone automated comparison tool packaged as an executable (.exe) file. This tool automates repetitive cross-verification and historical comment migration, effectively reducing operational workload. Its advantages become particularly pronounced in the later stages of the project, when data is relatively stable, as it enables rapid comparisons, thereby saving the cost of manual checks, helping the team focus on analyzing and resolving critical issues, significantly improving report review efficiency, while also promoting collaborative closure of data-related issues between SP and DM.

Practice has shown that business personnel—without needing to be proficient in programming—can, by clearly describing requirements, providing complete business context and data structure, and attaching existing code and restating the premises during iteration, quickly complete the full development of a tool from prototype to iterative updates with AI assistance. This truly achieves an efficient working model that is "business-driven and AI-empowered." In the future, we will continue to optimize the tool's capabilities, for example, by supporting deeper-level information matching such as matching issue explanations, to further facilitate cSDRG automation.

ACKNOWLEDGMENTS

Many thanks to my leader Wei Diao for the valuable guidance and my colleague Lin Liu for the great support and feedback during the development and implementation of this automation tool.

CONTACT INFORMATION

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

<Huanhuan Sun>
<Fosun Pharma>
<18234191169>
<sunhuanhuan@fosunpharma.com>