



DH11 - Ingenious Raw Data Tracking in SDTM Programs: A New Perspective

PHUSE US Connect 2024

**By :— Bhanoji Duppada
Sumanth Goud Bheemagani**



Ingenious Raw Data Tracking in SDTM Programs: A New Perspective

- Introduces a semi-automated method to enhance tracking of raw datasets within SDTM programs, simplifying traditional complex SAS programming challenges.
- Improves consistency checks between annotated CRFs and raw datasets in SDTM programs, significantly reducing manual effort and time.
- Utilizes MS DOS prompt, batch scripts, and Microsoft Excel to present raw dataset information efficiently, streamlining data analysis and cleaning practices.

Simplifying Complexity:

- Our method minimizes the complexities of SAS programming, allowing for a user-friendly way to track raw dataset names, SDTM variables, and keywords without intensive program modifications.

Unleashing Flexibility:

- Innovates raw dataset tracking with text file inputs and batch scripts, providing flexibility to operate outside study folder constraints and from any accessible location.

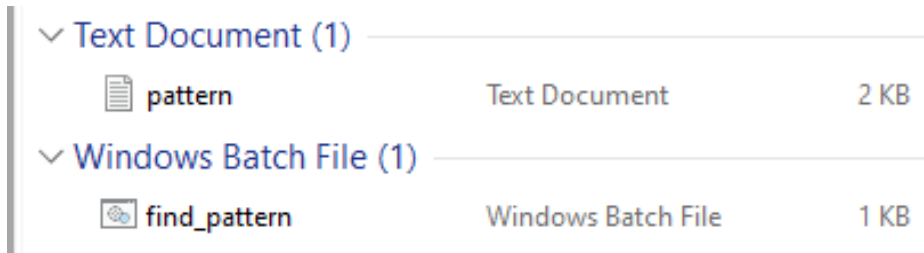
THE PROCESS UNVEILED:

Central to our method is a multi-step process designed for maximum efficiency:

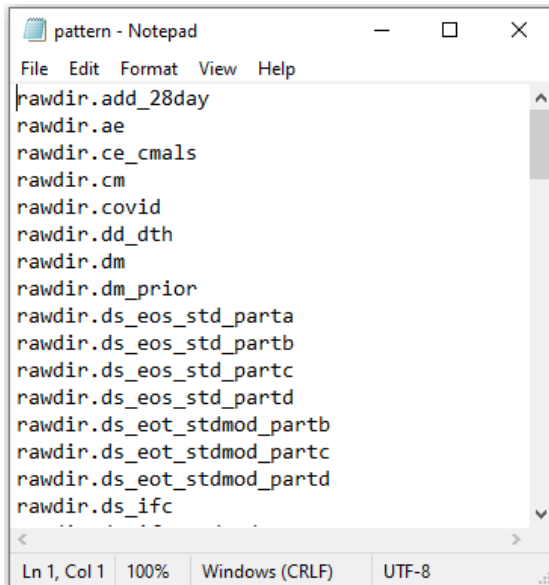
- **1. Pattern File Creation:** A "pattern.txt" file is generated, cataloging raw dataset names.
- **2. Batch Script Crafting:** A ".bat" script is meticulously crafted to scan SDTM program files for dataset occurrences.
- **3. Script Execution:** The batch script is executed, unearthing files containing dataset names, with results seamlessly copied to the "Source_tracker.csv" in the same location of batch script.
- **4. Excel Refinement:** Refine the CSV file using Excel or VBA techniques to copy and filter data, add, and populate auxiliary columns, and remove unnecessary values. Utilize Excel's array features and the IF function to get the desired output.
- **5. Harnessing Excel Functions:** Utilizing Excel's capabilities, including TEXTJOIN, UNIQUE, and IF functions, yields refined dataset details and counts.

PATTERN FILE CREATION

- Create a text file named "pattern.txt" in the same directory as the program and batch files. This file should list all the raw dataset names that will be searched within the program files.



- The text file should display the data in a manner similar to that shown in Figure below

A screenshot of a Notepad window titled 'pattern - Notepad'. The window contains a list of raw dataset names, one per line: rawdir.add_28day, rawdir.ae, rawdir.ce_cmals, rawdir.cm, rawdir.covid, rawdir.dd_dth, rawdir.dm, rawdir.dm_prior, rawdir.ds_eos_std_parta, rawdir.ds_eos_std_partb, rawdir.ds_eos_std_partc, rawdir.ds_eos_std_partd, rawdir.ds_eot_stdmod_partb, rawdir.ds_eot_stdmod_partc, rawdir.ds_eot_stdmod_partd, and rawdir.ds_ifc. The status bar at the bottom shows 'Ln 1, Col 1', '100%', 'Windows (CRLF)', and 'UTF-8'.

BATCH SCRIPT CRAFTING

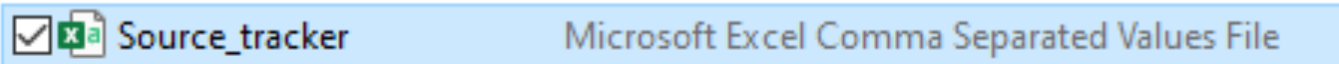
- Generate a text file and paste the provided code into it. Save this file with a '.bat' extension as demonstrated in the accompanying image. Ensure that this batch file is in the same directory as the program files and the pattern file.

```
1 @echo off
2 REM Create or clear the file at the beginning
3 > Source_tracker.csv echo File Search Results
4 for /F "tokens=*" %%i in (pattern.txt) do (
5 if not "%%i"==" " (
6 echo Files containing %%i>>Source_tracker.csv
7 echo Files containing %%i
8 findstr /I /M /C:%%i /S *.sas >> Source_tracker.csv
9 findstr /I /M /C:%%i /S *.sas
10 )
11 )
12 echo Check "Source_tracker.csv" for the results
13 cmd/k
```

Script Execution:

- Double click the batch file to execute and the csv file named Source_tracker will be created.
- The script performs an automated search for specified patterns within SAS program files. It operates by iterating over a list of search terms and scanning through .sas files in the current directory and its subdirectories. The results are compiled into a CSV file for easy review.

▼ Microsoft Excel Comma Separated Values File (1)



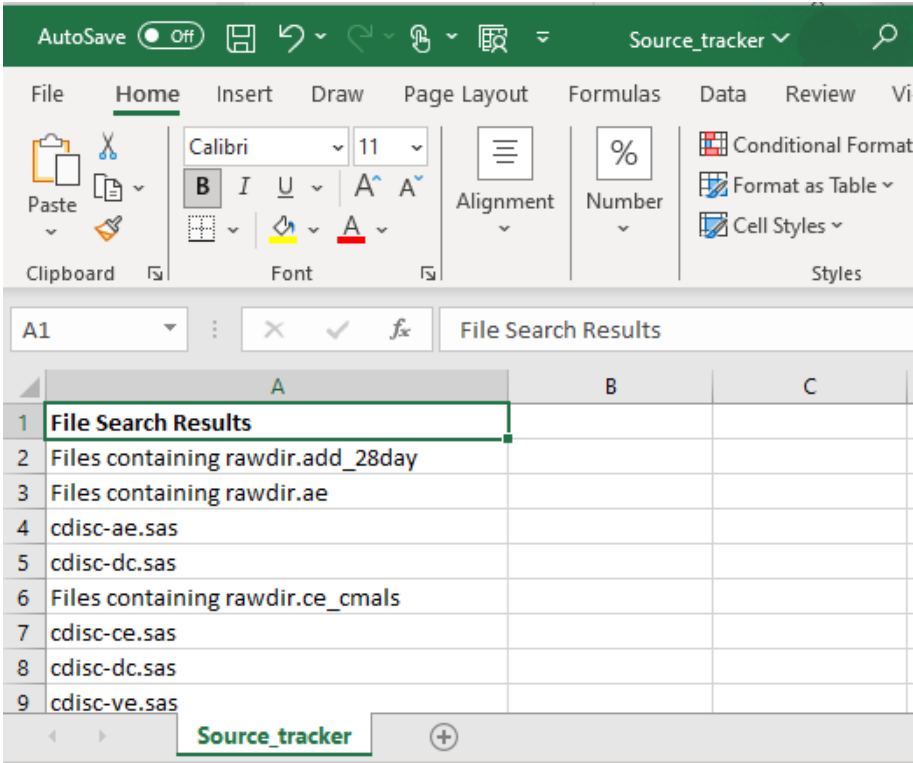
FUNCTIONALITY BREAKDOWN

- 1. Initialization:** The script starts by disabling the display of each command in the command console to maintain a clean output window.
- 2. Result Compilation:** A CSV file named "Source_tracker.csv" is created at the script's start. If this file already exists, it is overwritten to ensure fresh results are captured.
- 3. Pattern Iteration:** The script reads each search term from "pattern.txt", ensuring no line is overlooked.
- 4. Search Execution:** For each non-empty search term, the script:
 - Appends a descriptor line to "Source_tracker.csv", indicating the beginning of results for a new search term.
 - Echoes this descriptor line to the console, providing real-time progress feedback.
 - Executes a case-insensitive search for the term in all .sas files within the current and nested directories, appending the matching file names to "Source_tracker.csv".
 - Outputs the same search results to the console, offering immediate visibility of matches.
- 5. Completion Notification:** Upon completing the searches, the script prompts the user to check "Source_tracker.csv" for a compiled list of results.

4. Excel Refinement:

- The CSV file should display the data in a manner similar to that shown in Figure 1. In Column A, raw dataset names are enumerated, with the names of any corresponding SAS files listed beneath each if the datasets are utilized within those programs.

Figure 1



A	B	C
File Search Results		
Files containing rawdir.add_28day		
Files containing rawdir.ae		
cdisc-ae.sas		
cdisc-dc.sas		
Files containing rawdir.ce_cmals		
cdisc-ce.sas		
cdisc-dc.sas		
cdisc-ve.sas		

Figure 2

A	B
File Search Results	
cdisc-ae.sas	rawdir.ae
cdisc-dc.sas	rawdir.ae
cdisc-ce.sas	rawdir.ce_cmals
cdisc-dc.sas	rawdir.ce_cmals
cdisc-ve.sas	rawdir.ce_cmals
cdisc-ae.sas	rawdir.cm
cdisc-cm.sas	rawdir.cm
cdisc-dc.sas	rawdir.cm
cdisc-mh.sas	rawdir.cm
cdisc-dc.sas	rawdir.covid
cdisc-se.sas	rawdir.covid
cdisc-ve.sas	rawdir.covid
cdisc-dc.sas	rawdir.dd_dth
cdisc-dd.sas	rawdir.dd_dth
cdisc-ds.sas	rawdir.dd_dth

- Apply basic Excel data cleaning methods to refine Column A. This process includes copying and filtering data, creating auxiliary columns, filling in new or supplemental columns, and removing superfluous values. Employ Excel's array functionalities and using IF function and find and replace the string "Files containing" to missing to achieve the desired final output as shown below in Figure 2.
- Column A should be split to 2 columns so that column A contains the sas file names and column B should contain the raw dataset names.

	A	B	C	D
1	File Search Results	File Search Results		
2	Files containing rawdir.add_28day	Files containing rawdir.add_28day	1	Files containing rawdir.add_28day
3	Files containing rawdir.ae	Files containing rawdir.ae	1	
4	cdisc-ae.sas	cdisc-ae.sas		
5	cdisc-dc.sas	cdisc-dc.sas		
6	Files containing rawdir.ce_cmals	Files containing rawdir.ce_cmals	1	

	A	B	C	D
1	File Search Results	File Search Results		
2	Files containing rawdir.add_28day	Files containing rawdir.add_28day	1	Files containing rawdir.add_28day
3	Files containing rawdir.ae	Files containing rawdir.ae	1	Files containing rawdir.ae
4	cdisc-ae.sas	cdisc-ae.sas		
5	cdisc-dc.sas	cdisc-dc.sas		
6	Files containing rawdir.ce_cmals	Files containing rawdir.ce_cmals	1	Files containing rawdir.ce_cmals
7	cdisc-ce.sas	cdisc-ce.sas		
8	cdisc-dc.sas	cdisc-dc.sas		
9	cdisc-ve.sas	cdisc-ve.sas		

Formula Bar: `=IF(C2=1,A2,"")`

A	B	C	D
File Search Results	File Search Results		
Files containing rawdir.add_28day	Files containing rawdir.add_28day	1	Files containing rawdir.add_28day
Files containing rawdir.ae	Files containing rawdir.ae	1	Files containing rawdir.ae
cdisc-ae.sas	cdisc-ae.sas		
cdisc-dc.sas	cdisc-dc.sas		
Files containing rawdir.ce_cmals	Files containing rawdir.ce_cmals		Files containing rawdir.ce_cmals
cdisc-ce.sas	cdisc-ce.sas		
cdisc-dc.sas	cdisc-dc.sas		
cdisc-ve.sas	cdisc-ve.sas		
Files containing rawdir.cm	Files containing rawdir.cm		Files containing rawdir.cm
cdisc-ae.sas	cdisc-ae.sas		
cdisc-cm.sas	cdisc-cm.sas		
cdisc-dc.sas	cdisc-dc.sas		
cdisc-mh.sas	cdisc-mh.sas		
Files containing rawdir.covid	Files containing rawdir.covid		Files containing rawdir.covid
cdisc-dc.sas	cdisc-dc.sas		
cdisc-se.sas	cdisc-se.sas		
cdisc-ve.sas	cdisc-ve.sas		
Files containing rawdir.dd_dth	Files containing rawdir.dd_dth		Files containing rawdir.dd_dth
cdisc-dc.sas	cdisc-dc.sas		

Go To Special

Select

☐ Notes
☐ Constants
☐ Formulas
☒ Numbers
☒ Text
☒ Logicals
☒ Errors
☒ Blanks
☐ Current region
☐ Current array
☐ Objects

☐ Row differences
☐ Column differences
☐ Precedents
☐ Dependents
☒ Direct only
☐ All levels
☐ Last cell
☐ Visible cells only
☐ Conditional formats
☐ Data validation
☒ All
☐ Same

OK Cancel

A	B	C	D
File Search Results	File Search Results		
Files containing rawdir.add_28day	Files containing rawdir.add_28day	1	Files containing rawdir.add_28day
Files containing rawdir.ae	Files containing rawdir.ae	1	Files containing rawdir.ae
cdisc-ae.sas	cdisc-ae.sas		
cdisc-dc.sas	cdisc-dc.sas		
Files containing rawdir.ce_cmals	Files containing rawdir.ce_cmals	1	Files containing rawdir.ce_cmals
cdisc-ce.sas	cdisc-ce.sas		
cdisc-dc.sas	cdisc-dc.sas		
cdisc-ve.sas	cdisc-ve.sas		
Files containing rawdir.cm	Files containing rawdir.cm	1	Files containing rawdir.cm
cdisc-ae.sas	cdisc-ae.sas		
cdisc-cm.sas	cdisc-cm.sas		
cdisc-dc.sas	cdisc-dc.sas		
cdisc-mh.sas	cdisc-mh.sas		

SUM			
Functions	A	B	C
1	File Search Results	File Search Results	
2	Files containing rawdir.add_28day	Files containing rawdir.add_28day	1 Files containing rawdir.add_28day
3	Files containing rawdir.ae	Files containing rawdir.ae	1 Files containing rawdir.ae
4	cdisc-ae.sas	cdisc-ae.sas	=D3
5	cdisc-dc.sas	cdisc-dc.sas	
6	Files containing rawdir.ce_cmals	Files containing rawdir.ce_cmals	1 Files containing rawdir.ce_cmals
7	cdisc-ce.sas	cdisc-ce.sas	
8	cdisc-dc.sas	cdisc-dc.sas	
9	cdisc-ve.sas	cdisc-ve.sas	

File Search Results	File Search Results	
Files containing rawdir.add_28day	Files containing rawdir.add_28day	1 rawdir.add_28day
Files containing rawdir.ae	Files containing rawdir.ae	1 rawdir.ae
cdisc-ae.sas	cdisc-ae.sas	rawdir.ae
cdisc-dc.sas	cdisc-dc.sas	rawdir.ae
Files containing rawdir.ce_cmals	Files containing rawdir.ce_cmals	1 rawdir.ce_cmals
cdisc-ce.sas	cdisc-ce.sas	rawdir.ce_cmals
cdisc-dc.sas	cdisc-dc.sas	rawdir.ce_cmals
cdisc-ve.sas	cdisc-ve.sas	rawdir.ce_cmals

A	B	C
File Search Results		rawdir.add_28day
cdisc-ae.sas	rawdir.ae	rawdir.ae
cdisc-dc.sas	rawdir.ae	rawdir.ce_cmals
cdisc-ce.sas	rawdir.ce_cmals	rawdir.cm
cdisc-dc.sas	rawdir.ce_cmals	rawdir.covid
cdisc-ve.sas	rawdir.ce_cmals	rawdir.dd_dth
cdisc-ae.sas	rawdir.cm	rawdir.dm
cdisc-cm.sas	rawdir.cm	rawdir.dm_prior
cdisc-dc.sas	rawdir.cm	rawdir.ds_eos_std_parta
cdisc-mh.sas	rawdir.cm	rawdir.ds_eos_std_partb

Desired output shown below

A	B
File Search Results	
cdisc-ae.sas	rawdir.ae
cdisc-dc.sas	rawdir.ae
cdisc-ce.sas	rawdir.ce_cmals
cdisc-dc.sas	rawdir.ce_cmals
cdisc-ve.sas	rawdir.ce_cmals
cdisc-ae.sas	rawdir.cm
cdisc-cm.sas	rawdir.cm
cdisc-dc.sas	rawdir.cm
cdisc-mh.sas	rawdir.cm
cdisc-dc.sas	rawdir.covid
cdisc-se.sas	rawdir.covid
cdisc-ve.sas	rawdir.covid
cdisc-dc.sas	rawdir.dd_dth
cdisc-dd.sas	rawdir.dd_dth
cdisc-ds.sas	rawdir.dd_dth

HARNESSING EXCEL FUNCTIONS

To track the raw dataset names and their corresponding sas programs

- Copy unique raw dataset names into column C, then use the formula =TEXTJOIN(", ", TRUE, UNIQUE(IF(C1=\$B\$1:\$B\$245,\$A\$1:\$A\$245,""))) in column D to list program names where the raw dataset is used.
- Drag the formula down column D to display all program names associated with each raw dataset name.
- To count the number of programs using a raw dataset, input =IF(D1<>"",LEN(D1)-LEN(SUBSTITUTE(D1,"",""))+1,"") in column E and extend it to all cells in the column.
- The formulas aggregate unique program names and count them
- Adjust the cell ranges as necessary for different data sets.

D2				=TEXTJOIN(", ",TRUE,UNIQUE(IF(C2=\$B\$1:\$B\$245,\$A\$1:\$A\$245,"")))	
	A	B	C	D	E
1	File Search Results		rawdird.add_28day		
2	cdisc-ae.sas	rawdird.ae	rawdird.ae	cdisc-ae.sas, cdisc-dc.sas	2
3	cdisc-dc.sas	rawdird.ae	rawdird.ce_cmals	cdisc-ce.sas, cdisc-dc.sas, cdisc-ve.sas	3
4	cdisc-ce.sas	rawdird.ce_cmals	rawdird.cm	cdisc-ae.sas, cdisc-cm.sas, cdisc-dc.sas, cdisc-mh.sas	4
5	cdisc-dc.sas	rawdird.ce_cmals	rawdird.covid	cdisc-dc.sas, cdisc-se.sas, cdisc-ve.sas	3
6	cdisc-ve.sas	rawdird.ce_cmals	rawdird.dd_dth	cdisc-dc.sas, cdisc-dd.sas, cdisc-ds.sas	3
7	cdisc-ae.sas	rawdird.cm	rawdird.dm	cdisc-dc.sas	1
8	cdisc-cm.sas	rawdird.cm	rawdird.dm_prior	cdisc-dc.sas	1
9	cdisc-dc.sas	rawdird.cm	rawdird.ds_eos_std_parta	cdisc-dc.sas, cdisc-ds.sas, cdisc-se.sas	3
10	cdisc-mh.sas	rawdird.cm	rawdird.ds_eos_std_partb	cdisc-dc.sas, cdisc-ds.sas, cdisc-se.sas	3
11	cdisc-dc.sas	rawdird.covid	rawdird.ds_eos_std_partc	cdisc-dc.sas, cdisc-ds.sas, cdisc-se.sas	3
12	cdisc-se.sas	rawdird.covid	rawdird.ds_eos_std_partd	cdisc-dc.sas, cdisc-ds.sas, cdisc-se.sas	3

To track the sas program names and their corresponding raw dataset names

- Copy unique raw dataset names into column C, then use the formula =TEXTJOIN(", ", TRUE, UNIQUE(IF(C1=\$B\$1:\$B\$245,\$A\$1:\$A\$245,""))) in column D to list program names where the raw dataset is used.
- Drag the formula down column D to display all program names associated with each raw dataset name.
- To count the number of programs using a raw dataset, input =IF(D1<>"",LEN(D1)-LEN(SUBSTITUTE(D1,"",""))+1,"") in column E and extend it to all cells in the column.
- The formulas aggregate unique program names and count them.
- Adjust the cell ranges as necessary for different data sets.

A	B	C	D	E
cdisc-ae.sas	rawdir.ae	cdisc-ae.sas	rawdir.ae, rawdir.cm, rawdir.ho_hospt, rawdir.pr_ndt	4
cdisc-dc.sas	rawdir.ae	cdisc-dc.sas	rawdir.ae, rawdir.ce_cmals, rawdir.cm, rawdir.covid, ra	60
cdisc-ce.sas	rawdir.ce_cmals	cdisc-ce.sas	rawdir.ce_cmals	1
cdisc-dc.sas	rawdir.ce_cmals	cdisc-ve.sas	rawdir.ce_cmals, rawdir.covid, rawdir.ds_ifc, rawdir.ds	45
cdisc-ve.sas	rawdir.ce_cmals	cdisc-cm.sas	rawdir.cm	1
cdisc-ae.sas	rawdir.cm	cdisc-mh.sas	rawdir.cm, rawdir.mh, rawdir.pr_ndt, rawdir.sb_sod1	4
cdisc-cm.sas	rawdir.cm	cdisc-se.sas	rawdir.covid, rawdir.ds_eos_std_parta, rawdir.ds_eos_	48
cdisc-dc.sas	rawdir.cm	cdisc-dd.sas	rawdir.dd_dth	1
cdisc-mh.sas	rawdir.cm	cdisc-ds.sas	rawdir.dd_dth, rawdir.ds_eos_std_parta, rawdir.ds_eo	18

ADVANTAGES

- The input search values were provided dynamically through a text file and can be modified easily as needed.
- No need to change or update the batch script after creation.
- No need to learn the excel advanced functions, clean the excel sheet and maintain the data in proper columns as shown previously and copy and paste the formula. So that anybody can easily obtain the results with basic excel knowledge.
- The batch file assumes the sas files and pattern.txt file in the same location of the batch files, so no need to provide the path manually.
- If the results when compared with the annotated CRF, so that discrepancies will be found if any exists.
- The batch file can be adapted in various ways to look for strings, specific keywords, and regular expressions. Additionally, users can input search words in the pattern.txt file according to their requirements based on the SAS programs, maximizing the effectiveness of this process.

LIMITATIONS

- Ensure user access to the study folder to run the batch script; if not available, copy the SAS files, pattern.txt, and batch script to an accessible workspace.
- User must have access to the command prompt, which is typically available by default in most Windows environments.
- Getting the unique name of raw dataset names or program names seems difficult but this can be achieved by using simple CMD commands like DIR /B /A *.*.
- Remove irrelevant comments containing raw dataset names from SAS programs to maintain program cleanliness and ensure accurate results.
- Initially map programs using both library name and raw dataset name for effective monitoring of raw datasets.
- If initial mapping is not done, modify batch scripts to include regular expressions for precise results

CONCLUSION

- Introduces an efficient method for tracking raw datasets in SDTM programs, overcoming complex SAS programming and accessibility issues.
- Utilizes a combination of MS DOS prompt, batch scripts, and Microsoft Excel to simplify the data tracking process. Offers a user-friendly and flexible solution that improves data analysis and cleaning practices.
- Enables precise dataset tracking and cross-referencing with annotated CRFs for discrepancy detection and effective SDTM mapping validation.
- Requires well-structured program files and minimal extraneous comments for optimal performance.

References

<https://trumpexcel.com/multiple-lookup-values-single-cell-excel/>

CONTACT INFORMATION

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

Bhanoji Duppada
Vita Clinical Solutions
281 Winter Street, 1st Floor
Waltham, MA 02451
Work Email: bduppada@softworldinc.com

Sumanth Goud Bheemagani
Merck & Co., Inc. USA
351 N Sumneytown Pike,
Upper Gwynedd, North Wales, PA 19454
Work Email: Sumanth.goud.bheemagani@merck.com

THANK YOU