Data Profiling Made *Shiny*

Accelerating Clinical Study Delivery & Submission

Marc Hughes marc.hughes@astrazeneca.com Hao Chen hao.chen15@astrazeneca.com
Ruhul Amin ruhul.amin@astrazeneca.com

PHUSE PP07

AstraZeneca — What science can do

Abstract

Knowing the study data is the first step for every statistical programmer before creating Tables, Listings and Figures for reporting and regulatory submission. Currently, the data is opened manually, or ad-hoc programs are written to explore data insights. This time-consuming and repetitive process highlighted the need for an automation tool that can generate statistical summaries about a given data set regardless of the therapeutic area in an efficient manner. This innovative solution is designed and developed with flexibility, the user has the choice to either generate basic but fast results about the contents of data sets or they can opt for a customised solution involving an RShiny app presenting a more user-friendly and visually appealing experience. This data profiling tool is saving valuable programming time by replacing manual & repetitive work, thus accelerating clinical study delivery & submission milestones and is scalable across all studies and therapeutic areas.

The Motivation

Before a programmer wishes to work with some data, it's methodical to begin by gaining an understanding about the basics of that data. NULL values, variable names/types and number of observations are key concepts to consider before any quality programming can begin. Some of this information can be viewed by simply opening the dataset and scrolling back and forth, up and down – a tedious process should you forget this information and have to go searching again. SAS® macros can be created to try and automate this process but they will likely be created every time a user needs to profile their data. It was deemed appropriate to have a centralised program that can run within a few seconds and produce a report summarising the results that the user could refer to again...

A&R (SCE)-Based Solution

The biggest challenge of all has and will always be regulation as we always put patients first here at AstraZeneca. To get an early understanding of the value that the tool could bring, it was clear that AstraZeneca's GxP-qualified Scientific Computing Environment (SCE) platform should be the home for the tool as this reduces the amount of development time since the tool can work alongside the clinical data. This solution however throws in a couple of key development challenges...

Challenge 1: Limited to Validated Packages

Having R functionality within a GxP environment means that any packages available must be validated for use on clinical data. There may also be file extension constraints such as .Rmd which cannot be saved/executed. Some packages not validated for use include packages designed for creating reports and attribute finding.

Challenge 2: User Interaction

The SCE application offers a very simplistic user interface. Interaction is a key selling point for this tool and therefore the solution needs to reflect the desire for quick and simple data profiling.

Quick Results:

Users can gain quick insights into their data simply by running the program and entering the location path holding the SAS® datasets the user wishes to profile in the "data_area" parameter box within the execution pop-up window (see Figure 2). After executing, the user will be presented with a report of the findings.

Parameters

name	related to	in	out	ma...	type	quotes	value
data_area		☑	☐	☑	Path	NO_QUOTES	NULL/NULL
config		☑	☐	☐	File	NO_QUOTES	<undefined>
max_length		☑	☐	☑	Numeric	NO_QUOTES	30
max_width		☑	☐	☑	Numeric	NO_QUOTES	200
centralized		☑	☐	☑	Yes,No;Y,N	NO_QUOTES	Yes
col_n		☑	☐	☑	Numeric	NO_QUOTES	3

Customised Results:

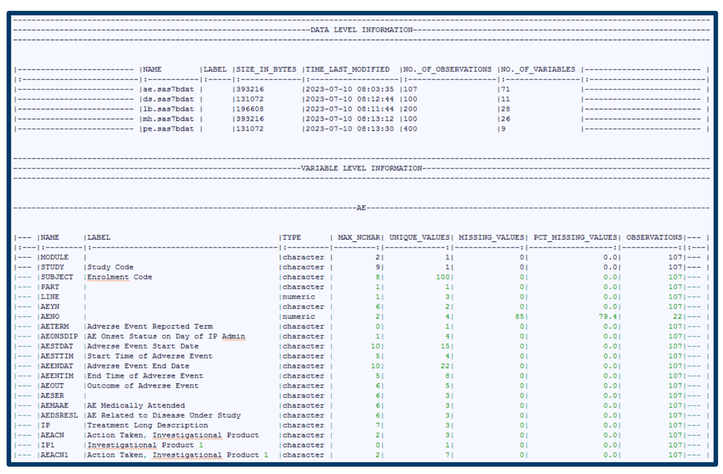
Users can customise the output of results from this tool using a JSON configuration file. Users can include or exclude the information they input into the file. This includes the ability to select the files to display variable level information as well as selecting variables to display value level information.

Other parameters available can be used to personalise the report, which includes dictating the number of results/columns to present the data in and whether to truncate any long data values.

Configuration File

```
{
"include": true,
"variable_level": [
"ae","eg","mh","se","ho"],
"value_level": [{
"dataset": "ae",
"variables": ["STUDYID","AETERM","AEBODSYSCD","AESEQ","AEOUT","EPOCH"]},
{"dataset": "rs",
"variables": ["RSTEST","RSORRES"]
}]
}
```

Example Output



DATA LEVEL INFORMATION

NAME	LABEL	SIZE_IN_BYTES	TIME_LAST_MODIFIED	NO._OF_OBSERVATIONS	NO._OF_VARIABLES
ae.sas7bdat		393216	2023-07-10 08:03:35	107	71
ds.sas7bdat		131072	2023-07-10 08:12:44	100	11
lb.sas7bdat		196608	2023-07-10 08:11:44	200	28
mh.sas7bdat		393216	2023-07-10 08:13:12	100	26
pe.sas7bdat		131072	2023-07-10 08:13:30	400	9

VARIABLE LEVEL INFORMATION — AE

NAME	LABEL	TYPE	MAX_NCHAR	UNIQUE_VALUES	MISSING_VALUES	PCT_MISSING_VALUES	OBSERVATIONS
MODULE		character	2	1	0	0.0	107
STUDY	Study Code	character	9	1	0	0.0	107
SUBJECT	Enrolment Code	character	8	100	0	0.0	107
PART		character	1	1	0	0.0	107
LINE		numeric	1	3	0	0.0	107
AEYN		character	6	2	0	0.0	107
AENO		numeric	2	4	85	79.4	22
AETERM	Adverse Event Reported Term	character	0	1	0	0.0	107
AEONSDIP	AE Onset Status on Day of IP Admin	character	1	4	0	0.0	107
AESTDAT	Adverse Event Start Date	character	10	15	0	0.0	107
AESTTIM	Start Time of Adverse Event	character	5	4	0	0.0	107
AEENDAT	Adverse Event End Date	character	10	22	0	0.0	107
AEENTIM	End Time of Adverse Event	character	5	8	0	0.0	107
AEOUT	Outcome of Adverse Event	character	6	5	0	0.0	107
AESER		character	6	3	0	0.0	107
AEMAAE	AE Medically Attended	character	6	3	0	0.0	107
AEDSRESL	AE Related to Disease Under Study	character	6	3	0	0.0	107
IP	Treatment Long Description	character	7	3	0	0.0	107
AEACN	Action Taken, Investigational Product	character	2	3	0	0.0	107
IP1	Investigational Product 1	character	0	1	0	0.0	107
AEACN1	Action Taken, Investigational Product 1	character	2	7	0	0.0	107

Example output report does not include real data.

The Tool

The data profiling tool allows users to quickly generate commonly sought-after information about SAS® datasets. Users can customise the information presented to them in various ways and a report is generated for users to conveniently refer to the information at any time. Split into 3 tiers, users can view data level, variable level and value level information:

DATA LEVEL INFORMATION:
Dataset Name, Label, Size, Time last modified, Number of Variables/Observations.

VARIABLE LEVEL INFORMATION:
Variable Name, Label, Type, Length, Format, # Unique/Missing Values, % Missing Values.

VALUE LEVEL INFORMATION:
Character Variables: Value Frequencies. Numeric Variables: Summary Tables.

The A&R Scientific Computing Environment (SCE) platform and Shiny platform versions both produce this information, however the solution offers different levels of customisation and user interaction.

The Value

This tool provides functionalities that facilitate swift reviews and queries of data profiles, while also enabling documentation for future use. It promotes efficient planning and data analysis, with early data insights helping to mitigate future problems and reduce debugging efforts. By generating a clean, consistent data summary reports, the tool bolsters the reliability of data-driven decisions. This tool is designed to automate manual statistical programming tasks and could expand to other data analysis functions, saving programmers' time and reducing delivery timelines. This tool sets the foundation for future AI-driven data analysis, using metadata to guide Large Language Models in creating complex code for insightful data analysis while preserving source data confidentiality.

Making It *Shiny*-er

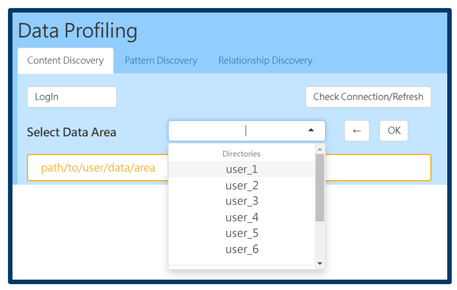
Turning this tool into a Shiny app offers users a much more interactive and hassle-free experience. The app will be presented with a host of widgets to customise what information is presented to users. The app is separated into three navigation tabs...

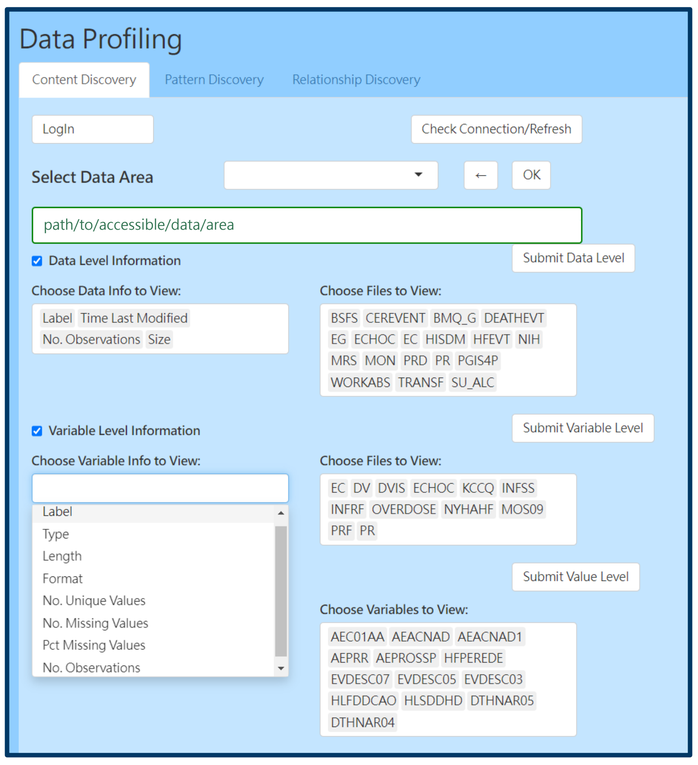
Content Discovery: The base function of the data profiling tool. Offers the user to view data level, variable level and value level information about their datasets.

Relationship Discovery: The user can use cross-table calculations to determine correlations between variables.

Pattern Discovery: The user can identify value patterns in character/date variables.

The majority of the app will run on clickable/selectable widgets meaning that it will be very simple to use. Upon successful login of the GxP SCE platform through the app, the user will be presented with a data_area navigation box which can be used to locate an area holding SAS® datasets. When the currently selected path has no accessible SAS® datasets, the path is highlighted orange. Once a valid data area has been found, the path will turn green and the user can subsequently click the 'OK' button to confirm the path. The user also has the option to revert to the parent directory using the back (←) button as well as selecting a path that has already been previously submitted by the user in a previous session.
After submitting the data area, the user is then presented with the options allowing for customising the output of the tool.





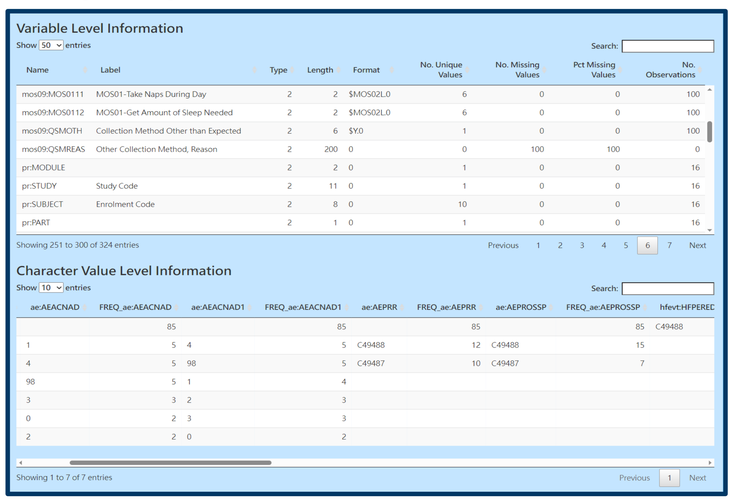
The above graphic does not include real file/variable names.

#1 Challenge: Moving Clinical Data Out of GxP Validated Environment

The main challenge for this version of the tool is to remove all risks posed with the transfer of any clinical data being transferred across platforms.
This includes:

- Making sure each user only has access to the data they have permission to view,
- Any data removed from GxP platform is not manipulable and any changes made are tracked within the GxP area,
- Platform data storage areas are wiped after a period of non-use,
- Maintaining as much of the original dataset as possible but removing any patient distinguishing information.

Example Output



Variable Level Information

Name	Label	Type	Length	Format	No. Unique Values	No. Missing Values	Pct Missing Values	No. Observations
mos09:MOS0111	MOS01-Take Naps During Day	2	2	$MOS02L0	6	0	0	100
mos09:MOS0112	MOS01-Get Amount of Sleep Needed	2	2	$MOS02L0	6	0	0	100
mos09:QSMOTH	Collection Method Other than Expected	2	6	$Y.0	1	0	0	100
mos09:QSMREAS	Other Collection Method, Reason	2	200	0	0	100	100	0
pr:MODULE		2	2	0	1	0	0	16
pr:STUDY	Study Code	2	11	0	1	0	0	16
pr:SUBJECT	Enrolment Code	2	8	0	10	0	0	16
pr:PART		2	1	0	1	0	0	16

Showing 251 to 300 of 324 entries — Previous 1 2 3 4 5 6 7 Next

Character Value Level Information

ae:AEACNAD	FREQ_ae:AEACNAD	ae:AEACNAD1	FREQ_ae:AEACNAD1	ae:AEPRR	FREQ_ae:AEPRR	ae:AEPROSSP	FREQ_ae:AEPROSSP	hfevt:HFPEREDE
	85		85		85		85	C49488
1	5	4	5	C49488	12	C49488	15	
4	5	98	5	C49487	10	C49487	7	
98	5	1	4					
3	3	2	3					
0	2	3	3					
2	2	0	2					

Showing 1 to 7 of 7 entries — Previous 1 Next

Example output report does not include real data.

Next Steps...

There are a couple of desired features that are yet to be included in the current iteration of the tool. These include:

Dataset Metadata Comparison

Being able to compare dataset metadata across versions is a useful feature in the current A&R environment which this data profiling tool looks to implement.

Cross-Table Calculations

Cross-table analyses allow users to gain basic insights into relationships between variables. A simple but useful feature to include in this app.

Pattern Detection

The current version of the tool can give you information about variable type, format name and length which are all vital pieces of information, but do not give any perception about possible common traits between values. This is where pattern detection can be extremely useful.