PharmaSUG China 2017 - Paper 34

Data Explore Matrix – Quality Control by Exploring and Mining Data in Clinical Study

Yongxu Tang, Merck, Beijing, China
Yu Meng, Merck, Beijing, China
Yongjian Zhou, Merck, Beijing, China
Yang Chong, Merck, Beijing, China
Yangfei Ma, Merck, Beijing, China

ABSTRACT

General data review mainly focus on two aspects: one is for in-house study review, the other is for traditional outsource surveillance review.

This paper introduces a new method called Data Explore Matrix (DEM), which is based on SAS® Version 9.3 to support quality control in clinical studies which consists of 1) High Risk Data Points Identity (HRDI) 2) Experience and knowledge input (E&K) 3) Tools and Models (T&M). Combined with those fields of knowledge towards visualization methods Data Explore Matrix (DEM) shows users with a novel way of reviewing data and detecting signals in a quick time.

In this paper we will share some examples using Data Explore Matrix (DEM) focus on Query, Domain Data, and Laboratory. The author is convinced that Data Explore Matrix (DEM) is useful to support integrated information review, fast issue identification and perspective decision making.

INTRODUCTION

Nowadays, quality control in clinical studies are more and more important, considering and measuring time and cost are far more simple and with lots of experiences. However, to measure and manage quality is a hot and interesting topic which attracts more and more attention especially on fast decision making. The current status of data review mainly focus on two aspects: one is for in-house study review, the other is for traditional outsource surveillance review. Most of data review for in-house study are using Data Validation Plan (DVP) as main source, the outputs of DVP are either queries fired in the system or some detailed listings including specific subject information, which is very detail and hard to find the root cause of the problem. In traditional surveillance data review, we are mainly relay on basic Tables, Listings and Figures (TLFs) which is a kind of simple visualization of data issues but still with some restrictions on extensibility and compatibility. In this paper we will share a new method called Data Explore Matrix (DEM), which is based on SAS® Version 9.3 to support quality control in clinical studies.

CONCEPT OF DATA EXPLORE MATRIX (DEM)

The concept of Data Explore Matrix (DEM) is consist of three elements:

- High Risk Data Points Identity (HRDI)
- Experience and Knowledge input (E&K)
- Tools and Models (T&M)

HRDI is a data list which summarized from edit checks or data listings through statistical analysis, consist of critical data points with high potential risk. HRDI should not only be based on single study but be run retrospectively on all studies that with same indication or similar study design/purpose, in order to obtain objective, appropriate and effective result. HRDI enables DEM more focus on the key data points that matters and improves efficiency. E&K is diversity and related to specific people experience. Brain storming, continuous training and catching up with industrial advanced technology are good ways to collect more ideas in E&K field. T&M is general graphs, tables and models developed based on variety platforms, there are mature tools and models could be referenced in pharmaceutical industry, however, the best practice of T&M is to keep an open mind to all other industries. There is no limitation or bottle neck across the industries to use those Tools and Models.

These three elements are fundamentals of the DEM, which can be freely combined with each other (Figure 1).

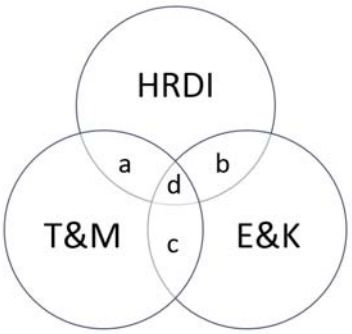

Figure 1. Concept of DEM

Data visualization ideas can be generated from the elements combination part. As shown in Figure 1, field a, b, c and d are all potential production areas of the visualization ideas. In the following part of this paper we will show some examples developed based on those potential production areas.

COMPARISON BETWEEN IN-HOUSE METHOD, TRADITIONAL SURVEILLANCE METHOD AND DATA EXPLORE MATRIX (DEM)

Before starting with data visualization using DEM model, we need to clearly understand why we choose this one compared with other methods. As mentioned previously, in-house study review and traditional outsource surveillance review are mostly used in current pharmaceutical industry for data review. It is convinced that using DEM can bring more benefits compared with other methods, as summarized in Table 1 below.

	In-house Method	**Traditional Surveillance Method**	**Data Explore Matrix (DEM)**
Layout	Detailed listings (DVP, etc.)	Basic TLFs (scatter plot, bar chart, etc.)	Integrated visualized figures
Route Cause Analysis	Very hard	Hard	Fast and directly
Explore Unknown	No	Normal	Good
Compatibility	No, trial specific	Normal, with modification	Good
Perceptiveness	No	No	Yes

Table 1. Comparison between In-house Method, Traditional Surveillance Method and Data Explore Matrix (DEM)

It should be aware that the utilizing of DEM should be accordance with customers' perspective. For example, it could not be considered as a benefit when looking into detailed data issues using DEM, however, as integrated visualization, DEM could be a good tool to detect signals in a fast and efficient way.

EXAMPLES

In this chapter we will introduce 3 examples following DEM model, which are generated by SAS® Version 9.3, and in which including both clinical data and operational data. Those 3 examples focus on top 5 query disposition, issue severity and difficulty classification in domain, singularity lab normal range check.

TOP 5 QUERY DISPOSITION

The operational data includes enrollment, query status, SDV etc., the traditional figures of operational data is bar chart or line chart which includes only two or three dimensional information. However, towards introducing DEM model the new graph is built based on HRDI and T&M, from which four or five dimensional information could be shown in one figure, and considered as an integrated information of all trials and site performance overview compare with KPI, also as a trend for ongoing study and perspective instruction to site and CRA. For example, in most of the studies below, CRF A has more issues than other CRF pages which need to raise attention to 1) retrain CRA on this page; 2) maybe CRF completion guideline is not clear in this page; 3) the CRF design for the CRF A page is not good enough to reduce the query rate, etc.

The figure below shows the TOP 5 query disposition.

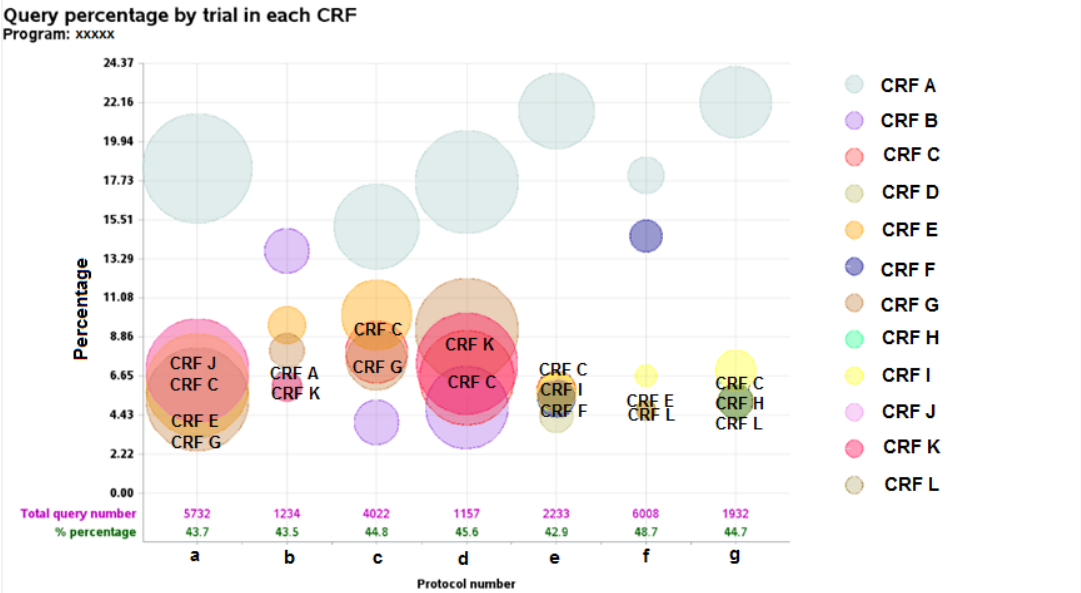

Figure 2. Query percentage by trial in each CRF

In the figure above, the x-axis is trial number within one compound/indication, y-axis is the percentage of the query rate of each CRF page. Each bubble represents a unique CRF page in each study with top 5 query rate, the different color of the bubbles shows different CRF page name, and different size of the bubbles shows the absolute query number of each CRF page. The total query number and percentage of top 5 queried CRF page are annotated at the bottom of this figure. As mentioned previously, from this figure we could easily see that most of the studies with the biggest number of query rate in CRF A, what's more, the distribution of top 5 CRF pages with high query rates could be identified in each single study, for example, in study b, the top 5 CRF page provides 43.5% query rate and all the CRF pages with similar query contribution. However, in study e, the top 5 CRF page provides 42.9% query rate but the CRF A page contributes queries far more than others.

There are some SAS code hints which considered important in this figure. For example, we had used annotation of SAS/GRAPH to draw the bubbles, which could be changed in size and color to show the different dimensional information together.

The SAS code of annotation is as below:

```
data anno_data;
 set final;
 length function $8 color $15 style $35 text $100;

%macro fm(fm=,k=,c=,label=);
 hsys='3'; when='a';
 function='pie'; style='psolid'; rotate=360;
 xsys='2'; x=count;
 ysys='2';

 if &fm.^=. then do;
       y=&fm.;
       if num&k. gt 50000 then size=9;
       else if num&k. gt 10000 then size=8.5;
       else do;
             if protn="xxxxx" then size=(sqrt(num&k./3.14)/sqrt(10000/3.14))*9;
             else size=(sqrt(num&k./3.14)/sqrt(6000/3.14))*9;
       end;
       if size<1 then size=5.5;
       size=2;
```

```
      if not missing(flg&k.) then do;
            y=y+flg&k.; function="label"; text="&label." ;htext=1;ftext="Albany
amt/bold";color='black';
            output;
            text="";h=.;html="";htext=.;ftext="";
      end;
 end;
%mend fm;


 %do e=1 %to &nofm;
      %fm(fm=%str(&&frm&e..),k=&e.,c=%str(&&c&e..),label=%str(&&slbl&e..));
 %end;

 function='label';  position='5'; size=.;
 ysys='3';
 y=15; text=trim(left(put(numt,best.)));color='darkpurple'; output;
 y=12; text=trim(left(pct));color='darkgreen'; output;
run;
```

However, some bubbles may overlap each other which caused difficulty in recognize, one solution is as the example shown just to annotate the CRF short name on the figure to distinguish the overlapped bubbles, the other option is to use different symbols (bubble pattern) to identify the bubbles. In this data step we can create an annotated dataset contains the pie function to draw bubbles on the graph page. In column style normally we could use PSOLID for circle filled or PEMPTY for circle, and we could also use PxNy to draw bubble shadow pattern (x stands for density of pattern and y stands for the angel of pattern). Below are the detailed information and example of PxNy.

Pdensity<style<angle>>[1]

a shaded pattern:

- density can be 1...5
- style can be X | N
- angle can be 0...360

```
data anno_data1;
 length function $8 style $35;
 function='pie';  size=1.5;
 hsys='3'; when='a'; rotate=360; xsys='5'; ysys='5'; x=79;
 y=21; style='P5N180'; output;
 y=27; style='P5N90';  output;
 y=40; style='P5N105'; output;
 y=80; style='P5N15';  output;
run;
```

In figure 3 you will see that for the elements of CRF C, CRF I, CRF K and CRF L are drew by PxNy in style column.

Figure 3. Examples of symbols draw by PxNy

ISSUE SEVERITY AND DIFFICULTY CLASSIFICATION IN DOMAIN

Some domains' data are very important safety data and some are very important efficacy data, so it is quite essential to check those domains without serious issue before database lock. In the past, many checks had been ran in those important domains to check the data issue in single data point or cross check the data points against other pages' items, even for some visualizations in those domains developed recently are also based on the two points above. So, to quickly have an overview on the whole domain's data in order to identify the data quality is good or not is not easy for the traditional data review. In this new figure developed towards DEM model (HRDI and T&M), we could have a quick overview on all domain's data to see if there is any risk on domain data quality, is there any improvement room of domain data quality and how to evaluate the data cleaning work on domain data. Furthermore, this figure is also a dynamic one which could be run among different data extracts to have the trend of each specific domain's data quality in order to support the decision making.

The figure below shows the distribution of domain in two dementional (Severity and Probability of Failure).

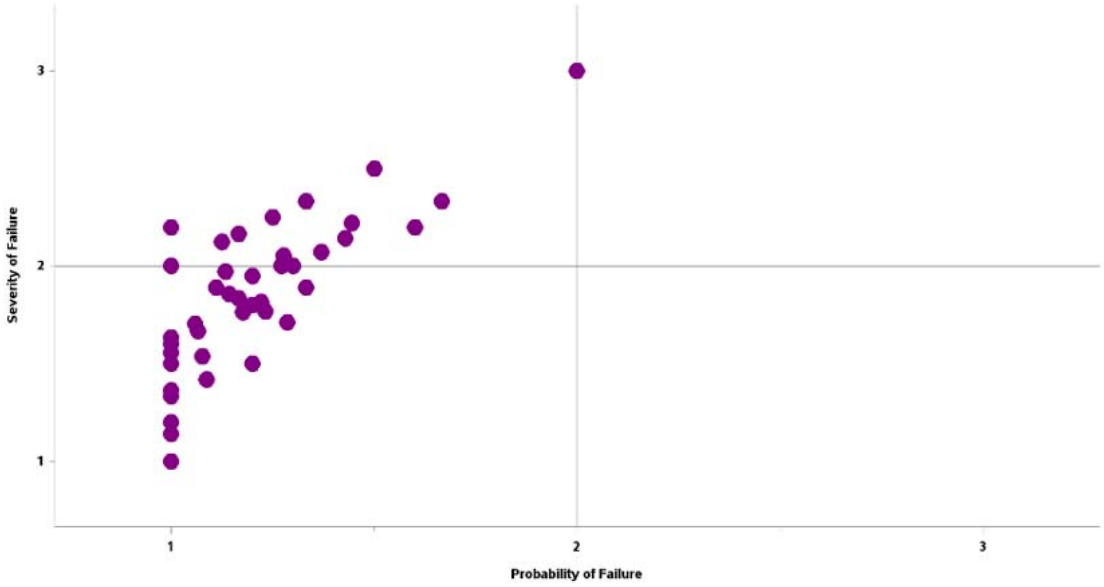

Figure 4. Issue severity and difficulty classification in Domain

Each domain observation can be a dot in this figure. The x-axis is the probability of failure, from 1 to 3 stands for low probability of failure to high probability of failure. The y-axis is the severity of failure, from 1 to 3 stands for low severity of failure to high severity of failure. Towards this classification the area is divided into four parts which stands

for high severity and low probability of failure, high severity and high probability of failure, low severity and low probability of failure and low severity and high probability of failure, each domain's observation dot will be showed in different areas based on the real status calculated. Thus first all, there should be a detailed listing which defines the classification of severity and probability of failure for each issue before drawing this issue map.

The example shows most of the dots concentrated in the field of low severity and low probability of failure, and parts of the dots concentrated in the field of high severity and low probability of failure which means an improved space for this domain's checking. What's more, while calculated and printed with time moving forward, the trend of issues could also be identified which is not showed in this paper.

The SAS programming of this figure is to use scatter plot in SAS/GRAPH or ODS/GRAPHICS for drawing it. However, before programming, the figure specification should be carefully designed and reviewed. The mandatory elements in the specification are: check description/logic, check item, severity score and probability of failure score, which allows programmers easily consolidates all the checks' result based on each record to calculate the mean score of severity and probability of failure for each record. In figure 5 shows an example of AE domain's specification, in which the "Description of the failure/risk" is the check/logic description, which could be single data point check (eg. missing AE reported term) or the cross check with other data points (eg. AE outcome is change in toxicity but there is no follow-up AE recorded). The second column is the checked item which means if the AE record is failed on this check, which item should be calculated for score in the next two columns (probability of failure and severity of failure). The programmer need to implement SAS code based on the check description against AE data, and collects all the failure items to calculate the average scores of probability of failure and severity of failure on each AE observation.

Description of the failure/risk	Checked item	Probability of Failure (1-3)	Severity of Failure (1-3)
Missing AE	AETERM	2	3
UNK start or stop dates	AESTDTC, AEENDTC	2	1
AE is change in toxicity and no follow up AE is entered	AEOUT	2	2

Figure 5. Specification Example

DISTRIBUTION OF LAB ORIGINAL UNITS

In current clinical studies, laboratory data plays a quite important role in efficacy and safety analysis, so the data quality of laboratory data is also a very hot topic. Normally in clinical studies, we can choose either central lab or local lab to collect lab results used for analysis. In local labs, the result unit can be diversity based on different lab center. Most of the checks run on the local lab original units are a kind of listing for manual review, which is time consuming and also increasing the risk of making mistakes to result in poor data quality. To overcome this question, we had introduced DEM model (HRDI, E&K) to develop a new figure in order to quick identify the issues in the local lab original units. The idea of this figure is not the traditional way to think from the perspective of sponsor or CRO, but a novel way from the perspective of the person who conduct the clinical trial (investigator). The traditional way always check the data point from either logic or compliance aspect, which is effective and easy for execution. But some data points could not get into such logic for checking, therefore, we need to think in another way which is from the investigators' perspective to execute the check. There are several characteristics during the choice made by investigators. First, the investigators' choice not always accurate, sometimes they may make some mistakes by accident which is inevitable. Second, majority of the choice made by investigators are correct and objective, so only a small amount of choice need to be reviewed and double confirmed. Third, as the incorrect choices are made by accident so they are not in a structural or logical way. After having those in mind, we can develop a figure to check the lowest frequency of the lab original units because they stands for the unconscious mistakes made by the investigators.

The figure 6 below shows the distribution of lab original units, the absolute count number of unit in 1 or 2 is highlighted in red in this figure allows user to focus on those parts which is high potential of issues.

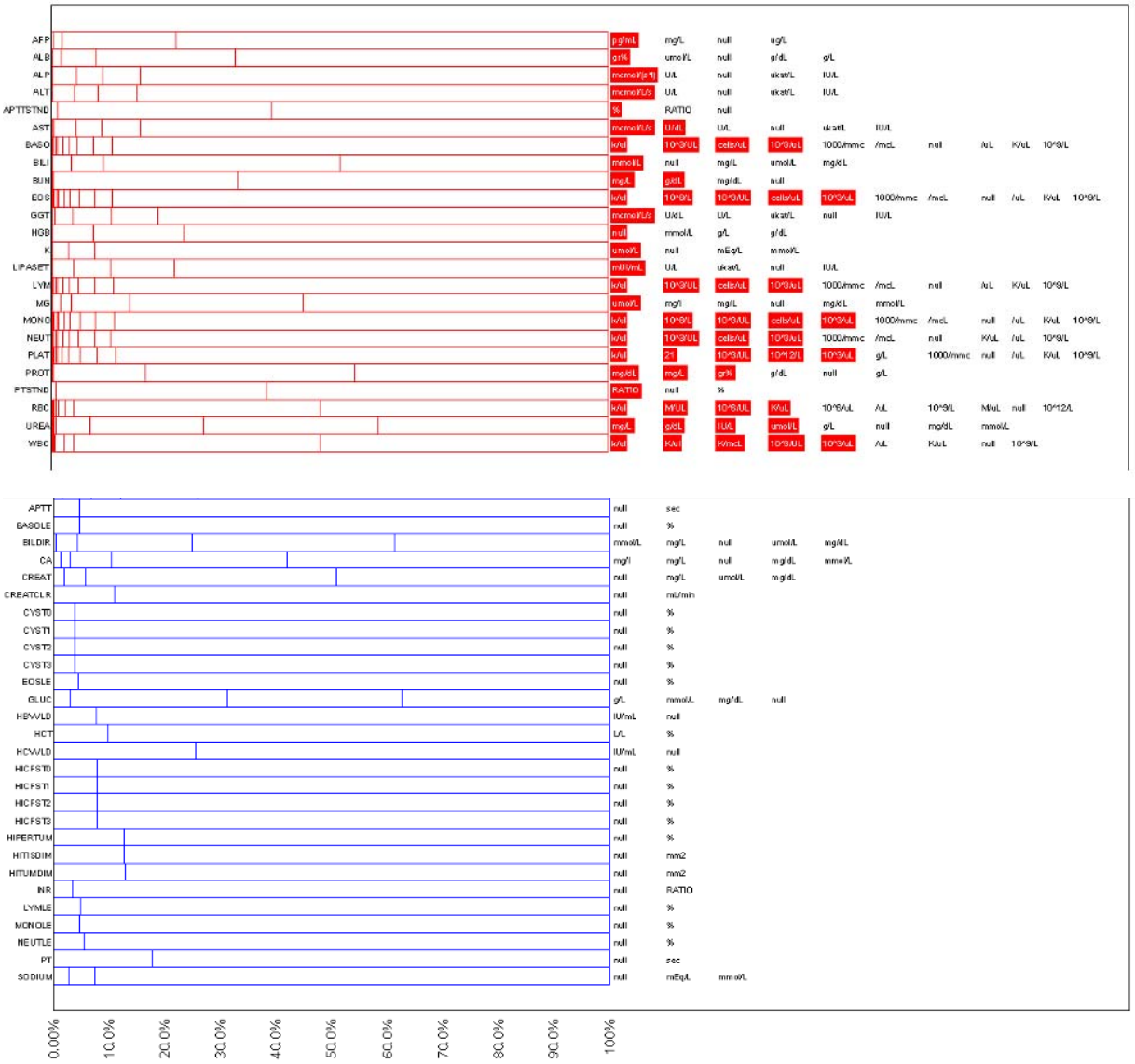

Figure 6. Distribution of Lab Original Units

Each lab test can be a bar in this figure. The x-axis shows different percentage of the lab original units. Detailed information of lab original units are annotated in the right part of this figure. The figure is divided into two parts, the first one is in red which shows lab tests with high potential issues in the original units, and the units recommended to be checked are highlighted in red; the second part in blue is the other lab tests which is with low risk than red ones in lab original units issue. The lines in the bar shows different percentage of the lab original units. Normally, in each local lab for one test, one consistent lab original unit should be used, so the lab original units could be a kind of reasonable distribution, the extreme values of lab original units count (eg. absolute count is 1 or 2) is considered to be an issue as they are far away from the reasonable distribution.

The above figure 6 shows in test AFP, there are 3 different original units, the majority of the choose are concentrated on mg/L and ug/L only 1 unit is selected as pg/mL which is quite different from others and need further check and review on this data point because it is with high potential risk of data issue (eg. choose by mistake from the investigator). Although there are some cases that some rare units could be chose by investigators which result that not all lab original units absolute count is 1 or 2 are all real issues; but most of the cases are, since the design of each local lab is to use the consistent unit to make it possible to see the value changes among different time points and

also the human behavior is always to be consistency instead of diversity. So this figure can help a lot in data review when the traditional logical or compliance way could not works.

Below are some SAS code hints which considered important in this figure. To use proc gchart to draw the bars and annotated the unit details at the right side of the graph. The code is as below:

```
%let c1=red;
%let c2=blue;
%let v=e;

pattern1 v=&v  c=&c1 repeat=135; /*use repeat to loop the pattern in two categories*/
pattern2 v=&v  c=&c2 repeat=62;

proc gchart data=graph_data_group anno=lbcd_lborresu_anno;
 hbar lbtestcd /  type=sum sumvar=pct
 subgroup=out_code  nozero nostats  /*use subgroup to draw lines in each bar*/
 group=color_type space=0          /*use group to divide the graph into two parts*/
 maxis=axis1 raxis=axis2 gaxis=axis3  name="&name"
 nolegend;
run;
```

The lab test short name, original units details are marked on the figure by using annotation. Below figure 7 is an example of the annotation dataset.

function	color	text	hsys	i	j	xsys	x	ysys	y	position	size	lbtestcd	cbox
label	white	pg/mL	3	51.818181818	93.303018868	3	52.818181818	1	96.803018868	6	0.6	AFP	red
label		mg/L	3	56.636363636	93.303018868	3	57.636363636	1	96.803018868	6	0.6	AFP	
label		null	3	61.454545455	93.303018868	3	62.454545455	1	96.803018868	6	0.6	AFP	
label		ug/L	3	66.272727273	93.303018868	3	67.272727273	1	96.803018868	6	0.6	AFP	
label	white	gr%	3	51.818181818	91.576037736	3	52.818181818	1	95.076037736	6	0.6	ALB	red
label		umol/L	3	56.636363636	91.576037736	3	57.636363636	1	95.076037736	6	0.6	ALB	
label		null	3	61.454545455	91.576037736	3	62.454545455	1	95.076037736	6	0.6	ALB	
label		g/dL	3	66.272727273	91.576037736	3	67.272727273	1	95.076037736	6	0.6	ALB	
label		g/L	3	71.090909091	91.576037736	3	72.090909091	1	95.076037736	6	0.6	ALB	
label	white	mcmol/(s*l)	3	51.818181818	89.849056604	3	52.818181818	1	93.349056604	6	0.6	ALP	red
label		U/L	3	56.636363636	89.849056604	3	57.636363636	1	93.349056604	6	0.6	ALP	
label		null	3	61.454545455	89.849056604	3	62.454545455	1	93.349056604	6	0.6	ALP	
label		ukat/L	3	66.272727273	89.849056604	3	67.272727273	1	93.349056604	6	0.6	ALP	
label		IU/L	3	71.090909091	89.849056604	3	72.090909091	1	93.349056604	6	0.6	ALP	

Figure 7. Example of Annotated Dataset

In which the value in function variable is “label” stands for using the label function to annotate the text value on the graph, in text variable contains the annotated values, and xsys, ysys and hsys defines the coordinate system in the x, y and size. The details of those three variable values could refer to the instruction from the SAS/GRAPH Reference below, figure 8.

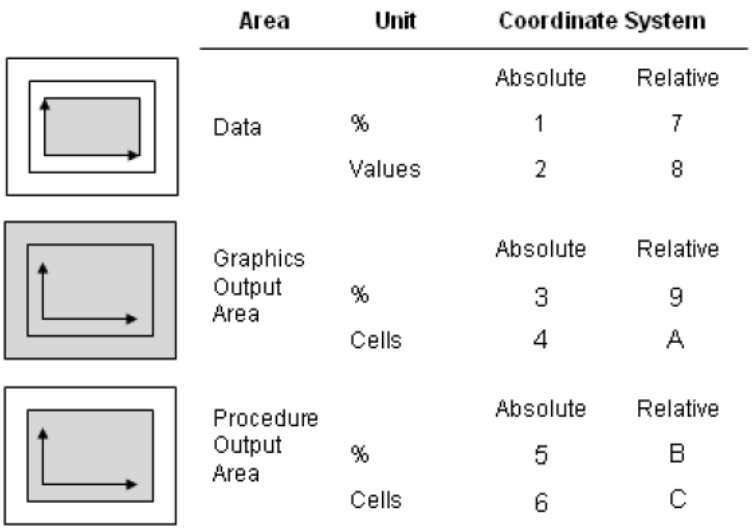

Figure 8. Areas and Their Coordinate Systems [1]

CONCLUSION

While using the DEM model, many new figures can be developed as above examples showed, which can be very helpful. Still we need to further think about how to improve the DEM model to make it more powerful. First, we can enlarge the existing elements in the DEM model to develop more and more innovative figures. For example, adding more information to the pool of tool and models, increasing the experience and knowledge, and identifying more high risk data points. Second, as shown in figure 1, DEM model is not only including those two strategies but also with other undeveloped areas. Especially the cross center of HRDI, T&M and E&K, as it could consolidate many information together so that the figures developed in this field could have combined merits of essential, target driven, innovation and user friendly. In future, new tools and technologies could be introduced into the DEM model to make it more powerful. For example, we could use other visualization tools such as SAS Visual Analytics or Spotfire to develop new figures or even we could use big data analysis as a new research method to develop figures based on those tools and technologies.

Last but not the least, the current DEM model is based on elements of HRDI, T&M and E&K, across those we could have develop the visualization figures in a novel way. Besides those three elements, there are more elements can be considered and introduced into this model. Through introducing new elements, the DEM model can bring more values in the visualization innovation, which shows strong ability in integrated information review, fast issue identification and perspective decision making.

REFERENCES

1 <Author name: Cary>. <2012>. <*SAS/GRAPH® 9.3: Reference, Third Edition* >. <2082>. <NC>: <SAS Institute Inc.>

ACKNOWLEDGMENTS

This paper is produced under guidance of Yongjian Zhou, who is lead clinical data manager in Merck, and many ideas in this paper are coming from the data management team in Global Clinical Data Sciences in Merck, the programming efforts are paid by the programming team within Global Clinical Data Sciences in Merck. Thanks them all to the contribution of making this paper.

CONTACT INFORMATION

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

Name: Teresa Tang
Enterprise: Merck
Address: 21F Nuo Center, No. 2 Jiangtai Road, Chaoyang District
City, State ZIP: Beijing, 100016
Work Phone: +86 10 5903 1448
E-mail: Teresa.tang@merckgroup.com