PharmaSUG 2017 - Paper DA02

Let's Check Data Integrity Using Statistical (SAS®) Programmers with SAS®

Harivardhan Jampala, Chiltern International

ABSTRACT

The success of any clinical trial depends on the accuracy and integrity of the study conduct and the data produced from the trial. As in any experiment, data plays the central role and almost everybody involved in a clinical trial generates, maintains, or explains data. Hence, we can all agree: It is vital that the data is clean and, more importantly, that it is fully utilized to make everyone's job easier and efficient.

A plethora of software, programing languages, and tools are employed by various contributors in clinical research across the industry to help make sense of clinical and operational data. The scope of and investment in such tools depends on the budget and organizational priorities. There are organizations that operate with moderate or minimalistic software resources. This paper will explore various ways in which SAS programmers (either statistical programmers or clinical programmers) can help other departments in such organizations.

The paper will list some examples of such offerings, sometimes with a sample approach, like, data listings, CRF tracking metrics, patient profiles, and site summary metrics; patient profiles with specific data points to help CRAs; patient safety summaries; and custom safety narrative templates. This paper also deals with some of the graphical representation of the data before the actual statistical programming starts for tables, listings, and figures. It also describes unique ways for summarizing clinical trial data to make it easy to spot errors in data about individual subjects or clinical sites.

INTRODUCTION

Clinical research is a collaborative process where various departments with varied skills and responsibilities work together with a common goal. This collaboration culminates in a successful, quality submission with biostatisticians and statistical programmers playing a key role in summarizing the reports reviewed by regulatory authorities. The quality of the data-driven decisions is limited by the quality of the data. If your data have errors, your decisions will be error prone as well. Detecting inadvertent errors and fraudulent data is paramount at every step. In general, CROs or sponsors often hire people to handle data entry and analysis, but how can you tell if you have the right staffing, especially with the type of reports that are being developed to see data more clearly? Can the database programmer produce these reports? Or, should we go to a statistical programmer who can utilize a gamut of capabilities in SAS? What about a useful software that isn't available because of financial limitations? One of the best ways to ensure data integrity is to create the reports that are discussed in this paper with appropriate data management plans from your clinical data management team with the help of a statistical programmer.

The idea is to use statistical programmers in addition to the database programmer in the initial stages of trial.

Why is it important? Data integrity is what enables organizations to get a clear picture of the trial, which, in turn, makes decision making efficient.

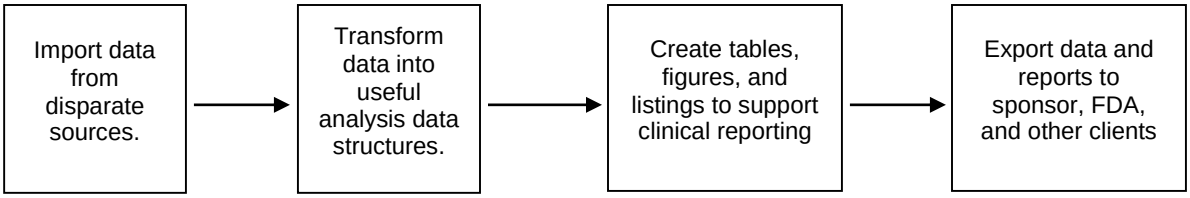


Figure 1. Following the Data Trail and Traditional Statistical (SAS®) Programmer Role: From the book SAS Programing in pharmaceutical industry.

WHAT ARE SOME OF THE THINGS THAT STATISTICAL (SAS®) PROGRAMMERS CAN OFFER

Quality of Data - Clinical Data Management

- CRF completion/tracker
- Site summary and query rate

Patient Safety and Efficacy - Drug Safety

- Safety review report
- RECIST listings

Decision-Making Visuals by Statistical Programmer (web-based reports) - Clinical Operations

- Simple demographics
- Sites with AE/SAE count
- Complete site analysis
- Laboratory data
- Patients meeting inclusion/exclusion criteria

Consumers of these reports could be data managers, biostatisticians, medical writers, clinicians, CRAs, or other decision makers on the sponsor's team. These reports are more than the DVS listings, edit check programs, or patient profiles/narratives that a programmer is doing on regular basis.

QUALITY OF DATA

CRF Completion

The earlier the missing CRF pages report (Display 1) is created and implemented, the sooner we will be alerted to any problems and patterns in the data collection process. Additionally, regular and thorough review of this type of report results in a more accurate trial database, a shorter time between the end of a trial and a final database and, ultimately, a higher quality trial.

Site Id	PI Name	Subject ID	Visit ID	Visit Name	Missing Page ID	Missing Page Name
1	Dan	1001	70	Cycle 2 Day 1	90	Maintenance Hydration
1	Dan	1001	80	Cycle 2 Day 2	50	Maintenance Hydration
1	Dan	1001	90	Cycle 2 Day 3	50	Maintenance Hydration
2	Ren	1002	90	Cycle 2 Day 3	70	DCE-MRI
2	Ren	1002	130	Cycle 3 Day 3	50	Maintenance Hydration
2	David	1003	240	Cycle 6 Day 2	10	Patient Visit
3	David	1003	240	Cycle 6 Day 2	20	Vital Signs
3	David	1003	250	Cycle 6 Day 3	10	Patient Visit

Display 1: CRF Missing Pages

The specification from the data management team may look like this example in Table 1.

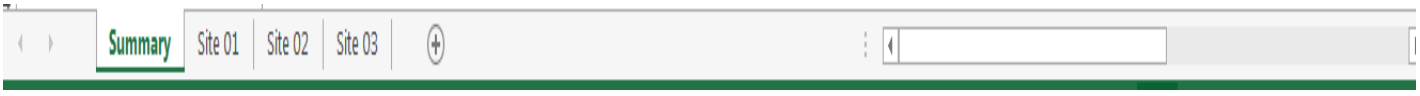
80	Cycle 1 Day 2	10	Patient Visit	Flag as missing if not present and current date is > 1 day after Cycle 1 Day 1 visit date (SV.SVSTDT where SV.VISITNUM = 20). Do not flag as missing if Conclusion of Subject Participation is entered and DS.DSSTDAT is less than the calculated visit date for C1D2 (DS.DSSTDAT<(SV.SVSTDT+1 where SV.VISITNUM = 20))

Table 1. Example of Specification

Site Summary and Query Rate

This report will give us the metrics and the status of a subject, e.g., whether the subject has entered the study and what the current CRF status is for a site or a subject, like pages received, entered, reviewed and cleaned, queries open, queries closed, etc. See Display 2.

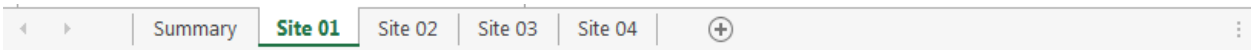
Site	PI Nar	Total Subjects Enrolled in EDC	CRFs Expected	CRFs Missing	Total CRFs Entered	Total CRFs Completed	Total CRFs Monitored	% CRFs Monitored	CRFs to be Monitored	Total CRFs DM Reviewed	% CRFs DM Reviewed	Total CRFs Frozen	% CRFs Frozen
01	Ren	11	3669	194	3475	3475	3459	94	210	3444	94	3415	93
02	Dan	5	1043	67	976	972	951	91	92	902	86	889	85
03	David	8	1857	68	1789	1750	1702	92	155	1688	91	1641	88
04	Roony	1	192	3	189	189	189	98	3	187	97	178	93
	TOTAL	25	6761	332	6429	6386	6301	93	460	6221	92	6123	91



Columns continue.

Total CRFs Signed	% CRFs Signed	Open Queries	Open more than 25 days	Answered Queries	Closed Queries	Total Queries Issued

Site ID	PI Name	Subject ID
'01	Ren	1001
'01	Ren	1002
'01	Ren	1007
'01	Ren	1008
'01	Ren	1010



Display 2. Site Summary Report

We can produce this type of reports by sponsor and study. An example of this can be seen in Table 2.

Sponsor	Study	NA	No Data	Entered	Completed	Monitored	Issue	DM Reviewed	Frozen	Signed	Comments
Sponsor1	Study	0	1216	85	36	0	0	2	41	0	
Sponsor2	Study	15	804	43	1584	63	0	133	2972	0	
Sponsor3	Study	0	1194	179	1904	3	0	49	2943	0	

Table 2. Site Summary Report by Sponsor

PATIENT SAFETY AND EFFICACY

Safety Review Report

This report will help us identify all safety information in a single shot for each subject. Similar reports can be used across various studies. See Display 3.

<Let's Check Data Integrity>, continued

PATIENT

Patient # / Initials	Dosing Group	Date ICF Signed	Date of Screening Visit	DOB	Sex	ECOG at Screening	ECOG at C1D1
01001/R-V	SAFETY RUN-IN PHASE	14FEB2013	14FEB2013	10DEC1940	MALE	0	0

VITAL SIGNS

Visit	Date	Time	Sys/Dia	HR	Resp	Temp	Temp	Weight	Weight Unit	Height	Height Unit
Screening	14FEB2013	10:15	128/80	62	18	98	F	134	LB	66	IN
Cycle 1 Day 1 Pre-Dose	19FEB2013	07:55	138/90	62	18	97.6	F	132.2	LB		
Cycle 1 Day 1 Post-Dose	19FEB2013	13:25	118/72	66	17	97.8	F				
Cycle 1 Day 1 Post-Dose	19FEB2013	14:25	132/72	63	16	97.3	F				
Cycle 1 Day 1 Post-Dose	19FEB2013	15:25	120/68	56	16	97.7	F				
Cycle 1 Day 1 Post-Dose	19FEB2013	16:25	126/70	87	17	97.5	F				

12-LEAD ECGS

Visit	Date	Time	QTc	Method	Overall Interpretation
Screening	14FEB2013	10:51	379	MACHINE CALCULATED	NORMAL

PT INFO | LAB DATA | EVENTS | MEDS | STUDY DRUG | DEATH - SURVIVAL

Patient # / Initials - 01001 / R-V

	Screening			C1D1			C1D8			C1D15		
Laboratory Test Short Name	Result	Unit	CS - Y/N	Result	Unit	CS - Y/N	Result	Unit	CS - Y/N	Result	Unit	
CHEMISTRY												
ALB	3	G/DL	N	3	G/DL	N	2.8	G/DL	Y	2.8	G/DL	N
ALP	661	IU/L	N	661	IU/L	N	876	IU/L	Y	894	IU/L	Y
COAGULATION												
INR	1.2			1.2								
PT	12.2	SEC	N	12.2	SEC	N						
HEMATOLOGY												
BASO	0.2	OTHER:X10E6/UL		0.02	OTHER:X10E3/UL		0	OTHER:X10E3/UL		0.1	OTHER:X10E3/UL	
EOS	0.4	OTHER:X10E6/UL		0.23	OTHER:X10E3/UL		0.6	OTHER:X10E3/UL	N	0.4	OTHER:X10E3/UL	

PT INFO | LAB DATA | EVENTS | MEDS | STUDY DRUG | DEATH - SURVIVAL

Display 3. Safety Review Report

RECIST Listings

To verify when the response data is correct and to identify where the potential errors are, queries can be written and also used to identify areas that the CRAs might need retraining on. This can be a helpful as a real-time tool for the CRAs to use while monitoring. See Table 3.

Site ID	Subject ID	Primary Diagnosis	Histologic Diagnosis	Assign Dose Level	Completion/ Discontinuation	Visit Number	Visit	Date of Assessment	Lesion Type	Lesion Number	Lesion Site
1	001-01-1	OTHER: ESOPHAGUS	ADENOCARCINOMA	10	17-Jun-13	10	SCREENING	17-Apr-13	Target	1	LIVER
1	001-01-1	OTHER: ESOPHAGUS	ADENOCARCINOMA	10	17-Jun-13	10	SCREENING	17-Apr-13	Target	2	LIVER
1	001-01-1	OTHER: ESOPHAGUS	ADENOCARCINOMA	10	17-Jun-13	10	CYCLE 2	17-Apr-13	Target	1	LIVER

Columns Continue.

Method of Assessment	Non-Target Lesions Status	Lesion Measurement	Total Dimension of Lymphoma	Sum of Lesion Measurements or Products	% Change from Baseline	% Change from Smallest Sum	Overall Responses	Response Description

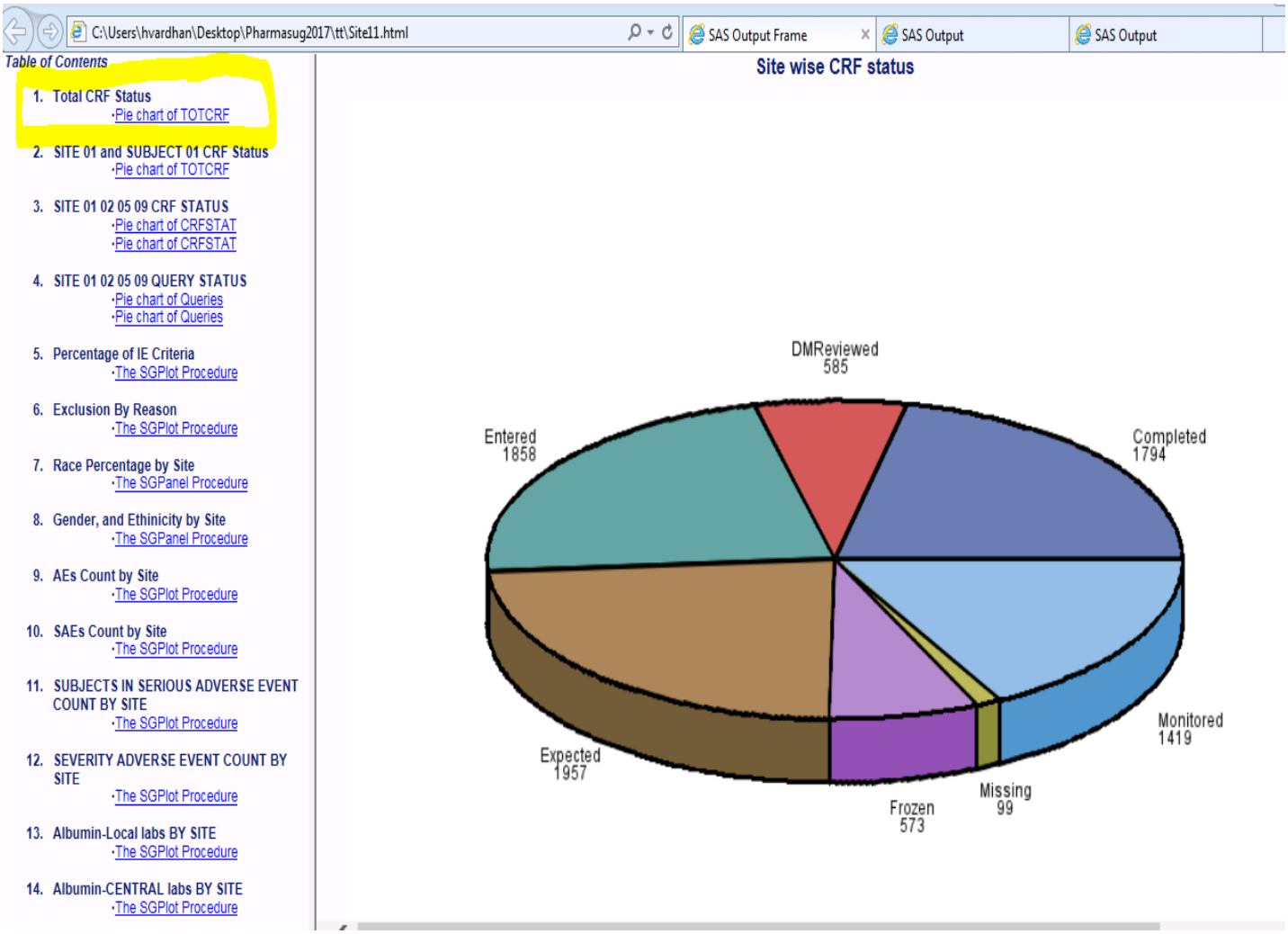
Table 3. RECIST Listings

CODE FOR ALL THE ABOVE REPORTS: The code for all these reports mostly uses ODS tag set or XML programing, which allows us to create multiple sheets for sites and subjects. However, the focus here should on ensuring correct patient data; the programing is an individual's own preference. Here is one option:

```
ODS TAGSETS.EXCELXP
    OPTIONS (sheet_name="&&site&nn"
absolute_Column_Width='8,15,10,10,10,10,10,10,10,10,10,10,10
frozen_headers='3' frozen_rowheaders="3");

Proc report;
 <CODE>
Run;

ODS TAGSETS.EXCELXP CLOSE;
ods listing;
```

CREATING DECISION-MAKING VISUALS BY STATISTICAL PROGRAMMER (SAS) (WEB-BASED REPORTS)

The consumers of these types of reports are CRAs, medical writers, clinical safety scientists, and physicians.



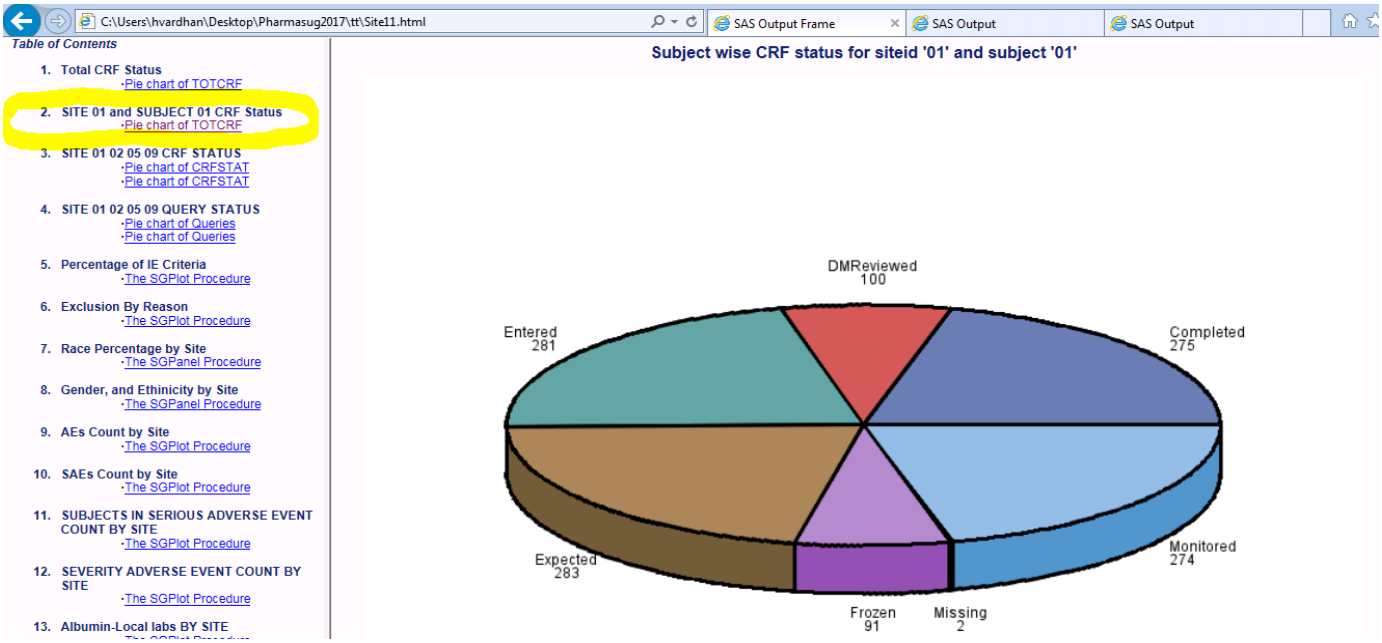
Display 4. Complete CRF Status of a Study

Here is an example of code that can be used to examine the complete CRF status of a study:

```
   filename odsout "C:\..\...\Desktop\Pharmasug2017";

ods html path=odsout frame="Site11.html"
   contents="Site_contents.html"
   body="Site_body1.html"
    nogtitle;
    title1 "SITE CRF STATUS";
    footnote j=r  "SITE INFORMATION ";
title1 "Site wise CRF status " h=9pt;
ODS PROCLABEL = "Total CRF Status";

<INDIVIDUAL PLOT STATEMENTS>
Proc SGPLOT;

<CODE>

Run;

Proc gplot;

<CODE>
```

```
Run;
<GTL STATEMENTS?
ods html close;
ods html;
```

Subject wise CRF status for siteid '01' and subject '01'



Display 5. CRF Status by Subject

Here is an example of a code that can be used to examine CRF status by subject:

```
proc gchart data=siteinfo;
  pie3d Site / sumvar=CRFSTATUS noheading woutline=2
               coutline=black discrete;
run;

QUIT;
```

Multiple Site CRF STATUS

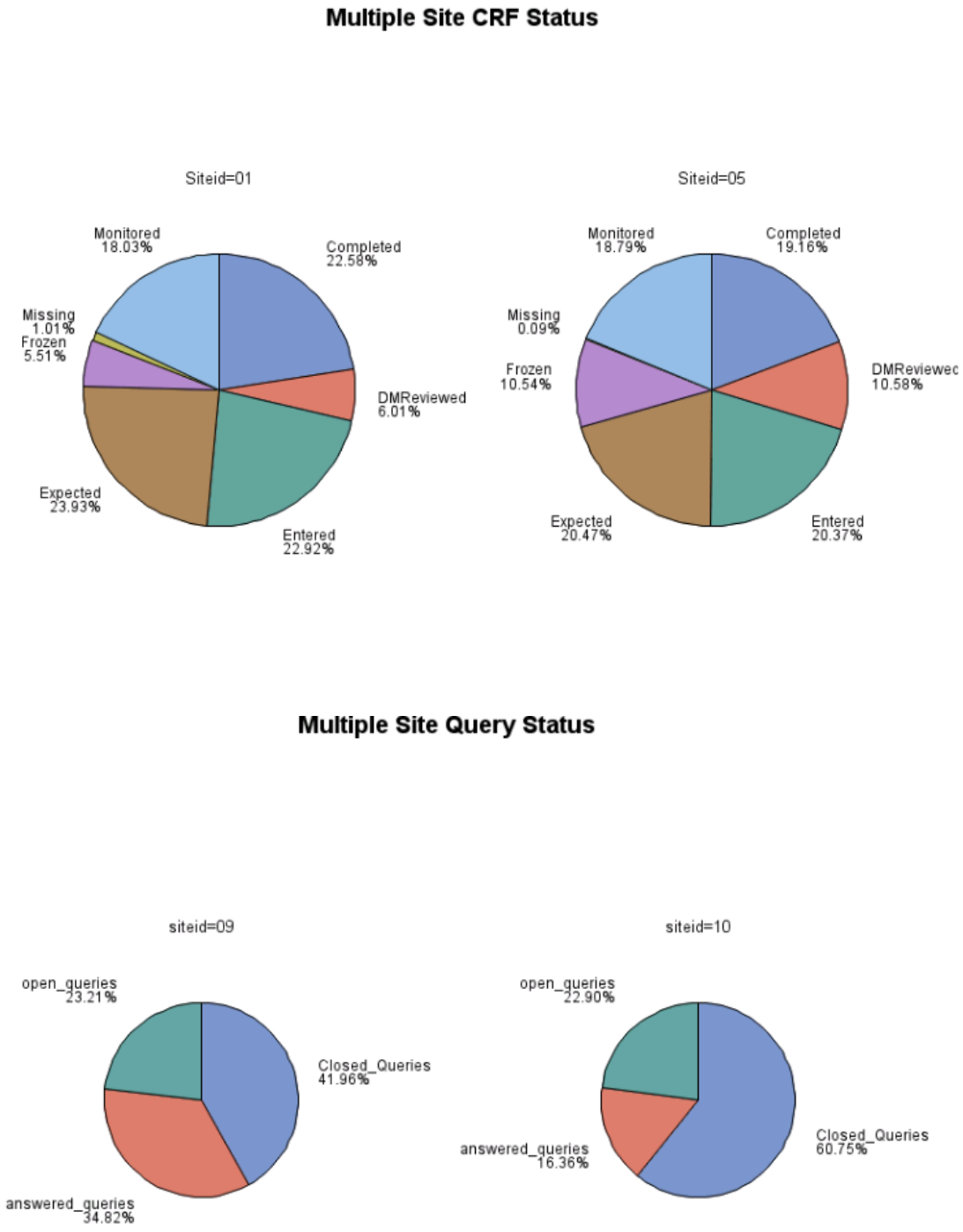


Figure 2. Multiple Site CRF and Query Status Information

The graphics in Figure 2 explain the details of the CRF status and query status for multiple sites at a glance.

Code for Figure 2:

```
proc gchart data= siteinfo;
   pie CRFSTAT / sumvar=SITE
                 other=5
                 otherlabel='Missing'
                 group=siteid
                 across=2
                 clockwise
                 value=none
                 slice=outside
                 percent=outside
                 coutline=black
                 noheading;
run;
quit;
```

ELIGIBILITY CRITERIA

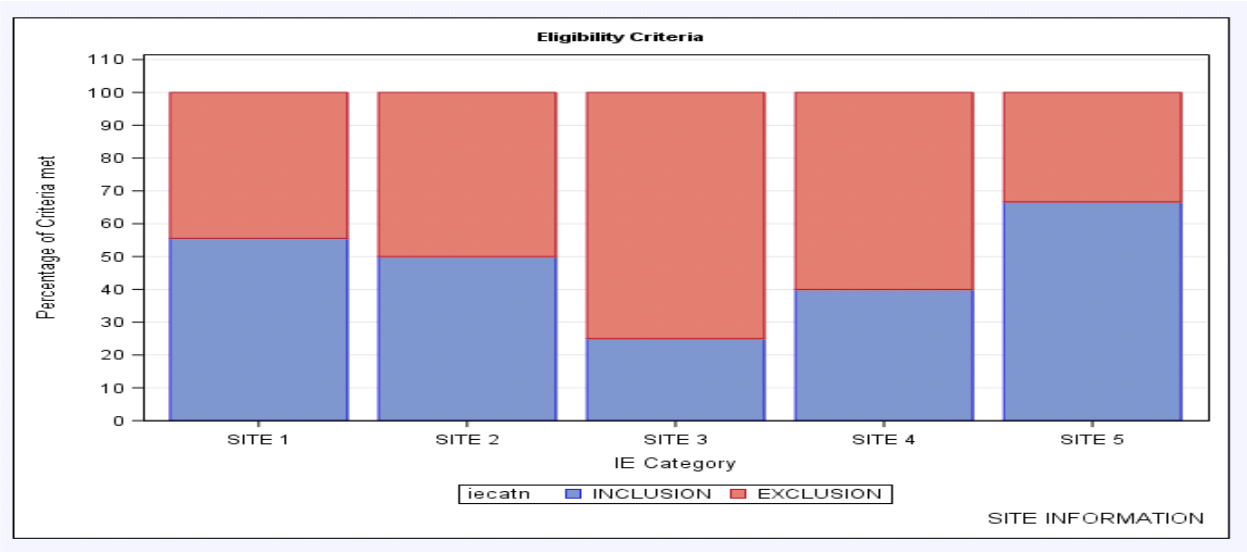


Figure 3. Inclusion/Exclusion Criteria by Site

It is important to note that inclusion and exclusion criteria are not used to reject patients personally, but rather to identify appropriate participants and avoid chances of higher patient-selection-related risks at the site. Figure 3 shows that SITE 3 has a higher exclusion rate and waivers are granted. Here is the applicable code:

```
proc sgplot data=siteinfo;
  format iecatn ie.;
vbar siteid / response=percent group=iecatn nostatlabel
       groupdisplay=stack
xaxis label="IE Category"
yaxis grid values=(0 to 110 by 10) label="Percentage of Criteria met";
```

```
run;
```

Figure 4 explains that the exclusion reason labeled as EXCL30 is the reason behind all the excluded subjects from SITE 3.

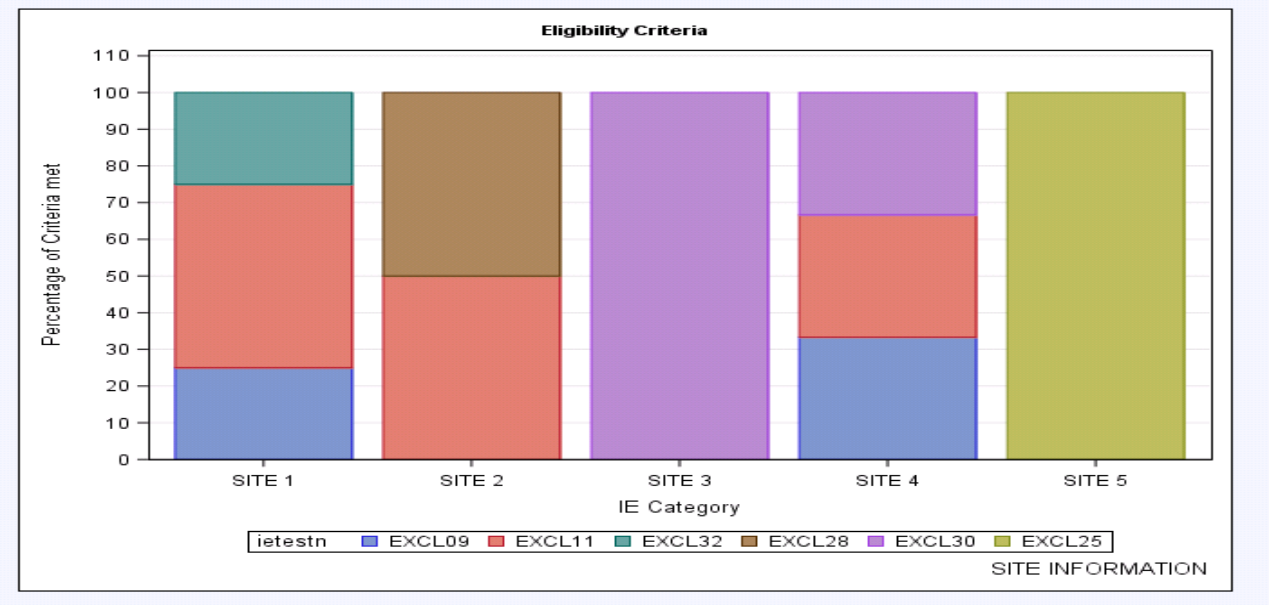


Figure 4. Exclusion Criteria by Site and Reason

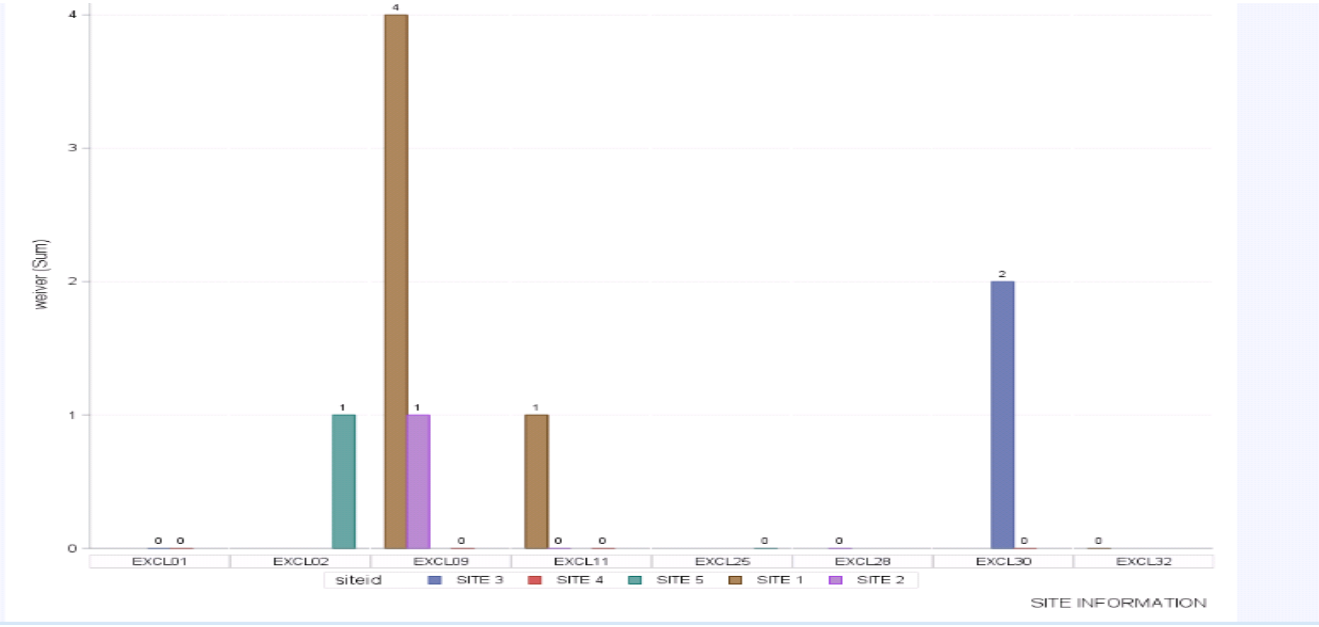
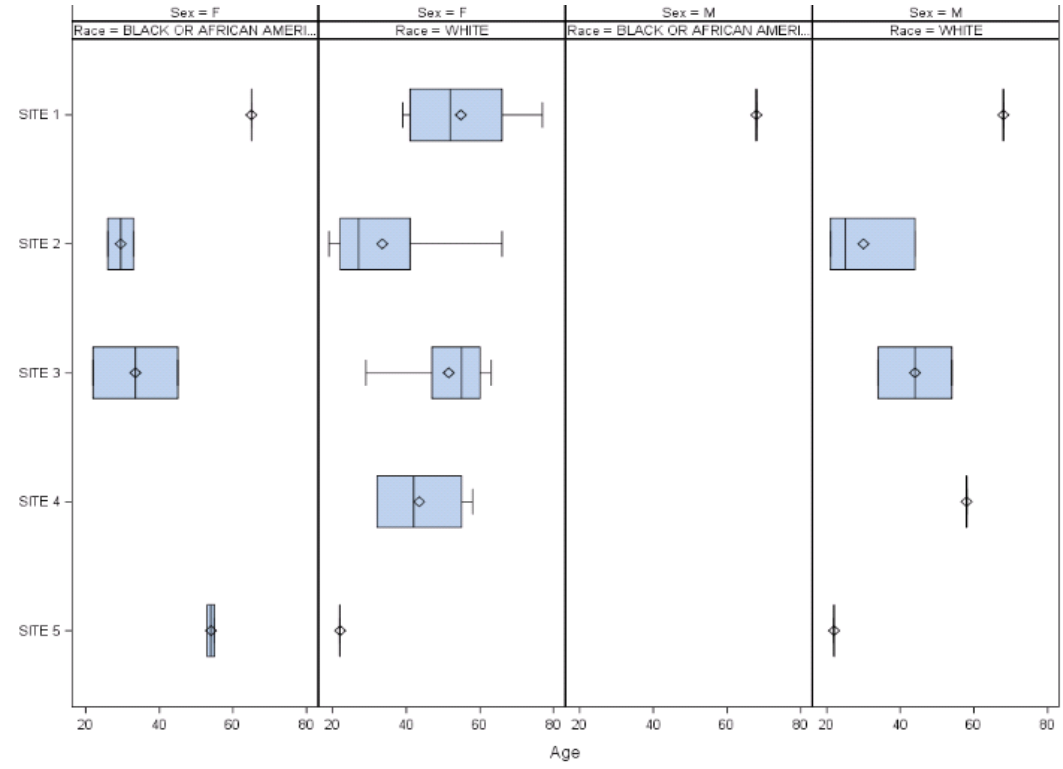


Figure 5. Exclusion Criteria With Waivers Count

Figure 5 explains that site 3 has two waivers granted for EXCL30.

DEMOGRAPHICS



Display 6. Demographics

To see simple demographics by gender, age, race, and site:

```
proc sgpanel data= siteinfo;
  panelby sex ethnic/
   layout=panel columns=4;
   hbox age / category=siteid;
    rowaxis display=(nolabel);
run;
```

ADVERSE EVENTS

Programmers can try multiple ways to show the AEs and SAEs, which, in turn, help medical writers, the safety department, and physicians. Some of the examples are mentioned.

Figure 6 explains which site has more AEs reported, with a reference line for 15 and 20 subjects.

Figure 7 explains which site has more SAEs reported, with a reference line for 5 and 2 subjects.

<Let's Check Data Integrity>, continued

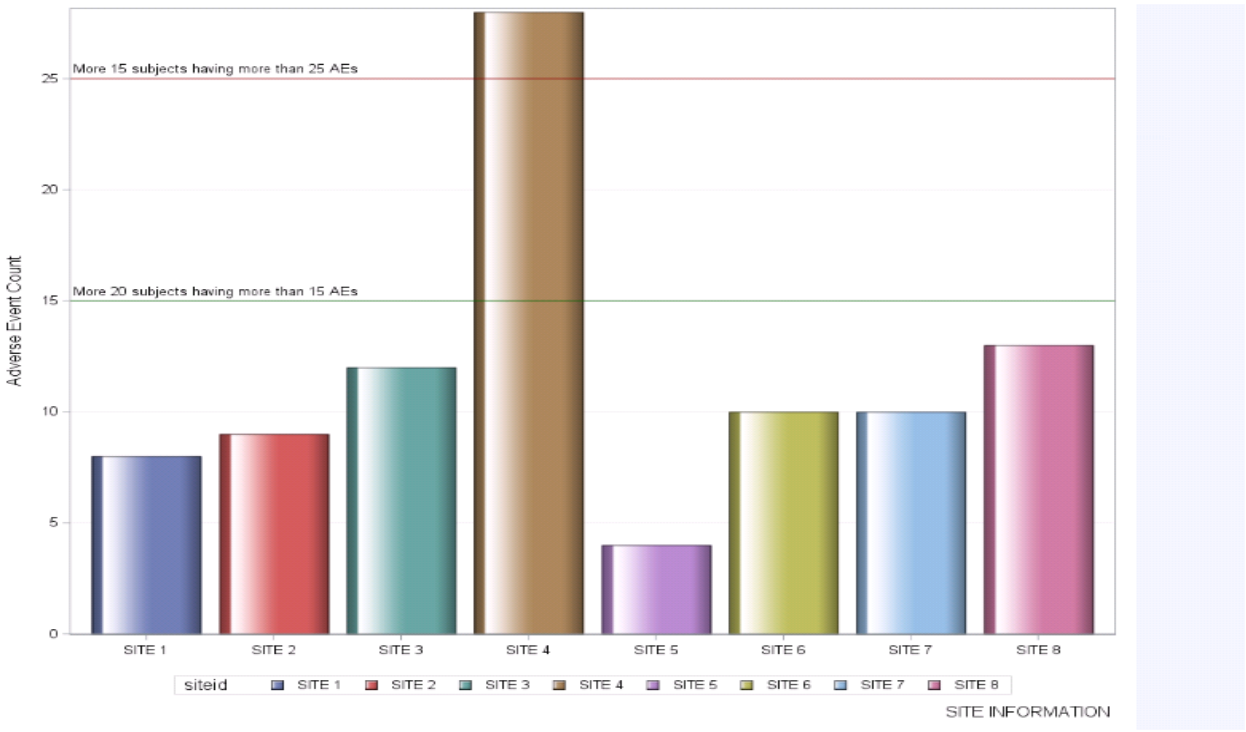


Figure 6. Adverse Event Count by Site

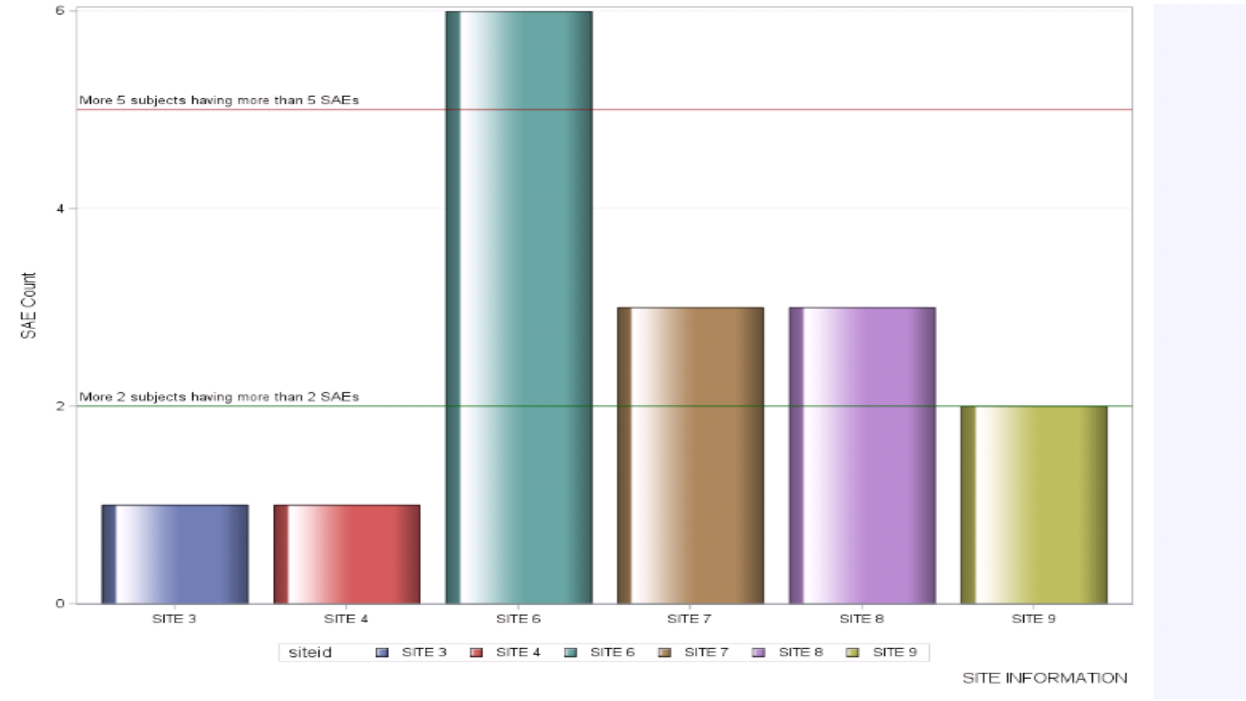
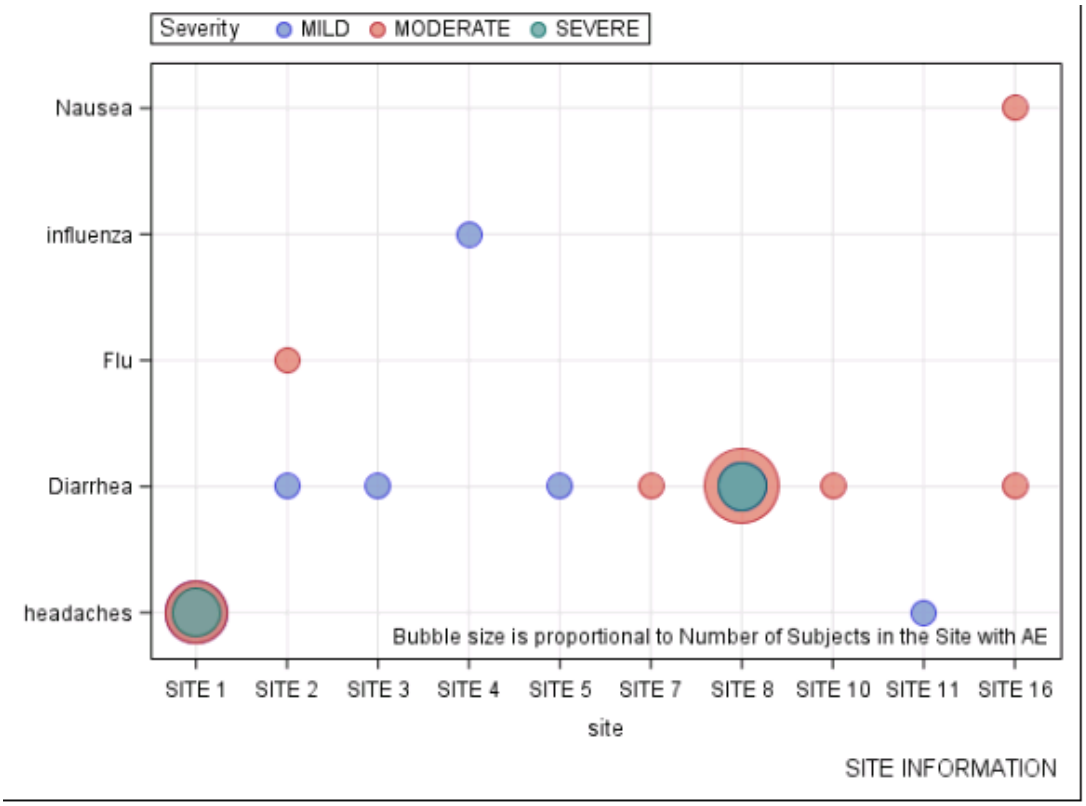


Figure 7. Serious Adverse Event Count by Site

CODE for Figure 6 (Same code can be used for Figure 7):

```
proc sgplot data=siteinfo;
vbar siteid / response=aenum stat=sum group=siteid nostatlabel
       groupdisplay=cluster dataskin=gloss;
 refline 15 / lineattrs=(color=darkgreen) label='More…… '
labelloc=inside labelpos=min;
 refline 25 / lineattrs=(color=darkred) label='More…..' labelloc=inside
labelpos=min;
xaxis display=(nolabel);
yaxis grid label="Adverse Event Count";

run;
```

Display 7 explains how the AEs are distributed by site and the size of the bubble indicates the number of subjects in the site with AE. We can see that SITE 1 has more subjects with headaches that are moderate and severe. We can create these types of plots for a single event and check the ratios.



Display 7. Adverse Events by Site and Severity

Here is an example of code that can be used to see adverse events by sites and by severity:

```
proc sgplot data=aebct;
format site. aesevn sev.;
bubble x=site y=aeterm size=aecount / group=aesevn transparency=0.2;
inset 'Bubble size is proportional to Number of Subjects in the Site with
AE' / position=bottomright;
```

```
xaxis grid type = discrete tickvalueformat = site.; yaxis grid
label="Subject count"; keylegend / title='Sevirity' location=outside
position=topleft;
run;
```

LABORATORY DATA:

Usually the labs data are collected by different labs and if we can show the difference between the ranges of results in those collections that will help in reconciliation of labs by data management.

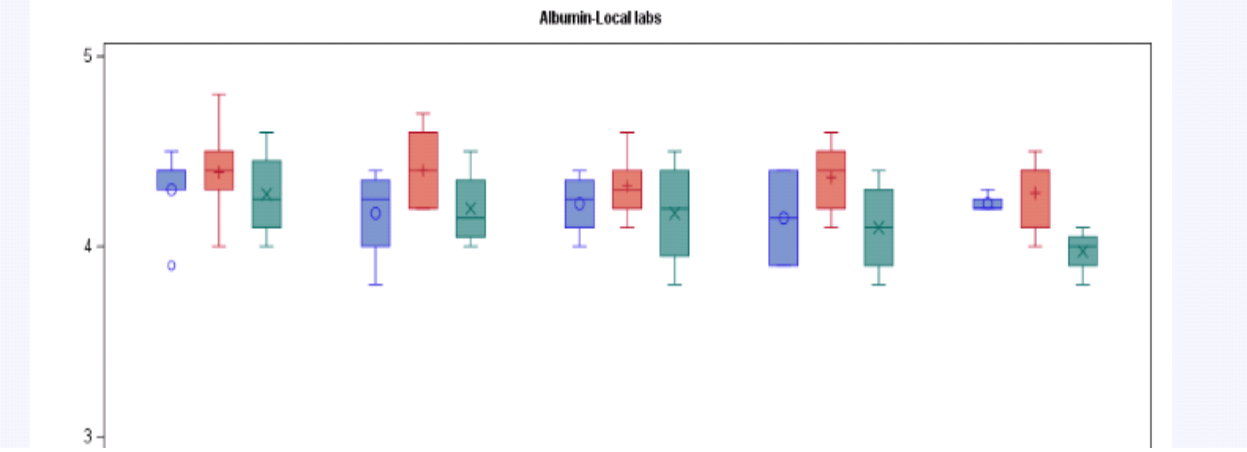


Figure 8. Lab with Local Values

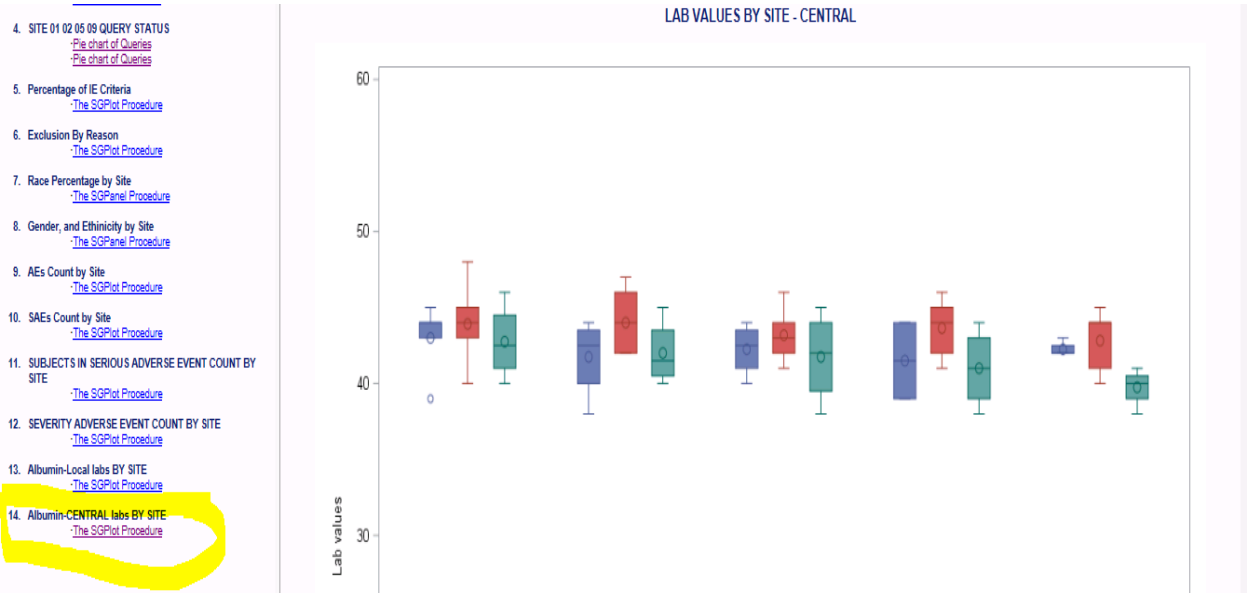


Figure 9. Lab with Central Values

Here is an example of code identifying lab values:

```
title1 "LAB VALUES BY SITE - LOCAL" h=9pt;
ODS PROCLABEL = "Albumin-Local labs BY SITE";

proc sgplot data = siteinfo;
```

```
    xaxis label="Visit" labelattrs=(size=9);
    yaxis  values=(0 to 5 ) label="Lab values"
                labelattrs=(size=9);
vbox  lborresn/  category  =  visitnum group=siteid;
keylegend/ title="Site:" position = bottom noborder;
run;
```

CONCLUSION

There are many software programs and programing languages available to check clinical data integrity like Tableau, Spotfire, JMP, Python, and SAS visual analytics options. But these are cost-intensive and time-consuming to learn. We know every company has an SAS license and every company has an SAS programmer(s). If we can use the knowledge of an SAS programmer using SAS, we might save a lot of time and money. The latest SAS version contains many possibilities to create all type of reports. I strongly believe that sponsor-level customizations can be done in SAS but not with other software programs.

REFERENCES

Matange, Sanjay. Google search for "Sanjay Matange SAS ODS SG Graphics" will produce a list of dozens of publications, too numerous to list here.

Matange, Sanjay; Heath, Dan. November 2011. *Statistical Graphics Procedures by Example: Effective Graphs Using SAS®.* SAS Institute.

"Example 8: Combining Graphs and Reports in a Web Page." S*AS Institute.* Available at http://support.sas.com/documentation/cdl/en/graphref/69717/HTML/default/viewer.htm#p0iexv7vslpxqqn1qrhrj6rcoys1.htm.

DelGobbo, Vincent. "More Tips and Tricks for Creating Multi-Sheet Microsoft Excel Workbooks the Easy Way With SAS®." *SAS Institute.* Available at http://support.sas.com/resources/papers/proceedings09/152-2009.pdf.

ACKNOWLEDGMENTS

I would like to thank my manager, Michael Wisniewski, in particular, who believed me and for providing encouragement and supporting PharmaSUG participation, and my friend and well-wisher, Sridhar Patel, who has immense patience to listen my ideas and provide his valuable feedback.

RECOMMENDED READING

- SAS® Programming in the Pharmaceutical Industry, Second Edition
- "Tips and Tricks for Clinical Graphs Using ODS Graphics." Available at http://jansenlex.readyhosting.com/pharmasug/2011/sas/pharmasug-2011-sas-ad01.pdf
- SAS Institute graphics resource. Available at http://support.sas.com/rnd/datavisualization/index.htm

CONTACT INFORMATION

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

Harivardhan Jampala
Chiltern International
Address: 4000 CentreGreen Way, Suite 300, Cary,
North Carolina, 27513, United States: +1 919 462 8867
Hari.Vardhan@Chiltern.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.