%DataAutoSum: A SAS® Macro for Automatically Producing a Customized Clinical Data Summary Report
Deqing Pei, M.S., Cheng Cheng, Ph.D.
St. Jude Children’s Research Hospital, Memphis, TN

ABSTRACT
Management of clinical trials data requires efforts of a large team, including clinicians, data entry personnel, database managers, and statisticians. For statisticians, to ensure the accuracy of the clinical data is the first critical step in analysis. No statistical method, regardless of its sophistication, can overcome major weaknesses arisen from inaccurate data. Results from analyzing inaccurate clinic data could be misleading and result in a wrong conclusion or decision. This paper introduces a macro, which can be used to generate data summary report with limited user input. Thus, it enables the statistician to check the quality of clinical data quickly and effectively.

INTRODUCTION
In the evaluation of any set of data, from whatever source, it is essential to begin with an investigation of the basic characteristics. These questions are in the realm of descriptive statistics. Clinical data are similar in forms to other data in that there are three types of variables: numerical, categorical or date. Numerical variables are numbers that can have any value within some acceptable range. The summary statistics (mean, median, max, min etc) of a numerical variable are helpful to identify any outliers or extreme observations. Frequency tables can provide valuable information on categorical variables. Earliest and latest date on each date-related variable and the number of missing values of each variable in the dataset are also important information to have for data quality checking. All information listed above could be generated in one simple step via calling the macro presented in this paper.

ARRAY SYNTAX

% DATAAUTOSUM (DS=,
DFROZEN=%STR(),
UNWANTNUM=%STR(),
UNWANTCHR=%STR(),
UNWANTDATE=%STR(),
DIRPATH=%STR(),
ID=%STR(),
NUM=,
PAGEF=%STR())

ARRAY PARAMETERS

DS: Provide name of the data set.

DFROZEN: Provide date when you received the data or retrieved the data from a database. This is very important information to keep on record, especially for clinical follow-up data.

UNWANTNUM: Provide a list of numerical variables for which summary statistics are not meaningful, such as patient’s id. You have an option to leave this parameter blank.

UNWANTCHR: Provide a list of character variables, for which frequency table are meaningless, such as patient’s first name, last name, etc. You have an option to leave this parameter blank.

UNWANTDATE: Provide a list of non-interested date variables, such as some defaulted date variables.

DIRPATH: Provide file directory where the summary report will be saved.

ID: Provide variable name which indicating individual subject, such as patient ID.

NUM: Provide the number of extreme values to identify for numerical variables. Default is set to 3.

PAGEF: Choose any options (OFF, ON, Never) based on the requirement of the summary report.

TECHNICAL AND REPORT DETAILS
DEFINE REPORT TEMPLATE
See code in Appendix Part I on defining customized template for generating data summary report. Any parameters defined in this template can be easily modified to suit the users’ requirements.

GET VARIABLE LIST
Code in Appendix Part II is a macro nested in the final macro %DataAutoSum. By calling this macro inside of macro %DataAutoSum, a list of variable names in any dataset can be acquired. Therefore it is a critical part of this entire macro.

SECTION I: REPORT TITLES
The title of data summary report includes name of SAS data checked, name of SAS library where dataset was stored, data frozen date, data quality checking date (SAS system date when macro was called), and the number of variables and observations in the dataset. The following is an example of report titles:

SAS Data: **DIAG_RESULTS** from the SAS Library: **BIMS2996**
Frozen or Download Date: **November 2, 2004**
Checking Date: **June 30, 2005**
The Number of Variables: **55**
The Number of Observations: **254**

SECTION II: SUMMARY TABLE1
This is an example of first summary table in report. Basic information on all variables such as variables name, format, label, etc can be found in this section. The most useful information from this table is the number of missing values for each variable. Code for this section can be found in Appendix Part III.

<table>
<thead>
<tr>
<th>Variable_Name</th>
<th>Number_Missing</th>
<th>Variable_Format</th>
<th>Variable_Lable</th>
</tr>
</thead>
<tbody>
<tr>
<td>DateInserted</td>
<td>0</td>
<td>DATE</td>
<td>DateInserted</td>
</tr>
<tr>
<td>ANC</td>
<td>23</td>
<td>Num</td>
<td>ANC</td>
</tr>
</tbody>
</table>

SECTION III: SUMMARY TABLE OF DATE VARIABLES
Date related variables such as patients following up date, adverse event date, etc. are critical information. Earliest and latest dates on each date-related variable are listed in the following example table from report. Extreme date values and the range for each date variable can be identified from this table. Code for this section can be found in Appendix Part IV.

<table>
<thead>
<tr>
<th>MRN</th>
<th>Extreme Date</th>
<th>Date_Situation</th>
</tr>
</thead>
<tbody>
<tr>
<td>11111</td>
<td>07JUL2000</td>
<td>Earliest_BMADATE</td>
</tr>
<tr>
<td>11111</td>
<td>08OCT2004</td>
<td>Latest_BMADATE</td>
</tr>
<tr>
<td>33333</td>
<td>06JUL2000</td>
<td>Earliest_CBCDATE</td>
</tr>
<tr>
<td>33333</td>
<td>17OCT2004</td>
<td>Latest_CBCDATE</td>
</tr>
</tbody>
</table>

SECTION IV: TABLES LISTING EXTREME OBSERVATIONS ON EACH NUMERICAL VARIABLE
A series of tables listing extreme observations for each numerical variable are presented in this section. The lowest and highest 3 values for each numeric variable along with their subject identifier such as medical record num (MRN) are shown in the tables by default. You can change default by providing desired number in the parameter 'NUM'. The number of such tables depends on the number of numeric variables in the dataset. You can avoid some tables by specifying a list of
uninterested numerical variables in parameter ‘UNWANTNUM’, which will eliminate tables of pre-specified unwanted variables. Code for this section can be found in Appendix Part V.

Table 3: An example of table listing extreme observations on numeric variables

<table>
<thead>
<tr>
<th>Value</th>
<th>MRN</th>
<th>Obs</th>
<th>Value</th>
<th>MRN</th>
<th>Obs</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>11111</td>
<td>212</td>
<td>97</td>
<td>44444</td>
<td>168</td>
</tr>
<tr>
<td>0</td>
<td>22222</td>
<td>200</td>
<td>97</td>
<td>55555</td>
<td>181</td>
</tr>
<tr>
<td>0</td>
<td>33333</td>
<td>195</td>
<td>99</td>
<td>66666</td>
<td>52</td>
</tr>
</tbody>
</table>

SECTION V: SUMMARY STATISTICS TABLE ON NUMERICAL VARIABLES

Summary statistics provides the most important information on numerical variables. All basic information such as mean, median, etc. for each numerical variable are presented in one table shown as follow. Therefore, you don’t need to run PROC MEAN or PROC UNIVARIATE for each numeric variable individually to get this information. It is convenient and saves programming time. Code for this section can be found in appendix part V.

Table 4: An example of five numbers summary statistics table on numeric variables

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Mean</th>
<th>Min</th>
<th>Median</th>
<th>Max</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANC</td>
<td>3149.7</td>
<td>0.00</td>
<td>836.00</td>
<td>59760.00</td>
<td>59760.00</td>
</tr>
<tr>
<td>Platelet</td>
<td>94.91</td>
<td>0.00</td>
<td>63.00</td>
<td>820.00</td>
<td>820.00</td>
</tr>
</tbody>
</table>

SECTION VI: FREQUENCY TABLES OF ALL CATEGORICAL VARIABLES

Number of tables depends on the number of character variables in the dataset. Similarly, you can also provide a list of uninterested character variables in parameter ‘UNWANTCHR’ to eliminate frequency tables of unwanted categorical variables, such as patient name, etc. Code for this section can be found in appendix part VI.

Table 5: An example of frequency table on character variables

<table>
<thead>
<tr>
<th>Immunophenotype</th>
<th>frequency</th>
<th>percent</th>
<th>cumulative frequency</th>
<th>cumulative percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>b-nos</td>
<td>5</td>
<td>1.97</td>
<td>5</td>
<td>1.97</td>
</tr>
<tr>
<td>early pre b</td>
<td>78</td>
<td>30.71</td>
<td>83</td>
<td>32.68</td>
</tr>
</tbody>
</table>

SAMPLE MACRO CALL

Clinical data sometime contain information on thousands of patients. Thus, it is very important to pre-specify a list containing unwanted numerical, character or date variables. The best procedure to call this macro is:

Step I. Identify a list of uninterested variable names using PROC CONTENTS.

Step II. Provide unwanted numeric variable names, such as patient ID, in parameter ‘UNWANTNUM’.

Step III. Provide unwanted character variable names, such as patient last name, in parameter ‘UNWANTCHR’.

Step IV: Specify a list of uninterested date variable names in parameter ‘UNWANTDATE’.


See the following for a specific example.
CONCLUSION
The main feature of this macro is that it can automatically generate data summary report with limited user input. Not only can it dynamically determine name, format and type for all variables in any input data set, it also can dynamically determine the number of missing values for each variable. The sample macro call above shows that syntax of this macro is very simple. Data summary report generated from this macro includes most of the basic and essential information, including the five number summary statistics for all numerical variables, frequency tables of all categorical variables, etc. In addition, this macro can help to identify extreme values and the number of missing values on each variable. Finally, this macro automatically saves a summary report in a nicely formatted Microsoft Word document in the pre-specified file directory for future use.

CONTACT INFORMATION
Your comments and questions are valued and welcome. Contact the authors at:

Deqing Pei  
Department of Biostatistics  
St. Jude Children’s Research Hospital  
332 North Lauderdale St.  
Memphis, TN 38105-2794  
Tel: 901-495-3076  
Fax: 901-522-8843

Cheng Cheng, Ph.D.  
Department of Biostatistics  
St. Jude Children’s Research Hospital  
332 North Lauderdale St.  
Memphis, TN 38105-2794  
Tel: 901-495-2935  
Fax: 901-522-8843

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.

APPENDIX:

Part I:

proc template;  
define style Styles.Custom;  
parent = Styles.Default;  
style fonts from fonts /  
'TitleFont' = ('Arial, Helvetica",10pt,Bold)  
'TitleFont2' = ('Arial, Helvetica",10pt,Bold)  
'StrongFont' = ('Arial, Helvetica",10pt,Bold)  
'EmphasisFont' = ('Arial, Helvetica",10pt,Bold)  
'headingFont' = ('Arial, Helvetica",10pt,Bold)  
'docFont' = ('Arial, Helvetica",10pt)  
'footFont' = ('Arial, Helvetica, Helv",8pt);  
replace color_list /  
'fgB2' = blue /* links */  
'fgB1' = darkmagenta /* visited links */  
'fgA1' = black /* table cell foreground */  
'bgA3' = white /* table cell background */  
'bgA1' = white /* table background - shows through if cellspacing>0 */  
'fgR' = black /* row header foreground */  
'bgR' = white /* row header background */  
'fgA2' = black /* column header foreground */  
'bgA2' = white /* column header background */  
'bgP' = white /* page background */  
'fgA' = black /* foreground for everything else: notes, proc titles,*/  
'bgA' = white /* background for everything else: notes, proc titles,*/
'Codings' Corner

Part II:

%macro VarList(dsn=);   %let dsid=%sysfunc(open(&dsn,i));   %let varlist = %sysfunc(varname(&dsid,1));   %do i = 2 %to %sysfunc(attrn(&dsid,nvars));   %let varlist = &varlist %sysfunc(varname(&dsid,&i));   %end;   %let rc = %sysfunc(close(&dsid));   %Quit:   &varlist   %mend Varlist;

option  nodate nonumber center;

Part III:

%macro DataAutoSum(ds=,dfrozen=%str(),unWantChr=%str(),unWantDate=%str(),unWantNum=%str(),dirPath=%str(),id=%str(),num=3,pageF=%str(never));    ods rtf file=&dirPath style=Custom bodytitle startpage=&pageF;    ods select all;    ods noproctitle;/*remove procedure titles*/    proc contents data=&ds out=overall(keep=Libname memname name varnum type label format informat nobs) noprint;run;    data dsName(keep=Libname memname nobs) outname(keep=name type label format informat);    set overall;if _N_ = 1 then output dsName;output outname;run;
    data _null_;    set dsName;    call symput("Tobs",trim(left(nobs)));    call symput("MName", trim(left(memname)));    call symput("LName",trim(left(LibName)));    run;
    proc contents data=overall out=o2(keep=nobs) noprint;run;    data o3;set o2;if _N_ =1;run;    data _null_;set o3;call symput("Nv",trim(left(nobs)));run;/*get number of variables in the dataset that we are quality checking*/    data _null_;call symput("PRTDT", left(put(TODAY(), WORDDATE.)));run;"get the date that we are checking data";

Coders' Corner

SUGI 31
%let ddsid=%sysfunc(open(&ds,i));
%do i=1 %to %sysfunc(attrn(&ddsid, nvars));
data t&i;
set outname;if _N_=&i;
call symput("varname",name);
call symput("vtype",type);
call symput("vformt",format);
call symput("vlabel",label);
run;

data miss&i;set &ds;
if (&vtype=1) && (&varname= . or &varname< 0) then output;
if (&vtype=2) && (&varname=' ') then output;run;quit;
proc contents data=miss&i out=m&i(keep=nobs) noprint;run;
data _null_;set m&i;call symput("missNo",trim(left(nobs)));run;
data missNo&i;
length No_Missing $35;
No_Missing="&missNo";
V_Name="&varname";
V_Type="&vtype";
V_Format="&vformt";
V_Lable="&vlabel";
run;
%end;
%do i=2 %to %sysfunc(attrn(&ddsid, nvars));
proc append base=missNo1 data=missNo&i force;run;
%end;
data missSum;set missno1;
if V_Format=" " then V_Format="Format Unclear";
run;
proc sort data=missSum; by V_Format;run;
proc print data=missSum;var V_Name No_Missing V_Format V_Lable;
TITLE1 color = blue h = 0.2 in "SAS Data Set: &MName from the SAS library: &LName";
TITLE2 "Frozen or Download Date: &dfrozen";
TITLE3 "Checking Date: &PRTDT";
TITLE4 "The Number of Variables: &Nv";
TITLE5 "The Number of Observations: &Tobs";
run;

Part IV:
data dateD;
set outname;
if format in ("DATE","DATE5.","DATE6.","DATE7.","DATE8.","DATE9.","DDMMYY2.","DDMMYY3.",
"DDMMYY4.","DDMMYY5.","DDMMYY6.","DDMMYY7.","DDMMYY8.","MMDDYY2.","MMDDYY3.","MMDDYY4.","MMDDYY5.",
"MMDDYY6.","MMDDYY7.","MMDDYY8.","DDMMYY","MMDDYY","DATETIME","YYMMDD2.","YYMMDD3.","YYMMDD4.",
"YYMMDD5.","YYMMDD6.","YYMMDD7.","YYMMDD8.","WORDDATE3.","WORDDATE4.","WORDDATE5.","WORDDATE6.",
"WORDDATE7.","WORDDATE8.","WORDDATE9.","WORDDATE10.","WORDDATE11.","WORDDATE12.","WORDDATE13.",
"WORDDATE14.","WORDDATE15.","WORDDATE16.","WORDDATE17.","WORDDATE18.","WORDDATE19.","WORDDATE20.",
"WORDDATE21.","WORDDATE22.","WORDDATE23.","WORDDATE24.","WORDDATE25.","WORDDATE26.","WORDDATE27.",
"WORDDATE28.","WORDDATE29.","WORDDATE30.","WORDDATE31.","WORDDATE32.","WORDDATE33.","WORDDATE34.",
"WORDDATE35.","WORDDATE36.","WORDDATE37.","WORDDATE38.","WORDDATE39.","WORDDATE40.","WARDATX4.",
"WORDDATAX5.","WORDDATAX6.","WORDDATAX7.","WORDDATAX8.","WORDDATAX9.","WORDDATAX10.","WORDDATAX11.",
"WORDDATAX12.","WORDDATAX13.","WORDDATAX14.","WORDDATAX15.","WORDDATAX16.","WORDDATAX17.","WORDDATAX18.",
"WORDDATAX19.","WORDDATAX20.","WORDDATAX21.","WORDDATAX22.","WORDDATAX23.","WORDDATAX24.","WORDDATAX25.",
"WORDDATAX26.","WORDDATAX27.","WORDDATAX28.","WORDDATAX29.","WORDDATAX30.","WORDDATAX31.",
"WORDDATAX32.");run;/*these are possible date format in SAS data set, there could be more*/
proc transpose data=dateD out=pei(drop=_NAME_ _LABEL_ &unWantDate);
id name;
idlabel name;
run;
%let J=1;
%do %while (%scan(%varlist(dsn=pei),&J) NE);
  %let var=%scan(%varlist(dsn=pei),&J);
  data temp;set &ds;run;
  data dd&var;
    set &ds;
    keep &id &var;
    where &var^=.;
  run;
  proc contents data=dd&var out=ddd&var noprint;run;
  data _null_;
    set ddd&var;
    call symput("TJ&j",nobs);run;
  proc sort data=dd&var;
    by &var;
    run;
  data f&var;
    set dd&var;
    if _N_=1 or _N_=&&TJ&j;/*get the first and last date for each date variable in the data set*/
    rename &var=ExtremeDate;
    run;
  data ff&j;set f&var;length DateSituation $35;
    if _N_=1 then DateSituation="Earliest_&var";else if _N_=2 then DateSituation="Latest_&var";run;
  proc append base=ff1 data=ff&J force;run;
  %let J=%eval(&J+1);
%end;

data DateDes;set ff1;if _N_=1 then delete;if _N_=2 then delete;run;/*deleted first two replicated rows because of append procedures*/
  proc print data=DateDes;var &id ExtremeDate DateSituation;

Part V:
data numD;
  set outname;
  if type in (1) and format not in ("DATE","DATE5.","DATE6.","DATE7.","DATE8.","DATE9.","DDMMYY2.",
"DDMMYY3.","DDMMYY4.","DDMMYY5.","DDMMYY6.","DDMMYY7.","DDMMYY8.","MMDDYY2.","MMDDYY3.","MMDDYY4.
","MMDDYY5.","MMDDYY6.","MMDDYY7.","MMDDYY8.","DDMMYY","MMDDYY","DATETIME","YYMMDD2.","YYMMDD3.",&
"YYMMDD4.","YYMMDD5.","YYMMDD6.","YYMMDD7.","YYMMDD8.","WORDDATE3.","WORDDATE4.","WORDDATE5.","WORDDATE6.",&
"WORDDATE7.","WORDDATE8.","WORDDATE9.","WORDDATE10.","WORDDATE11.","WORDDATE12.","WORDDATE13.",&
"WORDDATE14.","WORDDATE15.","WORDDATE16.","WORDDATE17.","WORDDATE18.","WORDDATE19.","WORDDATE20.",&
"WORDDATE21.","WORDDATE22.","WORDDATE23.","WORDDATE24.","WORDDATE25.","WORDDATE26.",&
"WORDDATE27.","WORDDATE28.","WORDDATE29.","WORDDATE30.","WORDDATE31.","WORDDATE32.","WORDDAT
X3.","WORDDATX4.","WORDDATX5.","WORDDATX6.","WORDDATX7.","WORDDATX8.","WORDDATX9.","WORDDATX1
0.","WORDDATX11.","WORDDATX12.","WORDDATX13.","WORDDATX14.","WORDDATX15.","WORDDATX16.",&n
"WORDDATX17.","WORDDATX18.","WORDDATX19.","WORDDATX20.","WORDDATX21.","WORDDATX22.","WORDDAT
X23.","WORDDATX24.","WORDDATX25.","WORDDATX26.","WORDDATX27.","WORDDATX28.","WORDDATX29.",&n
"WORDDATX30.","WORDDATX31.","WORDDATX32.");run;/*these are possible date format in SAS data set, there could be more*/
  proc transpose data=numD out=pei2(drop=_NAME_ _LABEL_ _unWantNum);
    id name ;
    idlabel name;
  run;
  %let J=1;
  %do %while (%scan(%varlist(dsn=pei2),&J) NE);
    %let var2=%scan(%varlist(dsn=pei2),&J);
  %end;
data dd&J;set &ds;
if   (&var2=.0) or (&var2<0) then delete;
keep &id &var2;run;

proc univariate data=dd&J noprint ;var &var2;output out=new&J min=MinValue max=MaxValue
mean=Mean median=Median range=Range;run;
data nnew&J;set new&J;length VariableName $30;
VariableName="&var2";
proc append base=nnew1 data=nnew&J force;run;
ods select extremeobs;
proc univariate data=dd&J  nextrobs=&num; var &var2;
id &id;
Title "Extreme values of Numeric Variable  &var2 from DataSet:&MName from SAS library &LName";
run;
%let J=%eval(&J+1);
%end;

proc sort data=nnew1 nodupkey out=NumDes;by VariableName;run;
proc print data=numdes;var VariableName Mean MinValue Median MaxValue Range;run;
TITLE "Descriptive Statistics of Numeric Variables in the DataSet: &MName from the SAS library 
&LName",run;

Part VI:

data chrD;
set outname;
if type in (2) ;run;
proc transpose data=chrD out=pei3(drop= _NAME_ _LABEL_ &unWantChr);id name ;idlabel name;
run;
%let J=1;
%do %while (%scan( varlist(dsn=pei3),&J) NE);
%let var3=%scan( varlist(dsn=pei3),&J);
data cc&J;
set &ds;
keep &id &var3;
run;
proc freq data=cc&J  ;table &var3;
TITLE "Frequency Tables of Categorical Variables &var3 in the Dataset: &MName from the SAS library 
&LName",run;
%let J=%eval(&J+1);
%end;
ods rtf close;
ods output close;
%mend DataAutoSum;