ACDAS - AN AUTOMATED CLINICAL DATA ANALYSIS SYSTEM

Donald D. M. Tong, The Upjohn Company
Lynn T. Julian, The Upjohn Company

INTRODUCTION

To quickly output all analyses specified by statisticians for any set of clinical data is a challenging and urgent problem in the pharmaceutical industry.

In clinical trials, information on numerous variables for drug safety and efficacy, etc, is gathered repeatedly for each study subject either at planned time points or over some unpredictable response time intervals (e.g. cancer trial). The variables, the treatment groups, and the experimental design, etc, are unique to each study and often vary substantially among studies. To analyze the information, statisticians request specific output that best depicts the data of the particular study. Unfortunately, each output format may require the use of SAS® procedures on some compatible data set structures. These structures may differ considerably from the structure of the database that has already been created for the study. For example, in order to do a linear model with a time factor included, a horizontal dataset where the time points are part of the variable names has to be converted into a vertical dataset where time points become a variable. On the contrary, in order to do an analysis of covariance, a vertical database has to be converted into a horizontal dataset. To build the datasets and tailor output repeatedly for hundreds of variables over each study not only is tedious but a nightmare for programming analysts. Errors creep in and pressure mounts as studies pile up. To clean up the vast number of studies, an efficient data analysis system is definitely needed.

To solve this problem, ACDAS - A Clinical Data Analysis System was developed recently in the Division of Medical Affairs of The Upjohn Company. The system, written in the SAS®[1] and SAS® MACRO language, is designed to analyze data of any clinical trial efficiently. It serves four functions: data editing, listing, analyses and output formatting. The design objective is aimed at keeping input information and programming effort to a minimum but at the same time providing a high degree of output flexibility and programming control. The remaining paragraphs describe the structure, operation, and features of the system.

The system consists of four parts and four job runs. The four parts are as follows:

1. A questionnaire for I/O information.
2. A variable information data set.
3. A restructuring of data values into new data sets.
4. A prototype of analysis and output programs.

THE FOUR PARTS

1. A Questionnaire for I/O Information

   At this moment, there are only eleven questions, and each requires a simple or blank answer. Information is asked about:

   1) The name(s) of the dataset(s) to be analyzed.
   2) Variables to be kept or dropped.
   3) Variable names for subject identification (e.g., COUNTRY INVestigator PATIENT).
   4) Medication code, (i.e., treatment group identification).
   5) BY or class variables.
   6) Time point variable.
   7) The value(s) which identify the baseline visit or control visits of each cycle.
   8) Time point label.
   9) The DD name for storing the reconstructed data sets (See Figure 1.), etc.

2. A Variable Information Data Set

   Key information of all input variables is condensed into a SAS® data set. The key information consists of distinct values of patient identifiers, medication code, BY or class variables, time points, and a list of the names of the analyses variables. (See Figures 2-5.) All subject visits or time points are printed with appropriate labels. The names of analysis variables are listed along with their position number (from left to right) in the input data set. The position number is particularly useful for data subsetting. This data set is used for checking information, data merging, data subsetting, and variable selection.

3. A Restructuring of Data Values into New Data Sets

   Restructuring data values into datasets which facilitate the easy use of SAS® PRDCs forms the heart of this automated system. The system converts each vari-

able into variables _VALUE, _DIF, and _LAGDIF, etc which represent the actual observed value, the difference from baseline, and the difference from previous visits, etc, respectively.

4. A Prototype of Analysis and Output Programs

Since the new datasets contain variable names like _VALUE, _DIF, and _LAGDIF, etc, which never change, a number of prototype programs can be easily written in four to five lines of code. Modification to meet specific needs is easily accomplished. In the prototype programs, the use of MACRO language or MACRO variables is not needed.

THE FOUR JOB RUNS

We shall describe the job runs of the system in what follows. To insure the integrity of the analysis data, we recommend submitting the job in four runs. The first two runs are primarily for data editing and listing. The last two runs are for restructuring the data set and analyzing the data. The runs are as follows:

Run #1 Preliminary edit check for medication code, BY or class variables and time point variable.

Run #2 Edit checks and data listings for analysis variables.

Run #3 Storing the restructured data sets and preliminary data analyses.

Run #4 Further analyses by subsetting or merging data sets or by modifying prototype programs.

Although there are four job runs, the answers to the entire I/O questionnaire does not change. Before each run, Macro or prototype programs could be selected or changed from the default code. The job runs are also recommended to be submitted only after any problems of the previous runs are resolved.

The jobs are OS batch jobs. From our experience, processing these jobs interactively affords no distinct advantage.

Run #1 Preliminary Edit Check for Medication Code, By or Class Variable and Time Point Variable

The first job run will check for crucial data set problems for analyses, namely, errors in the medication code, BY or class variables, and the time point variable. Printing the variable information data set which contains the distinct values of medication code, of BY or class variables and of the time point variable provides the first information check. In addition, for each subject, problems with duplicate or missing medication code and time points are automatically checked. If an error occurs, the data set name, the patient IDs, and the problems are printed. A random list of observations (the default number is twelve) is also printed for a quick detection of any gross data problems. (See Figure 6.) Five minimum and five maximum values for the analysis variables are printed without subject ID's. Also, the number of subjects by any combination of BY variables (usually by investigator and medication code) are counted for the whole study and across each visit. (See Figure 7.)

Run #2 Edit Check and Data Listings for Analysis Variables

The second job run primarily edit-checks the analysis variables for outliers and value consistency; however, any part of job run #1 could still be included or excluded. This run generates a preselected number (the default number is five) of min and max values of the actual observations, of the difference from baseline, and of the difference from previous visit for each analysis variable. The min and max values can be selected within any combination of BY variables (the default is by medication code) and they are identified with subject ID's and time of visit. (See Figure 8.) In addition, the observed values, the difference from baseline, and the difference from previous visit of each variable are listed for each subject across time points. Extra visits within a time period are printed toward the right side of the page. (See Figures 9-11.)

Run #3 Storing the Restructured Data Sets and Preliminary Data Analyses

The third job run is composed of storing the datasets, preliminary analyses and part of job run #1 and #2. After all data problems are corrected, the data is ready to be stored. Meanwhile, basic analyses such as ANOVA, listing of means at each time point, etc could be included in this job submission.

Run #4 Further Analysis by Subsetting or Merging Data Sets or by Modifying Prototype Programs

The fourth job run involves data subsetting and program modification

which are relatively simple in the system; also, additional stratification variables may be merged onto the datasets. Since each analysis variable has a position number associated with it, variable grouping or selection could be achieved by using this number in IF statements. In case a certain subject needs to be deleted, a single IF statement will also delete him/her from all analysis variables. BY variables could be merged at anytime, and they are immediately associated with all analysis variables. For programming modification, it is a matter of changing one or two words or writing a few lines of new code for the "entire" study. For example, a change from __VALUE to __DIF will generate output for difference from baseline instead of actual observations.

CONCLUSION

From our experience, this system proves to be extremely powerful and efficient. It eliminates the effort to key in individual variable name for analyses. It also eliminates the need to manipulate the data into different formats for various PROC's. It also provides a flexible analysis variable selection scheme. Any changes can be done in a few lines of code. Data listings across time are automatically printed in a nicely labeled fashion with time of visit being arranged in columns of ascending order from baseline visit. Above all, hundreds of clinical analysis variables can be processed in four to five lines of regular non-Macro SAS® code.

[1]SAS is the registered trademark of SAS Institute Inc., Cary, NC, U.S.A.

Figure 1

```
S A S   L'0 G    OS SAS 82.4        VS2/MVS JOB DMT9000V STEP SAS      PROC

  * 1? NAME(S) OF YOUR INPUT DATABASE(S):(MUST BE GIVEN):
       FOR EXAMPLE:  RESPDATA  SASDAT.MSTR2  LAB.MSTR1  ETC; %LET _INDATA=

                     VERT0  VERT1
                                                                        ;
  * 1A? VARIABLES TO BE KEPT:
       TYPE  (KEEP= VAR1 VAR2 ....)   WITH LEFT AND RIGHT PRN; %LET _KEEP=




                                                                        ;
  * 1B? VARIABLES TO BE DROPPED:
       TYPE  (DROP= VAR1 VAR2 ....)   WITH LEFT AND RIGHT PRN; %LET _DROP=

                    (DROP=INV_PAT)


                                                                        ;
    **********  ( BY / CLASS VARIABLES )   **********************;
  * 2? SUBJECT IDENTIFIERS: (MUST BE PROVIDED)
       EG. SITE INV PAT, OR COUNTRY CENTER PHYSI SUBJID ET; %LET _IDNAMES=

             COUNTRY  INV PAT
                                                                        ;
  * 3? MEDICATION CODE (I.E. TREAT GROUP IDENTIFICATION):
       A BLANK ANSWER MEANS NO COMPARATIVE GROUPS;              %LET  _MC=

                    MC
                                                                        ;
  *    FOR OPEN LABEL STUDY, PLEASE ALSO ANSWER QUESTION 3A IF APPLICABLE;
  * 3A? TREATMENT REGIMEN ASSIGNMENT OTHER THAN THE MEDICATION CODE ABOVE:
       LEAVE IT BLANK, IF IT IS NOT DONE;                       %LET  _TRTASGN=


                                                                        ;
  * 4? OTHER BY OR CLASS VARIABLES ( A GROUP VARIABLE SHOULD BE INCLUDED
       HERE IF THE STUDY HAS A CROSS-OVER DESIGN):
       E.G.  SEX GROUP BODYBLD RACE AGESPAN .. ETC;          %LET _BYNAMES=

                     SEX AGEGROUP V1
                                                                        ;
  * 5? ADDITIONAL DATASET NAME(S) CONTAINING SOME OF THE BY OR CLASS VARS:
       NAMES OF DATASETS BELOW SEPARATED BY BLANKS;          %LET _ABYDATA=

                    BY1 BY2
                                                                        ;
    **********  ( TIME / SEQUENCE VARIABLES )   **********************;
  * 6? TIME-SEQUENCE VARIABLE:  (MUST BE PROVIDED)
       E.G. SEQ OR WEEK OR MONTH, ETC. OSEQ FOR HORIZONTAL;    %LET _OSEQ=

                    OSEQ
                                                                        ;
  * 7? COVERSION CODES FOR TIME-SEQUENCE VARIABLES:
       PLEASE READ THE RULES PROVIDED IN THE DESCRIPTION SHEET;  * E.G.
       *  IF     OSEQ = .   THEN WEEK =9000;
       *  IF     OSEQ = 1   THEN WEEK =   0; *..ETC..ETC.  ANOTHER EXAMPLE
       *  IF _7<=TIMESEQ<7  THEN MONTH=   0; *..ETC..ETC;

                                    ETC.
```

Figure 2

```
THE  UPJOHN  COMPANY  DIVISION  OF  MEDICAL  AFFAIRS                    3
                        PROTOCOL 0000 -- TEST DATA          11:13 MONDAY, FEBRUARY 25, 1985
                AN IMAGINARY MULTICENTER, COMPLETELY RANDOMIZED STUDY
                     RX: (1) TEST
                         (2) PLACEBO
MONITOR: DENNIS THE MENACE, MD.  MRA: SNOOPY     BIOSTATISTICIAN: D.D.M. TONG, PHD.  PROGRAMMER: DONALD DUCK
                                INVESTIGATOR: MICKY MOUSE, MD.

                A LIST OF ID VARIABLES FROM DATABASE CLASID
                TO CHECK AGAINST THE RANDOMIZATION LIST, ETC

 OBS   IDVARS    IDLABEL              COUNTRY     INV     PAT    MC

   1   COUNTRY                      USA          4016      1    PLACEBO
   2   INV         INVEGSTIGATOR    USA          4418      1    TEST
   3   PAT         PATIENT          USA          4418      2    PLACEBO
   4   MC          MEDICATION CODE  CANADA       4419      1    PLACEBO
   5                                CANADA       4420      1    TEST
   6                                CANADA       4420      2    TEST
   7                                CANADA       4420      3    PLACEBO
```

Figure 3

A LIST OF BY/CLASS VARIABLES FROM DATABASE BYVATA
FOR DETECTING ANY UNWANTED CLASS VALUES, ETC

OBS	BYVARS	BYLABEL	MC	SEX	AGEGROUP	V1
1	MC	MEDICATION CODE	TEST	MALE		.
2	SEX	SEX	PLACEBO	FEMALE	TWENTIES	1
3	AGEGROUP	AGE GROUP	.	.	THIRTIES	2
4	V1	LOVE SCORE	.	.	FORTIES	4
5			.	.	SIXTIES	5
6			.	.	.	6

Figure 4

A LIST OF TIME-SEQUENCE VARIABLES FROM DATABASE SEQATA
FOR DETECTING ANY UNWANTED TIME VALUES, ETC

OBS	SEQVARS	SEQLABEL	OSEQ	VISIT_	HSEQ	HSEQLB	VISIT_B	HDFSEQ
1	OSEQ	WEEK	-3	WK-3	FIRSTVST	WEEK -3	WK-3_0	FIRSTDIF
2			-2	WK-2	WK-2	WEEK -2	WK-2_0	WK-2_0
3			0	WK0	BASELINE	WEEK 0 (BASELINE)	WK0_0	BASEDIF
4			1	WK1	WK1	WEEK 1	WK1_0	WK1_0
5			2	WK2	WK2	WEEK 2	WK2_0	WK2_0
6			3	WK3	WK3	WEEK 3	WK3_0	WK3_0
7			6	WK6	WK6	WEEK 6	WK6_0	WK6_0
8			11	WK11	LASTVST	WEEK 11	WK11_0	LASTDIF
9			8003	WK8003	WK8003	EXTRA WEEK 3	WK8003_0	WK8003_0

OBS	HDFSEQLB	VISIT_PV	HLGSEQ	HLGSEQLB
1	DIFF WEEK -3 FROM BASELINE	WK-3_P	FIRSTLAG	LAG DIFF OF WEEK -3 FROM PREVIOUS VISIT
2	DIFF WEEK -2 FROM BASELINE	WK-2_P	WK-2_P	LAG DIFF OF WEEK -2 FROM PREVIOUS VISIT
3	DIFF BASELINE (WEEK 0) FROM ITSELF	WK0_P	BASELAG	LAG DIFF OF BASELINE (WEEK 0) FROM PREVIOUS VISIT
4	DIFF WEEK 1 FROM BASELINE	WK1_P	WK1_P	LAG DIFF OF WEEK 1 FROM PREVIOUS VISIT
5	DIFF WEEK 2 FROM BASELINE	WK2_P	WK2_P	LAG DIFF OF WEEK 2 FROM PREVIOUS VISIT
6	DIFF WEEK 3 FROM BASELINE	WK3_P	WK3_P	LAG DIFF OF WEEK 3 FROM PREVIOUS VISIT
7	DIFF WEEK 6 FROM BASELINE	WK6_P	WK6_P	LAG DIFF OF WEEK 6 FROM PREVIOUS VISIT
8	DIFF WEEK 11 FROM BASELINE	WK11_P	LASTLAG	LAG DIFF OF WEEK 11 FROM PREVIOUS VISIT
9	EXTRA DIFF WEEK 3 FROM BASELINE	WK8003_P	WK8003_P	LAG DIFF WAS NOT CALCULATED FOR EXTRA VISIT

Figure 5

A LIST OF ANALYSES VARIABLES AND THEIR ORDERS FROM DATABASE ORDER
FOR REVIEWING ANY UNWANTED ANALYSIS VARIABLES, ETC

OBS	VARORDER	VARNAME	VARLABEL
1	1	WGT	WEIGHT
2	2	A1C	HEMOGLOBIN A1C
3	3	OK	OK

Figure 6

5 RANDOM OBSERVATIONS: 1. FOR CHECKING GROSS DATASET ERRORS
2. FOR EXAMINING VARIABLES WHICH ARE NOT TO BE ANALYZED, E.G. HAND TABLE CODE, CERTAIN DATES, ETC
SUGGESTION: DROP THOSE UNANALYZABLE VARIABLES IN THE REST OF JOB RUNS THRU QUESTION 1B

OBS	COUNTRY	INV	PAT	OSEQ	WGT	A1C	OK
1	CANADA	4419	1	0	143	5.4	45.3
2	USA	4418	2	0	135	6.2	5.3
3	CANADA	4419	1	3	143	5.4	45.3
4	CANADA	4420	2	6	220	0.5	51.4
5	USA	4418	1	11	127	.	35.5

Figure 7

NUMBER OF SUBJECT AT EACH VISIT IN THE STUDY

					WEEK								
					-3	-2	0	1	2	3	6	11	$003
			NPATIENT		PAT-IENT	PAT-IENT	PAT-IENT	PAT-IENT	PAT-IENT	PAT-IENT	PAT-IENT	PAT-IENT	PAT-IENT
			N	FCTN	N	N	N	N	N	N	N	N	N
COUNTRY	INVEGSTI-GATOR	MEDICATION CODE											
USA	4016	PLACEBO	1	14.3	.	.	1	1	.	.	.	.	.
	4418	TEST	1	14.3	.	1	1	.	.	1	1	1	1
		PLACEBO	1	14.3	1	.	1	1	1	.	1	.	.
CANADA	4419	PLACEBO	1	14.3	.	1	1	.	.	1	1	.	.
	4420	TEST	2	28.6	.	1	2	.	.	.	1	.	.
		PLACEBO	1	14.3	1	.	.	.	.	.	.	.	.
COUNTRY		MEDICATION CODE											
USA	ALL	TEST	1	14.3	.	1	1	.	.	1	1	1	1
		PLACEBO	2	28.6	1	.	2	2	1	.	1	.	.
CANADA	ALL	TEST	2	28.6	.	1	2	.	.	.	1	.	.
		PLACEBO	2	28.6	1	1	1	.	.	1	1	.	.
	INVEGSTI-GATOR	MEDICATION CODE											
ALL	4016	PLACEBO	1	14.3	.	.	1	1	.	.	.	.	.
	4418	TEST	1	14.3	.	1	1	.	.	1	1	1	1
		PLACEBO	1	14.3	1	.	1	1	1	.	1	.	.
	4419	PLACEBO	1	14.3	.	1	1	.	.	1	1	.	.
	4420	TEST	2	28.6	.	1	2	.	.	.	1	.	.
		PLACEBO	1	14.3	1	.	.	.	.	.	.	.	.
		MEDICATION CODE											
ALL	ALL	TEST	3	42.9	.	2	3	.	.	1	2	1	1
		PLACEBO	4	57.1	2	1	3	2	1	1	2	.	.
ALL			7	100.0	2	3	6	2	1	2	4	1	1

Figure 8

THE 2 MIN AND 2 MAX OF _VALUE BY MC FOR THE FOLLOWING VARIABLES

VARORDER=1 VARIABLE=WEIGHT

MEDICATION CODE	MM	___VALUES___	(<MED> SUBJ-ID TIME)
TEST	MIN	112	(<1> 2 4420 1 WEEK -2)
		112	(<1> 2 4420 1 WEEK 0 (BASELINE))
	MAX	220	(<1> 2 4420 2 WEEK 6)
		225	(<1> 2 4420 2 WEEK 0 (BASELINE))
PLACEBO	MIN	132	(<2> 1 4418 2 WEEK 6)
		134	(<2> 1 4418 2 WEEK -3)
	MAX	200	(<2> 1 4016 1 WEEK 0 (BASELINE))
		201	(<2> 1 4016 1 WEEK 1)

THE 4 MIN AND 4 MAX OF _DIF BY MC FOR THE FOLLOWING VARIABLES

VARORDER=1 VARIABLE=WEIGHT

MEDICATION CODE	MM	___VALUES___	(<MED> SUBJ-ID TIME)
TEST	MIN	-5	(<1> 2 4420 2 WEEK 6)
		-2	(<1> 1 4418 1 WEEK -2)
		-1	(<1> 1 4418 1 WEEK 3)
		0	(<1> 2 4420 1 WEEK -2)
	MAX	0	(<1> 2 4420 1 WEEK -2)
		5	(<1> 1 4418 1 WEEK 6)
		6	(<1> 1 4418 1 EXTRA WEEK 3)
		6	(<1> 1 4418 1 WEEK 11)
PLACEBO	MIN	-3	(<2> 1 4418 2 WEEK 6)
		-1	(<2> 1 4418 2 WEEK -3)
		-1	(<2> 2 4419 1 WEEK -2)
		0	(<2> 1 4418 2 WEEK 1)
	MAX	0	(<2> 1 4418 2 WEEK 1)
		1	(<2> 2 4419 1 WEEK 6)
		1	(<2> 1 4418 2 WEEK 2)
		1	(<2> 1 4016 1 WEEK 1)

Figure 9

DATA FOR THE FOLLOWING VARIABLE BY VISIT IN _HVAL
THE OBSERVED VALUE AT EACH VISIT

VARLABEL=WEIGHT

MC	COUNTRY	INV	PAT	WKM3	WKM2	WK0	WK1	WK2	WK3	WK6	WK11	WK8003
TEST	USA	4418	1	.	119	121	.	.	120	126	127	127
TEST	CANADA	4420	1	.	112	112	.	.	.	.	.	.
TEST	CANADA	4420	2	.	.	225	.	.	.	220	.	.
PLACEBO	USA	4016	1	.	.	200	201	.	.	.	.	.
PLACEBO	USA	4418	2	134	.	135	135	136	.	132	.	.
PLACEBO	CANADA	4419	1	.	142	143	.	.	143	144	.	.
PLACEBO	CANADA	4420	3	186	.	.	.	.	.	.	.	.

Figure 10

DATA FOR THE FOLLOWING VARIABLE BY VISIT IN _HDIF
DIFFERENCE FROM BASELINE VISIT

VARLABEL=WEIGHT

MC	COUNTRY	INV	PAT	WKM3_0	WKM2_0	WK0_0	WK1_0	WK2_0	WK3_0	WK6_0	WK11_0	WK8003_0
TEST	USA	4418	1	.	-2	0	.	.	-1	5	6	6
TEST	CANADA	4420	1	.	0	0	.	.	.	.	.	.
TEST	CANADA	4420	2	.	.	0	.	.	.	-5	.	.
PLACEBO	USA	4016	1	.	.	0	1	.	.	.	.	.
PLACEBO	USA	4418	2	-1	.	0	0	1	.	-3	.	.
PLACEBO	CANADA	4419	1	.	-1	0	.	.	0	1	.	.
PLACEBO	CANADA	4420	3	.	.	.	.	.	.	.	.	.

Figure 11

DATA FOR THE FOLLOWING VARIABLE BY VISIT IN _HALAG
LAG DIFFERENCE FROM PREVIOUS VISIT

VARLABEL=WEIGHT

MC	COUNTRY	INV	PAT	WKM3_P	WKM2_P	WK0_P	WK1_P	WK2_P	WK3_P	WK6_P	WK11_P	WK8003_P
TEST	USA	4418	1	.	.	2	.	.	-1	6	1	.
TEST	CANADA	4420	1	.	.	0	.	.	.	.	.	.
TEST	CANADA	4420	2	.	.	.	.	.	.	-5	.	.
PLACEBO	USA	4016	1	.	.	.	1	.	.	.	.	.
PLACEBO	USA	4418	2	.	.	1	0	1	.	-4	.	.
PLACEBO	CANADA	4419	1	.	.	1	.	.	0	1	.	.
PLACEBO	CANADA	4420	3	.	.	.	.	.	.	.	.	.