Paper 208-29

An Introduction to Matching and its Application using SAS®

Jayawant N. Mandrekar, and Sumithra J. Mandrekar
Division of Biostatistics, Mayo Clinic, Rochester, MN

ABSTRACT

When is a matching study done? How much matching is appropriate? These are but a few routine questions for clinical investigators, epidemiologists and biostatisticians who deal with matched case control or cohort studies. Clearly, this is a multidimensional problem. The decision on what is an ideal matching, i.e., 1:1, 1:2, 1:3 and so on, may sometimes involve the number of variables that should be matched on but often times is primarily based on an investigators experience. We propose to give a gentle introduction to matching covering some of the broad issues and scenarios where matching is appropriate, and summarizing the existing literature along with a discussion of their relative strengths and limitations. We will overview some of the existing SAS® macros developed within our institution for matching studies, illustrating their usage with a real life data set. Further, analyses will be illustrated using existing procedures in SAS®. In conclusion, we will discuss advantages and disadvantages pertaining to matching based on sample size, and cost considerations.

INTRODUCTION

A distortion of the association between an exposure (E) and a disease (D) can be caused by unaccounted confounding factors (C), i.e. C is associated with E and C is an independent risk factor for D. For example, smoking (C) confounds the relationship between alcohol consumption (E) and lung cancer (D), since alcohol (E) and smoking (C) are related, and smoking (C) is an independent risk factor for lung cancer (D). Confounding factors, when ignored can often lead to invalid and inefficient estimation of the true E to D association. Matching helps to address many such confounder-related distributional imbalance problems at the design stage of a study rather than at the analysis stage. It provides a means for efficient (increased precision of estimates or in other words reduced standard errors) stratified analysis when the matching variables are associated with both D and E. Thus, a matched analysis helps to assess the relationship of E to D having already taken the confounding factor(s) into account. It is especially useful in situations where the confounding factors are difficult to measure and hence to control, e.g. residence location.

Matching refers to the selection of the reference or comparison group that is similar to the index group with respect to the distribution of one or more potentially confounding factors. In follow-up or cohort studies, the reference group is the unexposed subjects and in case-control studies, the reference group is the subjects who do not have the disease of interest (controls). In a cohort study, matching is done without accounting for the disease status (i.e., only on exposed vs. unexposed), which is unknown at the start of follow-up and hence this alters the distribution of the matching factors on the entire source population from which the study subjects are chosen. On the other hand, in a case-control study, matching is done using the disease status (i.e., diseased vs. nondiseased) and this affects only the distribution of the controls. In an event that matching factors are associated with exposure, the selection process will be different with respect to both exposure and disease resulting in a selection bias (Rothman and Greenland, 1998; Kupper, 1998).

Matching is used to avoid or control for the sparse data problem that might arise from confounding factors measured on a nominal scale with many categories, for example, occupation, number of siblings etc. In the absence of matching, although many subjects may be available, most strata in a stratified analysis might have only one subject (a case or a control) and hence provide no effect information. Through matching however we can ensure that each case will have at least one or more matched controls. This is particularly useful in small studies, where there are not enough subjects to adjust for several confounding factors together. Matching is also advantageous in those situations where obtaining exposure and confounder information from the study subjects may be expensive in which case matching can optimize the amount of information obtained per subject rather than increase the number of subjects. Thus when cost of matching is small compared to cost of additional sample size, matching is desirable (Rothman and Greenland, 1998). Often times, a matching study design and the choice of the matching variables are mainly an investigator's decision based on his/her prior knowledge and experience about the disease.

MATCHING SCHEMES

The matching variables can be either continuous (age, weight etc.) or categorical (gender, ethnicity etc.). While matching on a continuous variable, we need to identify the range of values of the continuous variable that can be considered close enough to declare that an index subject's value is matched with a reference subject's value. Often times however variables are generally categorized for matching purposes in which case the index and reference subjects are matched on the category level of the variable.

Individual and frequency (or category) matching are the two matching schemes. As the name suggests, individual matching involves matching one (1:1) or more (1:m, where m is the number of controls matched per individual case) reference (controls or unexposed) subjects with a single index (case or exposed) subject on the matching factors within each stratum. On the other hand, frequency matching involves matching an entire stratum of reference subjects with an entire stratum of index subjects based on the matching factors. In the case when the criteria for the matching factor is broad, for instance matching on females whose ages are between 40-49, Individual matching is unnecessary and leads to a loss in statistical efficiency. Clearly a case or exposed subject in such a scenario can be matched with any control or unexposed subject from that stratum and hence forcing an individual pairing is unnecessary (Kupper, 1998).

It is important to pay attention to the "overmatching" paradigm while using a matching scheme. It refers to a situation when cases and controls are matched on variables that are related to the exposure i.e., cases and controls are similar with respect to the exposure status (Bland and Altman, 1994, 1995). A good strategy would be to match only on the most important risk factors that are highly likely to manifest as confounding factors and adjust for other confounders at the analysis stage.

In this paper, we discuss the individual matching schemes and analysis methods used in case-control studies, and illustrate them using SAS®. In addition, we will also highlight some advantages and disadvantages of matching.

SELECTING A MATCHING CONTROL FROM A POOL OF CONTROLS

Matching is a common method of adjustment in observational studies. Rosenbaum (1989) combined two essentially disjoint literatures on matching: statistical literature on the construction of matched samples for observational studies and literature from discrete mathematics, computer science and operations research on matching in graphs and networks. The first step is identifying which control is "best" for a particular case. This can be determined using a distance measure, D_{ij}, between the i^{th} case and the j^{th} potential control. Let $\underline{X}^1 = \{x_1^1, x_2^1, ..., x_p^1\}$, and $\underline{X}^0 = \{x_1^0, x_2^0, ..., x_p^0\}$ be the vector of matching variables for N cases and M controls $(M \geq N)$. Then, one possible definition for D_{ij} is based on the weighted sum of the absolute differences between the i^{th} case and the j^{th} potential control, i.e., $D_{ij} = \sum_{k=1}^{p} \left| x_{ik}^1 - x_{ik}^0 \right| \times W_k$ and the total distance $T = \sum_{i=1}^{N} D_{ij}$ is thus a natural way to evaluate how well the entire group of cases is matched to the controls. Rosenbaum (1989) discusses two algorithms to compute this distance measure: *greedy* and *optimal*. The greedy algorithm sorts the cases and controls randomly and matches the first case with the closest control using the smallest D_{ij}, and repeats the process until all cases are matched. This algorithm produces good matches but does not guarantee to minimize the total distance T. The optimal algorithm on the other hand produces the optimal set of matches based on minimizing T. Refer Rosenbaum (1989) for detailed discussions on these algorithms and their properties.

Bergstralh et al. (1995) developed a SAS® macro, *%match*, using the D_{ij} distance measure. This macro can be used to match 1 or more controls (from a total of M available controls) for each of the N cases as well as allows for the controls to be matched to the cases by one or more variables of interest. The *%match* SAS® macro has both greedy *(%greedy*) and optimal *(%optimal)* matching algorithms. The *%optimal* option uses the PROC NETFLOW procedure from SAS®. The macro call statement is as follows:

```
%match(case=, control=, idca=, idco=, mvars=, wts=, dmaxk=, dmax=, time=, method=, ncontls=,
seedca=, seedco=, mincont=, maxcont=, maxiter=, out=, outnmca=, outnmco=, print=);
```

This macro requires that the N cases and M controls are stored in two independent data sets, with an ID variable as well as other matching variables. One obvious constraint is that the matching variables in each data set have the same name. Details on the definitions of each parameter used in the above macro call function can be found in Bergstralh et al. (1995) as well as with the macro documentation online (see reference for SAS® macros resource).

ILLUSTRATION

A large hospital based case control study on benign breast disease was conducted through two hospitals in New Haven, Connecticut to examine the epidemiology of fibrocystic breast disease. Cases included women with a biopsy-confirmed diagnosis of fibrocystic breast and controls were selected from among patients admitted to the general surgery, orthopedic, or otolaryngologic services at these two hospitals. Trained interviewers administered a standardized structured questionnaire to collect information from each subject. More details on this study can be found in Pastides et. al. (1983,

1985). For the purposes of illustration, data from 50 cases and 150 age-matched (age at interview, denoted by AGMT) controls are used (Hosmer and Lemeshow, 2000). The variables of interest are patient identification (PTID), age at menarche (AGMN), weight at the time of interview (WT), presence or absence of regular medical check-ups (CHK), and marital status (MS: 1 = never married, and 0=ever married). We will refer to this data set as the benign breast disease (BBD) data.

As an illustration of the usage of *%match* macro using the BBD data set, we have selected a random sample of 6 cases and 15 controls from the pool of 50 cases and 150 controls. We treat the cases and controls as two separate data sets and use the 1:2 matching scheme with AGMT as the matching variable.

```
Data CASES;
Input PTID AGMT CHK AGMN WT;
Cards;
101 39 1 13 118
102 35 1 14 129
103 28 1 12 108
104 31 1 14 110
105 41 1 13 138
;

Data CONTROLS;
Input PTID AGMT CHK AGMN WT;
Cards;
201 39 2 11 175
202 62 1 11 170
203 35 2 11 170
204 42 1 13 118
205 30 1 14 130
206 39 2 12 135
207 27 2 12 127
208 58 1 10 140
209 36 1 14 110
210 41 2 12 129
211 55 2 13 193
212 45 1 11 154
213 61 2 13 153
214 31 2 11 97
215 28 2 11 145
;
```

Next, the *%match* macro is used to generate a 1:2 matched case-control data set, using AGMT as the matching variable. Here CASES and CONTROLS are the SAS datasets for cases and controls and PTID is the patient identification variable from each of the two datasets. In this example, the matching variable (mvars) is AGMT, weight (wts) is assumed to be 2 and the maximum allowable difference in AGMT between case and the corresponding matched control (dmaxk) is also assumed to be 2. The macro allows matching on several different variables with a variety of options for wts and dmaxk respectively. Maximum number of iterations for NETFLOW procedure is specified by maxiter=10000, which is also a default value. MTCH is a SAS dataset containing the results of the matching process for matched cases only and can be printed by specifying print=y option. The SAS code to perform matching using “Optimal” algorithm to match 2 controls per case (ncontls=2) is given below.

```
%match(case=CASES, control=CONTROLS, idca=PTID, idco=PTID, mvars=AGMT, wts=2, dmaxk=2,
method=optimal, ncontls=2, maxiter=10000, out=MTCH, print=y); run;
```

Below is the output data set MTCH from the *%match* macro:

OBS	PTID CASE	PTID CONTROL	CONTROL NUMBER	DISTANCE D_IJ	AGMT ABS. DIFF	AGMT CASE	AGMT CONTROL
1	101	201	1	0	0	39	39
2	101	206	2	0	0	39	39
3	102	203	1	0	0	35	35
4	102	209	2	2	1	35	36
5	103	207	2	2	1	28	27
6	103	215	1	0	0	28	28
7	104	205	2	2	1	31	30
8	104	214	1	0	0	31	31
9	105	204	2	2	1	41	42
10	105	210	1	0	0	41	41
				========			
				8			

This output displays the patient identification numbers for matched cases and controls, the weighted distance, the absolute difference for the matching variable, and the actual values of the patient's ages at the interview. The output data set is sorted by patient identification number of the cases. Total distance T is displayed under DISTANCE D_IJ. In this example, the total distance is 8 and all the 5 cases were matched to 2 controls respectively. In an event if some cases cannot be matched based on the initial specified parameters, one may have to relax some of the matching restrictions for the options mvars, wts, dmaxk or ncontls as appropriate.

CONDITIONAL LOGISTIC REGRESSION

Let **x** be the vector of independent variables, β be the vector of unknown coefficients corresponding to **x** and y be the disease outcome. Let us assume that this case-control data set has K strata (or matched sets) with n_{1k} cases and n_{0k} controls, where k=1, 2, …, K, where the subscripts 1 and 0 stand for cases (y = 1) and controls (y = 0) respectively. The total number of subjects be $n_k = n_{1k} + n_{0k}$. Then the conditional (exact) likelihood for the k^{th} stratum is given by:

$$l_k(ß) = \frac{\prod_{i=1}^{n_{1k}} P(x_i \mid y_i = 1) \prod_{i=n_{1k}+1}^{n_k} P(x_i \mid y_i = 0)}{\sum_{j=1}^{C_k} \left\{ \prod_{i_j=1}^{n_{1k}} P(x_{ji_j} \mid y_{i_j} = 1) \prod_{i_j=n_{1k}+1}^{n_k} P(x_{ji_j} \mid y_{i_j} = 0) \right\}}$$

where $C_k = {}^{n_k}C_{n_{1k}}$. Using the assumptions about the sampling of cases and controls and application of Bayes' theorem to each term in the above equation, we get:

$$l_k(ß) = \frac{\prod_{i=1}^{n_{1k}} \exp(ß' x_i)}{\sum_{j=1}^{C_k} \left\{ \prod_{i_j=1}^{n_{1k}} \exp(ß' x_{ji_j}) \right\}}$$

The full conditional likelihood is the product of $l_k(\beta)$ over K strata, $l(ß) = \prod_{k=1}^{K} l_k(ß)$. All of the above notations are consistent with those used in Hosmer and Lemeshow (2000), which can be referred for more details. Due to the remarkable similarities between the likelihood function from the conditional logistic regression model and the partial likelihood function from the Cox proportional hazard's regression model, one can use the Cox proportional hazards model to fit the dataset from a matched case-control study. This is done by treating each matched set as a stratum, assuming all cases within a given matched set to have the same event time. Then, the exact partial likelihood method can be used to fit the data (Vierkant et al., 2000).

Conditional logistic regression is thus used to investigate the relationship between an outcome and a set of risk factors in matched case-control studies. Conditional logistic regression is performed with the PROC PHREG in SAS® using the discrete logistic model and forming a stratum for each matched set. In addition, a dummy variable for the survival times is created such that all the cases in a matched set have the same event time value, and the corresponding controls are censored at the later times. However, this may not be necessary if the dummy time variable is created to have the same non-zero constant value for all cases and controls in the dataset (Vierkant et al., 2000).

ILLUSTRATION

In the BBD data set, the outcome variable of interest (FNDX) is whether the subject is a case (fibrocystic breast, denoted by 1) or a control (no fibrocystic breast, denoted by zero). The variable TIME*FNDX is the response (as in a survival analysis setting), where FNDX is the censoring variable and TIME is the dummy variable created as explained in the previous section. The unique strata variable, STR, is created from the matching variable, AGMT, which is then used in the STRATA statement. The explanatory variables include AGMN, CHK, MS and WT. The TIES=DISCRETE option requests the discrete logistic model that uses the exact partial likelihood to fit the data in the presence of tied event times. This option is crucial when there exists stratum that contains more than one case (for example, as in n:m where n>1). For matched case-control studies with only one case per matched set (1:m matching), the likelihood function for the conditional logistic regression reduces to that of the Cox model for the continuous time scale, and hence all the available methods for handling ties give the same result. The model building techniques, forward selection, backward elimination and stepwise, are similar as in any regression setting.

1:1 Matching

For the illustration of 1:1 matching, we use regular medical check-ups and weight of the subject at interview as independent variables. The PHREG procedure is used to fit this conditional logistic regression model using “discrete” option to handle ties. This procedure displays a summary of the number of event and censored observations for each stratum, which are the number of cases and controls for each matched set respectively. The SAS® commands and the results from the conditional logistic regression analysis are shown below:

```
PROC PHREG data=BBD;
MODEL TIME*FNDX(0)=CHK WT /ties=discrete rl;
STRATA STR;
RUN;
```

The PHREG Procedure

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
CHK	1	-1.25923	0.48664	6.6957	0.0097	0.284	0.109	0.737
WT	1	-0.03243	0.01303	6.1939	0.0128	0.968	0.944	0.993

Based on the Wald test for individual variables, both CHK and WT are statistically significant at 5% level of significance. The odds ratios from the conditional logistic regression model (given under the Hazard Ratio column in the above output) are computed by exponentiating the parameter estimates, and are useful in interpreting the results of the analysis. An estimated odds ratio of 0.28 (0.11, 0.74) indicates that women undergoing regular medical check-up are about one third as likely to be diagnosed of having benign breast disease, than women of the same weight who do not undergo regular medical check up. In other words, women who do not have regular medical check-ups are about 3.6 times more likely to be diagnosed of having benign breast disease. Similarly, heavier women are less likely to have benign breast disease. In summary, regular medical check-up and increased weight significantly reduce the odds of having benign breast disease.

1:3 Matching

For the illustration of 1:3 matching, we use regular medical check-ups (CHK), age at menarche (AGMN), weight of the subject at interview (WT) and marital status (MS) as the independent variables. The SAS® commands and the results from the conditional logistic regression analysis are shown below:

```
PROC PHREG data=BBD;
MODEL TIME*FNDX(0)=CHK AGMN WT MS /ties=discrete rl;
STRATA STR;
RUN;
```

The PHREG Procedure

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
CHK	1	-1.16120	0.44696	6.7497	0.0094	0.313	0.130	0.752
AGMN	1	0.35923	0.12788	7.8909	0.0050	1.432	1.115	1.840
WT	1	-0.02823	0.00998	8.0056	0.0047	0.972	0.953	0.991
MS	1	1.59326	0.73600	4.6862	0.0304	4.920	1.163	20.818

The results from Wald test for individual variables indicate that all the four independent variables are statistically significant at 5% level of significance. An estimated odds ratio of 0.31 (0.13, 0.75) indicates that women undergoing the regular medical check-up are about one third as likely to be diagnosed of having benign breast disease, than those who do not undergo regular medical check up. Women who are older at the time of menarche are about 43 percent more likely to have benign breast disease and the risk doubles with every two-year increase. Heavier women, as seen in the case of 1:1 matching, are less likely to have benign breast disease and an increase of 10 pounds in weight decreases the risk of getting benign breast disease by about 25 percent. Finally, women who are never married are about 5 times more likely to be diagnosed with benign breast disease. Of note is the relatively wider confidence interval in this case, which is due to the sparse data, i.e. due to fewer women in the never married status category. We refer the readers to Hosmer and Lemeshow (2000) for the discussion on model fitting and diagnostics.

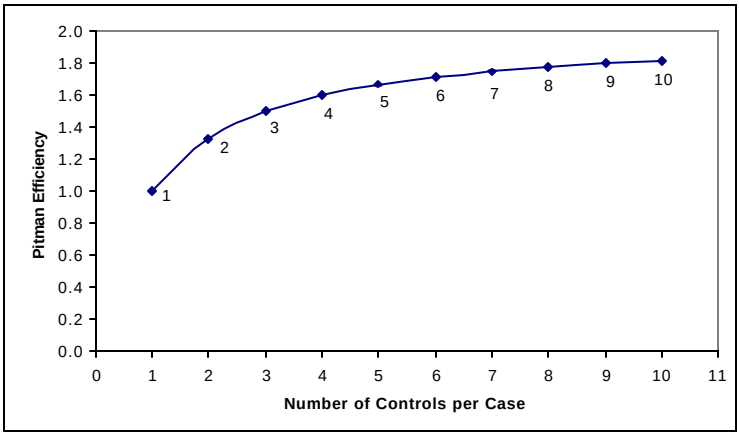
Vierkant et al. (2000) developed a SAS® macro, *%mcstrat*, that fits the conditional logistic regression model, and generates tables that describe the matched sets and the independent variables included in the model. Using the PROC PHREG and PROC IML procedures from SAS®, model diagnostics and fitted values are generated. Below is the macro call for performing the 1:1 and 1:3 matching discussed earlier in this section.

```
%mcstrat(data=BBD, setid=STR, case=FNDX, indvar=CHK WT, uni=yes, diag=yes, tables=CHK WT);

%mcstrat(data=BBD, setid=STR, case=FNDX, indvar= CHK AGMN WT MS, uni=yes, diag=yes, tables=
CHK AGMN WT MS);
```

EFFICIENCY CONSIDERATIONS

A detailed discussion of the gain in statistical efficiency of using multiple controls in a case-control study can be found in Ury (1975) and Miettinen (1969). It has been shown that the improvement is trivial beyond three or four controls per case. This can be explained based on the Pitman efficiency of the Mantel-Haenszel test for stratified data. It is given by $\frac{2m}{m+1}$, when the variances for the case and control measurements are assumed to be equal, where m is number of controls per case. More generally, $\frac{m_1(m_2+1)}{m_2(m_1+1)}$ gives the Pitman efficiency for m_1 controls as compared to m_2 controls. Using these definitions, the gain in Pitman efficiency from using 2 controls relative to 1 control per case is 33% (1 versus1.33) whereas the gain is only 4.2% (1.6 versus 1.667) from using 5 controls relative to 4 controls per case. The graph below gives a visual representation of the Pitman efficiency from 1 up to 10 controls per case in a matched case-control study.



Another aspect of efficiency in matching studies is the cost efficiency, which is judged by the amount of information per unit of cost involved in obtaining that information. In the case of individual matching, if several matching factors are used, then substantial effort (and expense) may be incurred in finding potential controls (often times this leads to gathering information on a number of eventually unmatched subjects) with same characteristics as the case. Cost efficiency is of significant concern and a careful assessment of both the cost and size efficiency is needed (see Rothman and Greenland, 1998, for more details).

DISCUSSION

Matching is essentially stratification done at the design stage to form strata that are sufficiently balanced that provides for an efficient statistical analysis. It is desirable in the following scenarios: a) when a random sample is difficult to find, b) when it is quick, easy and inexpensive to get controls, c) to control for confounding factors that are difficult to measure, all in an effort to improve the efficiency of the study. Considering the cost and time involved in matching the cases with the controls, it is important to think about potential drawbacks like a) misclassification bias, b) inability to examine the risk factor(s) associated with the matching variable, c) if controls are not identified then the data collected from the cases cannot be used and vice versa, and d) possible overmatching.

REFERENCES

1. Allison PD. (1999) Logistic Regression Using the SAS System: Theory & Application. SAS Institute Inc., Cary, NC, USA.

2. Bergstralh EJ, Kosanke JL (1995). Computerized matching of controls. *Technical Report Series No. 56*, Department of Health Science Research, Mayo Clinic, Rochester.

3. Bergstralh EJ, Kosanke JL, Jacobsen S (1996). Software for optimal matching in observational studies. *Epidemiology,* 7, 331-332.

4. Bland JM, Altman DG (1994). Statistics notes: Matching. *British Medical Journal*, 309, 1128.

5. Bland JM, Altman DG (1995). Matching in case-control studies. *British Medical Journal*, 310, 329-330.

6. Fleiss JL, Levin B, Paik MC (2003) Statistical Methods for Rates and Proportions. New York: John Wiley and Sons.

7. Hosmer DW, Lemeshow S (2000). Applied Logistic Regression, 2nd Edition. New York: John Wiley and Sons.

8. Kupper LL (1998). Matching in Encyclopedia of Biostatistics / editors-in-chief, Peter Armitage, Theodore Colton, 2441-2445, New York: John Wiley and Sons.

9. Miettinen OS (1969). Individual matching with multiple controls in the case of all-or-none response. *Biometrics*, 25, 339-355.

10. Ming K, Rosenbaum PR (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics,* 56, 118-12.

11. Ming K, Rosenbaum PR (2001). A note on optimal matching with variable controls using the assignment algorithm. *Journal of Computational and Graphical Statistics*, 10, 455-463.

12. Pastides H, Kelsey JL, LiVolsi VA, Holford T, Fischer D, Goldberg I (1983). Oral contraceptive use and fibrocystic breast disease with special reference to its histopathology. *Journal of the National Cancer Institute*, 71, 5-9.

13. Pastides H, Kelsey JL, Holford TR, LiVolsi VA (1985). The epidemiology of figrocystic breast disease. *American Journal of Epidemiology*, 121, 440-447.

14. Rosenbaum PL (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408), 1024-1032.

15. Rothman KJ, Greenland S (1998). Modern Epidemiology, 2nd Edition. Philadelphia: Lippincott-Raven Publishers.

16. Ury HK. (1975) Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data. *Biometrics*, 31(3): 643-9.

17. Vierkant RA, Therneau TM, Kosanke JL, Naessens JM (2000). Penalized Survival Models and Frailty. *Technical Report Series No. 66*, Department of Health Science Research, Mayo Clinic, Rochester.

SAS® MACROS RESOURCE

Mayo Clinic, Department of Health Sciences Research, Division of Biostatistics http://www.mayo.edu/hsr/sasmac.html

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Address all correspondences to:

Jayawant N. Mandrekar, Ph.D.
Mayo Clinic
200 First Street SW
Harwick 7
Rochester MN 55905
Work Phone: (507) 266 0573
Fax: (507) 284 9542
Email: mandrekar.jay@mayo.edu