A major function assumed by many of today's Corporate Human Resource Organizations is the provision of data processing application services and developments for Benefits Departments. A topic of much interest to benefit professionals is the cost factor associated with maintaining the physical and mental well-being of the workforce and their covered dependents. Executives nationwide have a growing concern over the perpetually increasing premiums and expenses of employee health insurance. And with this concern comes a very real need for expanded and flexible health care data analyses and reporting.

Our SAS application presents a sampling of approaches to providing benefit groups with descriptive statistics regarding the demographical profiles of their various health insurance participants and generating the results of comparative and multiple regression procedures concerning variables of specific interest (e.g., the relationship of premiums paid to employee age, sex, and/or number of dependents). The simulations utilize realistic data elements and historical-type variables of common interest to most benefits and claims groups. The advantages which may be derived from our development include the following:

1. Provides flexible, in-house approaches to accommodate the reporting and charting requirements of benefit managers in a cost-effective manner.

2. Establishes, at arm's length, a framework for monitoring health care costs on a comparative basis for time-series analyses (e.g., self-insurance versus health maintenance organizations or HMOs).

3. Improves internal control and integrity with respect to input data and enhances the response time for satisfying information requests—the basis from which decisions may be made.

Our project demonstrates SAS as a powerful and pragmatic statistical tool and users should find our examples readily adaptable. One may combine all or part of the sample techniques into a tailored set or model for determining participant demographical attributes and evaluating health benefits variable relations.

The importance and urgency for understanding and controlling health care costs cannot be overemphasized. According to a 1984 U.S. Chamber of Commerce survey, $610 billion are spent annually on employee health care [3]. Today, more and better health care is being sought by a work force that has grown to expect easy access to health care with little or no concern for the cost. And the reasoning: Why should employees care about cost when the bulk of the financial burden is borne by employers through their group medical insurance policies [4]? Well, employers are certainly reaching their limits in tolerating any such attitude and in fostering any such complacency toward the escalating costs of medical care coverage. They are taking the entire health care community to task in efforts to cut the fat—sources say there is $100 billion a year of it—out of the system [9].

Health care costs in America have soared more than 1,000 percent since 1950. It is startling to note that by the mid-seventies, medical care cost inflations had become so great and plans so liberal that General Motors was spending more for medical insurance than for steel. Another eye-opener is that more than nine percent of the Gross National Product is spent on health care. Despite all the cost containment measures employed by several companies, corporate America is still experiencing an average rate of increase in health care costs of 20 percent annually [4]. Though the statistics vary from organization to organization, the predicament remains the same: Health care costs are threatening the very solvency of many of America's most successful businesses [6].

So, who is to blame? The hospitals blame the physicians, the physicians blame the hospitals and malpractice insurance, the insurance carriers blame everyone in the system including the government, and the beat goes on [7]. And the situation is made even more complex for benefits professionals regarding the span of choices for providing health care; ranging from the traditional fee-for-service health insurance, to health maintenance organizations (HMOs), to preferred provider organizations (POPs), to diagnostic related groups (DRGs), to self-insurance, and on to numerous variations thereof.

With the well-documented awareness of the high costs associated with health care and the accompanying complexities for coping with its provisions [1], [2], [5], there exists the very real need for employers to make intelligent
choices. Many management decisions aimed at making selections and modifying employee health behavior are being made because they seem to "make sense" whether or not they are supported by good data. However, decisions should be made based on specific patterns of use and knowledge of the practice patterns of individual practitioners [8]. In meeting the growing needs for information in making health care decisions comes the expanding role which many corporate and public human resource organizations are assuming in data analysis and provision of benefits information reporting. The model of our Statistical Analysis System (SAS) software application will be described within the context of this environment.

Our focus is with a logical first step in developing a cost containment strategy: an analysis of participant demographical profiles, on a comparative basis, and their medical claims history. In this way, potential relationships among health insurance variables and medical expense patterns may be identified. We suggest, by some examples, how SAS procedures and statistics (e.g., descriptive statistics, charting and graphing, correlation, and regression) may be used in developing a prototype for evaluating sample data items. Of course, what variables and statistics are of specific concern will vary among benefits managers and the examples are provided with this need for flexibility in mind. Furthermore, the applicability of our writing is somewhat contingent upon the availability and reliability of benefits data bases and historical records maintained by an organization. Our purpose is to demonstrate how potential profiles may be determined and how potential relationships among health care variables may be assessed so as to facilitate appropriate future development of programs to improve efficiency and effectiveness of employee medical care usage.

METHODS AND CONTENT

The types of analyses conducted depends to a large extent upon what the specific benefits organization deems of most importance and relevance. Regarding the development of demographical profiles of health care members, SAS provides some powerful descriptive statistical tools. Procedures of interest here are MEANS, UNIVARIATE, SUMMARY, PROG, TABULATE, and CORR. Demographical variables of concern often include age, sex, race, years of service, marital status, number of dependents, and so on.

PROC TABULATE

To illustrate, the TABULATE procedure is very effective in depicting distributions of classification variables and is flexible in providing a variety of statistics (e.g., mean, standard deviation, range, sum, student's t value, percentage, etc.). Assume we are interested in finding the distribution of age categories between sex among employees subscribing to offered HMO plans versus the organization's self-insured medical plan (SIMP). And we would like this distribution to be shown by subsidiary (SUBSID) and to include a further breakdown of active versus retired employees. The SAS code to invoke this procedure would look something like the following:

```
PROC TABULATE;
CLASS SUBSID EMPTYPE INSTYPE SEX AGE;
FORMAT AGE AGECATS;
TABLES COMPANY, SEX*AGE
(ALL='TOTAL' EMPTYPE = 'ACTIVE/ RETIRED')*(ALL INSTYPE)
N = ' ' *F = 7 /RTS s 15'
```

PROC FREQ

Another example would be the use of PROC FREQ to summarize, say, the frequencies of medical care received (variable name = TYPECARE) by age and sex. The TYPECARE would be a nominal variable representing the type of health service received by a participant (e.g., inpatient/outpatient surgical utilization, maternity care, psychiatric care, substance abuse—and subcategories within these categories). It would again be useful to express age data as categories. The SAS code used to generate a two-way cross-tabulation table of the second and third variables as produced for each value of the first variable would be:

```
PROC FREQ;
       TABLES TYPECARE*AGECAT*SEX;
PROC FREQ;
       TABLES LEVEL*SEX;
PROC FREQ;
       TABLES LEVEL*SEX;
      WEIGHT ADMISSNS;
```

Our data set contains variables LEVEL (management and nonmanagement), sex, and ADMISSNS (frequency of hospital admissions). The above statements would produce a table showing the number of admissions (normally for a predetermined time period) for nonmanagement females, nonmanagement males, management females, and management males. Other procedures that may be used for counting and comparing purposes are: TABULATE (for more general table layouts), SUMMARY (for output data sets), CHART (for bar charts and other graphical representations), and FUNCAT (for statistical analysis of categorical models) [10].

PROC CHART & PROC PLOT

Picturing health benefits data within reports and presentations can have an effective impact upon readers and audiences. Printing columns of numbers will give you the exact values represented by your data, but this does not provide a clear picture of your data—where it peaks, where the numbers begin to rise slowly, where the isolated values are [10]. The impact of a histogram or graph is direct and the featured SAS tools relevant here are PROC CHART and PROC PLOT.

The CHART procedure produces vertical and horizontal bar charts, block charts, pie charts, and star charts. These charts are useful for showing pictorially a variable's values or the relationships between two or more variables. In each situation, the variable listed (the chart variable) determines the values that label the bars or sections. Options are available to control the kind of statistics presented and any grouping that is done. For example, you use TYPE= to select a measure to compute and display:

- frequency counts (TYPE=FREQ)
- percentages (TYPE=PCT)
- cumulative frequencies (TYPE=CFREQ)
- cumulative percentages (TYPE=CPCT)
- totals (TYPE=SUM)
- averages (TYPE=MEAN) [10].

The PLOT procedure graphs one variable against another, generating a printer plot. The coordinates of each point on the plot correspond to the two variables' values in one or more observations of the input data set. PLOT takes the values that occur for each observation in an input SAS data set on two variables, say FYEARS (Fiscal years) and MEDCOST (total cost medical care), and then represents the intersection of these values as points on the plot. This would provide a time-series display of the behavior of the MEDCOST variable. All you need to do to produce a plot is to tell the procedure which variables to plot [10]. For example, these statements:

```
PROC PLOT DATA=RETIRES;
PLOT MEDCOST*FYEARS;
```

would produce a simple plot of total dollar costs provided for an organization's retirees over the time period of fiscal years for which historical records exist.

Another illustration would be the use of CHART to highlight, say, the comparison of length of stay for maternity admissions (LOSMA) among the participating hospitals or medical centers (HOSPITAL) over the past fiscal years (FYEARS). The SAS code to produce a vertical bar chart display in this case would be:

```
PROC CHART DATA=FEMALES;
VBAR FYEARS/SUBGROUP=HOSPITAL SUMVAR=LOSMA DISCRETE
```

PROC T TEST

A useful tool for making comparisons is the t statistic as generated by the TTEST procedure. Means for a variable are computed for each of the two groups of comparative interest which are identified by values of a classification or CLASS variable. The t statistic tests the hypothesis that the true means are the same. This can be considered as a special case of a one-way analysis of variance with two levels of classification [11].

The TTEST procedure is based on the assumption that the variances of the two groups are equal and also generates an approximate ′t′ based on the assumption that the variances are unequal. Also, the t test generated here assumes that the variables are normally and independently distributed.

Assume, for our application, the grouping variable is SEX and the variable whose means are to be compared is IPR (inpatient utilization rate). The SAS code to generate the t statistic would be:

```
PROC TTEST;
CLASS SEX;
VAR IPR;
```

Depending upon the confidence interval selected for use within the analysis, the difference between the two means may or may not be statistically significant.

PROC CORR

The CORR procedure may be used to measure the degree of association between two variables. In other words, correlation measures the closeness of a linear relationship between two variables. If one variable X can be expressed exactly as a linear function of another variable Y, then the correlation is 1 (a strong positive association) or -1 (a strong negative association). When the correlation is equal to zero, the two variables are not related to each other at all—that is, the variables are independent of one another [10].

For example, the SAS code listed below instructs the procedure to compute the correlation between the variable on the WITH statement, MEDPAY (amount of dollars spent per employee for medical claims), and the variable listed on the VAR statement, NODEPNTS (the number of dependents claimed per employee for health benefits coverage). The Pearson product-moment correlation, the default correlation, is calculated. The Spearman and Kendall correlations are available upon request on the PROC statement.

```
PROC CORR;
VAR NODEPNTS;
WITH MEDPAY;
```
A valuable SAS tool for understanding potential relationships among health care variables is PROC STEPWISE. This is a regression analysis technique which analyses the relationships between a dependent variable and independent variables. The problem of deciding which of a set of independent variables to include in a regression model is difficult especially when highly correlated independent variables are present in a regression model. The researcher may want to include only one or two of the variables in the model. STEPWISE regression is one way of deciding which variables to include.

Generally, only a small number of a set of multicollinear independent variables will be included in the regression model by the STEPWISE procedure. This procedure tests the parameter associated with each variable in the presence of all the variables already in the model. STEPWISE regression assists the exploratory analysis by searching for the "best" model by bringing into the regression equation the independent variables one by one. PROC STEPWISE allows the user to choose one of five different methods for selecting variables for the model. The default selection, stepwise selection, is a combination of the forward and backward selection methods. The result is a linear equation which best fits the dependent and independent variables.

In the example below, the MODEL statement assigns HOSPCOST (hospitalization costs) as the dependent variable and LOS (length of stay), AGE, and MEDARATE (inpatient medical admission rate) as the independent variable. For purposes of comparison, the BY statement may be used to obtain separate analyses on observations in groups defined by the BY variable—the grouping variable here is HCPRGM (participating health care program). Also, depending upon the distribution of enrollment among the offered health care programs, the use of a weighting factor (W) may be required so as to represent the enrollments proportionately.

PROC STEPWISE;
  MODEL HOSPCOST=LOS AGE MEDARATE;
  WEIGHT W;
  BY HCPRGM;

The default stepwise selection is made since none of the other four methods is indicated on the MODEL statement. The regression equation selects only those explanatory independent variables which are most correlated with HOSPCOST and best explain the HOSPCOST data. The regression model may, therefore, contain all or any subset of those independent variables given on the MODEL statement. Once you have examined all possible models generated by STEPWISE, you can select from this screening the best predictive models and use GLM with them.

PROC GLM

As a last example, let us consider the potential application of PROC GLM to our health benefits concern—another technique that may be used to study relationships among variables. Statistical methods available in PROC GLM ("general linear models") include regression, analysis of variance, analysis of covariance, multivariate analysis of variance, and partial correlation. GLM deals with classification variables, which have discrete levels, as well as with continuous variables, which measure quantities. Our sample application pertains to GLM as used for multiple regression analysis [11].

The primary goal of multiple regression is to account for the changes in a dependent variable based on its hypothesized dependency on two or more independent variables. The basic output generated from GLM includes an overall analysis of variance table, miscellaneous statistics (e.g., R-square, coefficient of variation, mean and standard deviation of the dependent variable), results for Type I and Type IV tests, and estimates for the model parameters—the intercept and the coefficients. The benefits derived from an examination of such calculations are learning if one variable may be expressed in terms of another and predicting one variable’s values from another’s values.

For example, we might hypothesize that total medical payments made (MEDPAY) depends on both time (YEAR) and the annual employee head count of the organization (TOTEMP).

PROC GLM;
  MODEL MEDPAY=YEAR TOTEMP;

YEAR and TOTEMP appear on the right-hand side of the MODEL statement because they are the independent variables.

Another alternative within GLM is to treat polynomial regression as a multiple regression—where within your model you have a term whose value has been squared. The above example may be modified: (to include TOTEMP*TOTEMP and remove YEAR) and enhanced as follows:

PROC GLM;
  MODEL MEDPAY=TOTEMP TOTEMP*TOTEMP/P CLM;
  OUTPUT OUT=MT P=Mpaypred R=RESID;
  PROC PLOT DATA=MT;
  PLOT MEDPAY*TOTEMP=' A' Mpaypred*TOTEMP='P' /OVERLAY;

Within the MODEL statement are two important options for predicted values and residuals. The 'P' requests GLM to print observed, predicted, and residual values for each observation that does not contain missing values for independent variables. The CLM allows for printing of confidence limits for a mean predicted value for each observation. The PROC

692
PLOT will provide for an overlay graphical illustration of the actual and predicted values of MEDPAY (Y-axis) and the values of TOTEMP (X-axis). The goodness of fit of the model will be measured largely by the coefficient of determination $R^2$ and the accompanying regression residuals will reflect the predictive utility of the model.

CONCLUSION

The framework of the aforementioned SAS procedures is such that an organization may tailor their use into a prototype for analyzing their participant demographics and assessing potential relations among health care variables. Three important advantages which may be derived from such a development are outlined below.

First, the collection of SAS procedures we have suggested, or some combination thereof, provide flexible approaches to accommodate the reporting and charting requirements of benefits managers. From collecting, evaluating, and distributing health care data in the form of information reporting, benefits professionals can go far in educating participants and influence their patterns of usage. It appears clear that unit costs of health care services will continue to rise at an unusually high rate. A particular employer's health care costs will then increase to the degree that utilization of health services either remains the same or increases over time. When equipped with a set of revealing and insightful information, the employer has a strong reference base from which to influence the type and level of use of services by employees and their dependents with regard to less costly service alternatives and use of fewer units of services, (i.e., number of hospital days, doctor visits, laboratory tests, and prescription drugs) thereby decreasing their health benefit related expense over time.

Second, having a framework for monitoring health care costs on a comparative basis contributes to the quality of future implementation decisions. Prior to any attempts at developing specific cost containment programs, effective utilization management would be well advised to first obtain the results and study the information provided from a model designed to examine health care variables. The procedures and statistics presented in our model would prove quite effective in highlighting abusive practices and areas for improvement (e.g., excessive lengths of stay, unnecessary admissions for nonsurgical care of back pain, and frequencies of hospitalization for ambulatory—appropriate surgery and diagnostic tests). Without true focus, action plans and alternatives for cost containment can prove to be inappropriate and wasteful.

The third advantage derived from an established health care data analysis model is internal control and integrity with respect to input data and enhanced response time for satisfying information requests. This is especially relevant to benefits managers in their roles as negotiators—whether bargaining with representatives of health care providers or bargaining with representatives of labor unions. For example, having the appropriate model established the benefits manager can be far more timely and effective in selective contracting and negotiating new premiums with HMOs based on documented usage levels of the specific company's participants. Also, for instance, the provision of information obtained from a model's procedures would legitimize management's position when requesting from bargained-for employees required contributions to remain enrolled in a non-PPO care program.

In summary, the use of SAS pictorial and statistical procedures in the development of a health care data analysis prototype would certainly contribute to the benefits manager's decision making when selecting cost containment strategies and action plans. Its real impact and success, however, can be maximized from tailoring the available tools in a manner most responsive to the organization's specific needs.

REFERENCES


SAS is the registered trademark of SAS Institute Inc., Cary, NC, USA.

Bruce W. Eagle
Ameritech Services
3040 Salt Creek Lane
1-23
Arlington Heights, IL 60005
(312) 870-3298

Ralph F. Catalanello
Department of Management
College of Business
Northern Illinois University
DeKalb, IL 60115
(815) 753-6305