Annu. Rev. Public Health. 1999. 20:125–44

METHODS FOR ANALYZING HEALTH CARE UTILIZATION AND COSTS

P. Diehr,[1,2] *D. Yanez,*[1] *A. Ash,*[3] *M. Hornbrook,*[4,5] *D. Y. Lin*[1]

[1]Department of Biostatistics and [2]Department of Health Services, University of Washington, Seattle, Washington 98195; [3]General Internal Medicine, Boston University, Boston, Massachusetts 02118; [4]Center for Health Care Research, Kaiser Permanente, Portland, Oregon 97227; and [5]Population Based Nursing, Oregon Health Sciences University, Portland, Oregon 97201; e-mail: pdiehr@biostat.washington.edu, lin@biostat.washington.edu, yanez@biostat.washington.edu, aash@bu.edu, HORNBROOKMA@chr.mts.kpnw.org

ABSTRACT

Important questions about health care are often addressed by studying health care utilization. Utilization data have several characteristics that make them a challenge to analyze. In this paper we discuss sources of information, the statistical properties of utilization data, common analytic methods including the two-part model, and some newly available statistical methods including the generalized linear model. We also address issues of study design and new methods for dealing with censored data. Examples are presented.

INTRODUCTION

Important questions about health care are often addressed by analyzing health care utilization data. A finding that people with lower income or who live in certain areas of the country use fewer services can indicate problems with access to care. If patterns of use are found to vary by insurance plan, this may suggest positive or negative properties of managed care. Studies that show high variation among geographic areas in rates at which a surgical procedure is performed suggest that residents in some of those areas are not receiving optimal care. Another important area of research is prediction of total health care costs

0163-7525/99/0510-0125$08.00

for a group of people for a year, so that providers can be paid appropriate rates for caring for those people.

Multiple regression methods are used extensively to adjust analyses for important patient characteristics or to predict future utilization for individuals (1, 13, 14, 20, 31, 33, 34; MC Hornbrook, RT Meenan, DJ Bachman, MC O'Keefe Rosetti, MJ Goodman, et al, submitted for publication). Utilization data have several characteristics that make them a challenge to analyze. Over the years general approaches for analysis have evolved, and some new approaches are now possible because of advances in analytic methods and software. In this paper we discuss sources of data, the statistical properties of utilization data, common analytic methods, the use of newly available computing methods, study design, and methods for dealing with censored data. We use data from an evaluation of Washington State's Basic Health Plan to illustrate these methods (7, 24).

SOURCES AND QUALITY OF DATA

Fifty years ago, patient utilization data had to be obtained directly from patients or by abstracting medical records. Such data are expensive to obtain and of questionable quality. In health services research today, such data are used primarily to supplement information obtained from large administrative databases. Examples of administrative databases include the national Medicare and Medicaid databases as well as some state registries and claims files for privately insured patients or members of a particular health facility. Some databases have been created especially for health services research.

Data from the Medicare program, run by the Health Care Financing Authority (HCFA), are confidentiality-protected, research-quality, longitudinally linked, person-level records that track virtually all elderly US citizens from their 65th birthday onwards, through geographical moves and changes in providers, until death. (About 10% of Medicare enrollees are younger, disabled persons, who are tracked from their time of certification.) Available data include the types and amounts of health services used (e.g. hospitalizations, office visits, home health care, surgeries, and diagnostic tests), the medical problems being treated (diagnoses), provider characteristics (site of service and physician training), and charges. Information on long-term care services and outpatient prescription drugs, not covered by HCFA, is not available. Data for capitated managed care systems [health maintenance organizations (HMOs)] are currently incomplete, but should be adequate for analysis by the year 2000. In addition, the Medicare Current Beneficiary Survey collects longitudinal data on all types of health insurance coverage, on health status, and on sources of payment for all health care utilization for Medicare beneficiaries.

Data are also available about state Medicaid programs, which offer health coverage to low-income persons, most often women with children. These data are of varying quality and, because each state administers its own program, eligibility requirements are state specific and change over time. In addition, people move on and off the system as their eligibility changes, and the same person may appear under different identification numbers. More than 30 states currently submit person-level Medicaid data to HCFA, which monitors their quality and constructs and maintains standardized Medicaid research files known as the Medicaid Statistical Information System (MSIS).

Data about privately insured populations in the United States, usually under age 65, have been used extensively in health service research. Such data are reasonably complete but events such as job loss, geographical movement, marriage, divorce, and alternative coverage that becomes available (or is lost) to a spouse or child can disrupt data continuity. Information on services may be lost if the service was billed to a different insurance plan. There is often less information available about dependents than about the contract holder.

The quality of claims data is adequate for many purposes, but it is important to remember that claims are generated to justify reimbursement rather than to facilitate research. For example, a bill for tests to rule out cancer may contain a diagnostic code for "cancer" even if the tests were negative. Few administrative systems capture outside or "out-of-pocket" utilization, such as purchases of nonprescription drugs and use of alternative or noncovered services.

Some states maintain public use files of each hospital discharge. These can be valuable, although the amount of additional information known about a person is usually limited. It often is not possible to identify multiple admissions for the same person, and in some states only civilian hospitals are included. An excellent population-based data set, created by the Agency for Health Care Policy and Research and the National Center for Health Statistics, is the Medical Expenditures Panel Survey (MEPS) (5), which collects data from several sources to provide a complete picture of the health status and health care utilization of a random sample of citizens.

Descriptions of methods for collecting and analyzing utilization data may be found in several references (5, 19, 26).

MEASURES OF UTILIZATION AND COST

Utilization can be measured as the number of services provided to a patient, such as the number of X rays. More often, however, a variety of procedures and services are of interest, and some measure of "cost" is assigned to each service so that resource intensity can be summed over all provided services. Here we use the word "cost" for simplicity; however, our examples are based on billed charges, which differ from the amount paid because of discounts, deductibles, copayments, and coinsurance. Billed charges do not represent all of a person's costs, because patients may have other coverage and typically pay for many costs out of pocket.

Billed charges rarely represent the cost to the provider of providing a particular service, because the actual cost of a procedure is difficult to define and calculate. Some providers use step-down accounting systems that allocate fixed overhead expenses to particular services. The cost estimate for, say, a mammogram, which results from such a detailed accounting process is appropriate from the accounting perspective, but it may vary substantially in a short period of time, owing to changes in the volume of mammograms performed or to shortfalls in seemingly unrelated parts of the system. Such data must be used with care in research.

Because pricing systems are likely to vary among providers, investigators sometimes adjust the charges by a facility's "cost-to-charge" ratio. Such an approach assumes that costs are actually known and that the relationship of costs to charges can be represented very simply. Another approach is to summarize utilization by counting each unit of care and applying a "relative value" to each procedure performed (2) instead of using the facility charges. If two providers treat cases identically but have different charge structures, the relative-value approach will find the two providers to be equivalent. Depending on the goals of the analysis, it may be appropriate to study charges, costs, or relative values. These approaches will not always yield the same results.

DISTRIBUTIONAL PROPERTIES OF UTILIZATION DATA

Utilization data are generally analyzed by using ordinary least squares regression, but they do not satisfy the standard assumptions for that method. To illustrate these points, we present data from the evaluation of Washington State's Basic Health Plan, which provided subsidized health insurance for low-income residents, starting in 1989 (7, 24). The 6918 subjects in the study were enrolled in four health plans, 26% in a HMO and 74% in one of three independent-practice associations (IPA). Subjects were aged 0–65 years (mean 23 years) and were followed for an average of 22 months (range 1–44 months). The length of follow-up depended on when the person joined the program relative to the end of the evaluation period and is probably not related to the person's health. During the study period, 79% of subjects used some services, and annualized outpatient costs ranged from 0 to $22,452; the mean annualized cost was $390, and the standard deviation was $895.

Model Assumptions

It is tempting to treat the numbers of visits or admissions as count data and to analyze them by using Poisson models. However, the units counted are not independent, and such a model will usually underestimate the variance.

Ordinary least squares (OLS) regression that includes hypothesis testing requires the dependent variable to satisfy normality, homoscedasticity, and independence assumptions. (More formally, the residual error must satisfy these assumptions. Since the residuals in utilization data usually have a similar distribution to the original data, we discuss characteristics of the original data.)

Health care utilization variables are usually not normally distributed, as they tend to have a mode at zero and a distribution with a long, heavy right tail. The distribution often looks more like a lognormal than a Poisson distribution, even when the data are counts. Figure 1 shows the distribution of annualized outpatient costs for 6918 persons, which is heavily skewed to the right. Costs above $3000 were truncated at $3000 to make the graph easier to read. The average cost was $390, with a standard deviation of $895. Total costs are more variable than outpatient costs because of the relative rarity and high cost of hospital care.

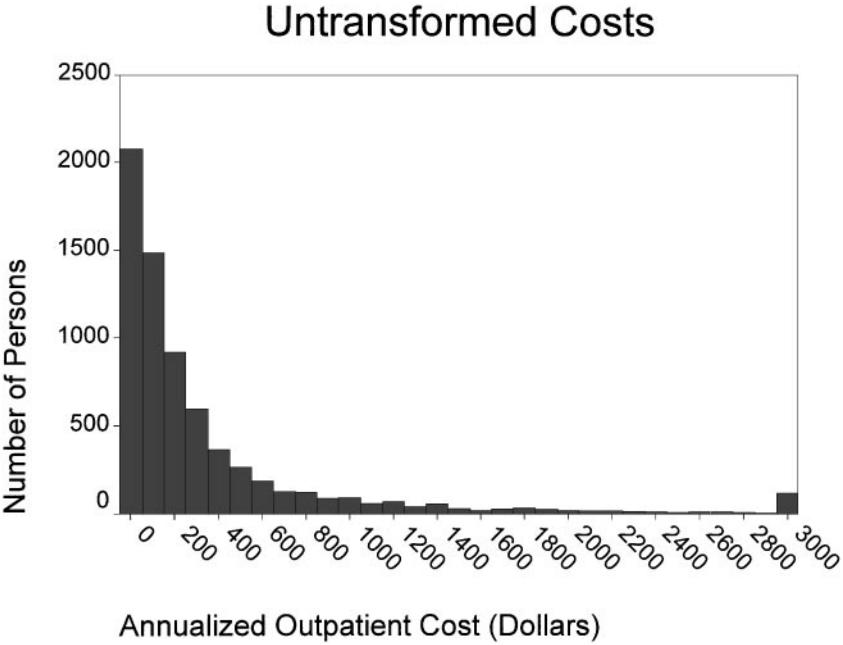

Figure 1 Distribution of annualized outpatient costs.

Counts of visits, admissions, inpatient days, laboratory procedures, and X rays also have distributions with many zeroes and a long right tail.

A second assumption of least squares regression is homoscedasticity, i.e. that the variance is the same for any fixed combination of the covariates. When we grouped the subjects in our example into 132 categories by age and sex, there were substantial differences in variance among the subgroups. There was a strong relationship between mean and standard deviation in the subgroups, which can be characterized as approximately $\sigma = 0.6\ \mu^{1.1}$. That is, the standard deviation increases approximately as the mean. This relationship has been noted elsewhere (3, 6; DK Blough & SD Ramsey, submitted for publication). In the Poisson distribution, in contrast, the standard deviation increases with the square root of the mean; in the normal distribution, the mean and variance are independent.

A third assumption of least squares regression is that the observations are independent. Utilization data can fail to be independent for several reasons. There may be multiple hospitalizations for the same patient. Study subjects may be clustered within families or use the same doctors or hospitals. Hierarchical modeling has been used to adjust for clustering (17). Generalized-estimating equations and random-effects models are also appropriate when the number of clusters is large (21).

Utilization data are often transformed to the log scale, which usually shortens the long right tail, lessens heteroscedasticity, and decreases the influence of outliers. Figure 2 shows the distribution of logarithm of “outpatient cost +1” (one dollar is added to costs to permit calculating a logarithm for people with no utilization). The distribution resolves into a spike representing the people who used no services and a somewhat normal-looking distribution for people who did use services. Models that analyze these two distributions separately are discussed below. The shape of the distribution will differ by population; a healthy insured population will have many nonusers, whereas a population of sick people will have few.

DESIGN OF STUDIES

Length of Follow-Up

When designing a study, investigators must decide how long to follow the subjects. To address this question, we divided the subjects into groups depending on the number of months they were followed, and we calculated the standard deviation of annualized costs for each group. There are at least 50 people in each group, and the smallest number in the groups with 1–4 months of experience is 143. Figure 3 is a plot of the standard deviation of annualized outpatient cost versus the number of months the person was followed. For

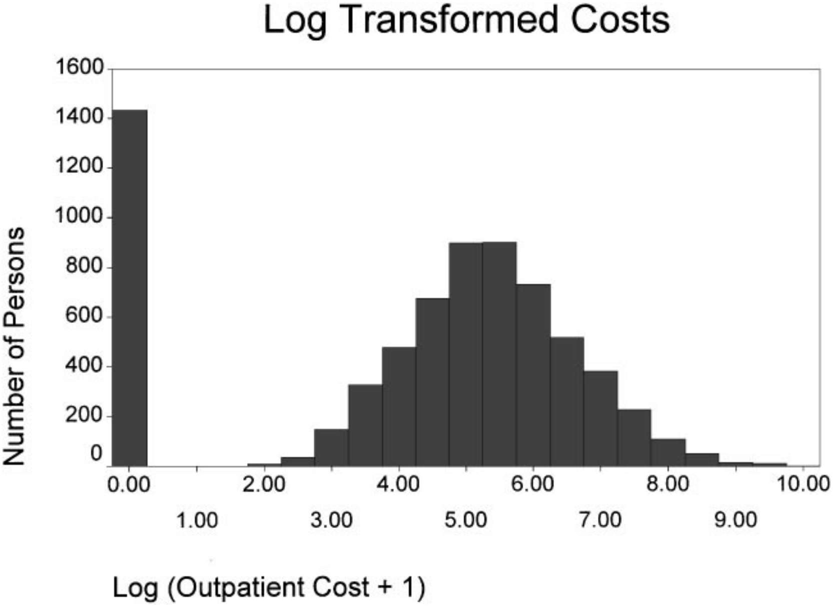

Figure 2 Distribution of log-transformed outpatient costs.

example, the leftmost point shows the standard deviation of annualized cost for people followed only one month; their annualized cost is thus their observed cost multiplied by 12. Note that there is very little association between the length of follow-up and the standard deviation. That is, following people for more than about 6 months does not reduce the variability in annualized outpatient cost. This phenomenon has been noted elsewhere (6) and is similar when inpatient and outpatient costs are combined. Power considerations (e.g. for a study comparing one insurance plan with another) thus do not require long follow-up; however, most investigators prefer to follow subjects for an integral number of years, to guard against seasonal trends in utilization.

Adjusting Utilization for Different Lengths of Enrollment

When study subjects are followed for different lengths of time, it is common to annualize the utilization rates. This may seem risky because a person with only 1 month of follow-up who had an admission during that month would be estimated to have 12 admissions per year. Figure 3 shows that the standard deviation for the annualized cost is fairly independent of the number of months of follow-up for these subjects, whose length of follow-up depended only on when they entered the study. If subjects have lower follow-up for reasons related

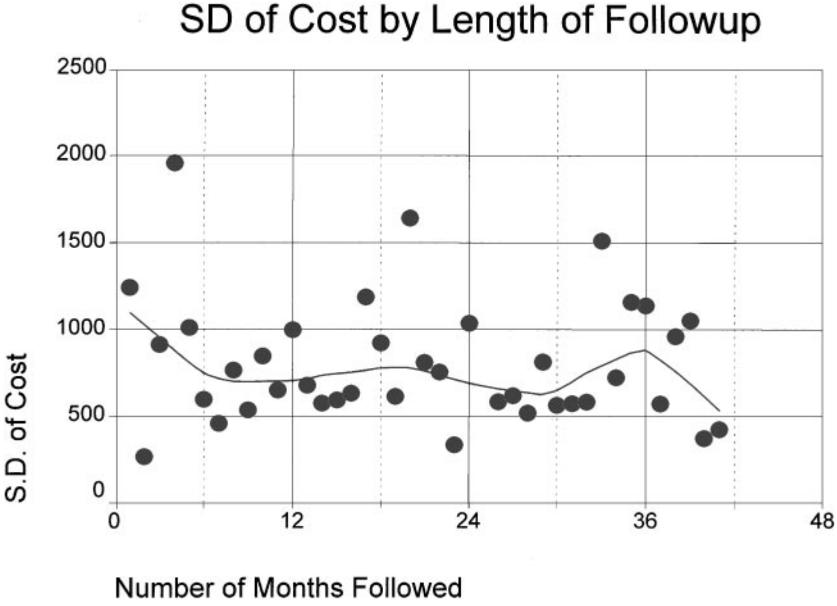

Figure 3 Standard deviation of annualized outpatient costs compared with months followed.

to their utilization or health, however, people with low follow-up may be a biased subset. Ellis & Ash suggest annualizing rates but then weighting persons based on the number of months they were actually observed (12). If length of follow-up differs among subjects, it is probably a good idea to include length of follow-up as a covariate as well, since annualization may not completely remove the effect of time. For example, new enrollees may have pent-up need, and short-term enrollees could have higher annualized rates because they have more use in the first few months. Related issues are described below in the section on censoring.

Sample Size

In the example above, the standard deviation of annualized cost is about $900. Based on the usual sample size calculation methods, a study with 80% power to detect a difference in mean cost of, say, 0.2 standard deviations ($180) would require about 392 people per group. Even though the "effect size" approach of Cohen (4) calls a difference of 0.2 standard deviations small, a difference in cost this large would often be unrealistic to expect. A more reasonable expected difference of $50 per year would require a sample size of about 5100 people per group. Cost studies usually require large samples before reasonable differences can be detected. If pilot data on standard deviations are not available, the relationship noted above in which the mean is approximately proportional to the standard deviation may be useful for sample size calculations.

The cost-effectiveness ratio for an intervention is the difference in mean costs between the treatment and control groups divided by the difference in mean effectiveness. Traditionally a cost-effectiveness ratio has been presented as a single number, with no associated measure of variability. In clinical trials, in which costs and benefits may be estimated for the same individuals, it is possible to use methods such as the bootstrap to provide a confidence interval for such a ratio (29). In our example, the standard error for the difference in mean costs for two groups of size 100 each is about \$131; with 1000 people it is about \$41. When this variability in the numerator is combined with the likely high variability of the denominator, it becomes clear that very large studies may be needed to provide a tight bound on the cost-effectiveness ratio.

ANALYTIC METHODS FOR UTILIZATION DATA

Adjustment for Patient Characteristics

Most analyses attempt to adjust for patient characteristics before testing hypotheses about cost. Age and sex are the most common covariates, because they are reasonable proxies for a person's need for services and are nearly always available. Health status, beliefs, and behaviors are sometimes elicited through surveys of the subjects. Health status is sometimes inferred from the diagnostic codes in a person's previous utilization (12, 19, 33, 34; MC Hornbrook, RT Meenan, DJ Bachman, MC O'Keefe Rosetti, MJ Goodman, et al, submitted for publication). A strong predictor of future utilization is previous utilization, based on survey or claims data. As in any regression analysis, it is important not to control for variables in the causal pathway. For example, in comparing two plans, adjusting for the enrollee's utilization in the previous year in that same plan could mask the differences between the plans.

Relationship of Cost to Age and Sex

Most analyses in the literature adjust for age and sex in a very simple way. One common model is cost $= a + b(\text{sex}) + c(\text{age})$; this model assumes that the relationship of age to utilization is linear and that regression lines are parallel for men and women. An interactive model, cost $= a + b(\text{sex}) + c(\text{age}) + d(\text{age})(\text{sex})$, assumes a linear relationship with different slopes for men and women. Adding a quadratic term, such as cost $= a + b(\text{sex}) + c(\text{age}) + d(\text{age})(\text{sex}) + e(\text{age}^2) + f(\text{age}^2)(\text{sex})$, assumes a quadratic relationship that is different for men and women. The actual relationship of cost to age is not

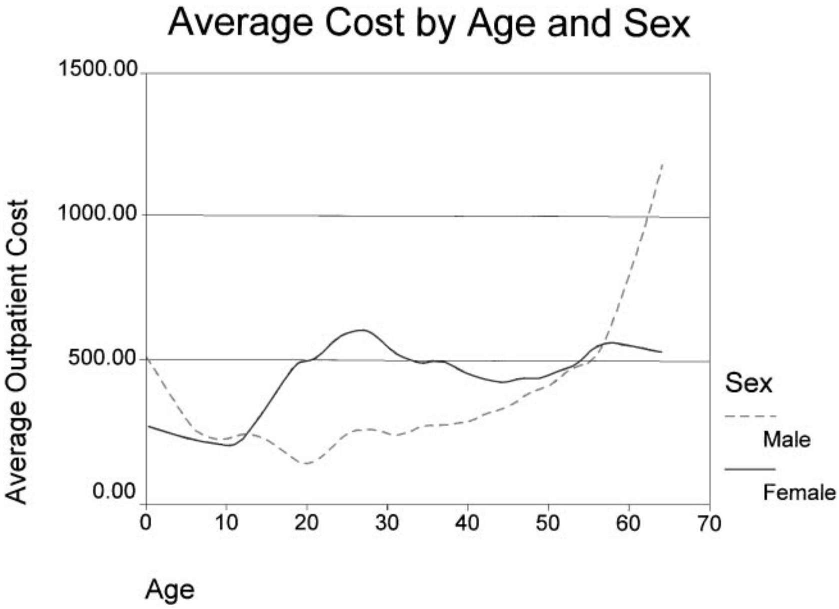

Figure 4 Average outpatient costs by age and sex.

simple, as is shown in Figure 4. The curves are nonlinear and differ by sex. Very young children have high utilization. Males and females have similar utilization until puberty, at which time women increase their utilization because of childbearing. Men's utilization is low until about age 40.

Figure 4 suggests the use of richer age and sex models than those noted above. The curves in Figure 4 can be fit adequately by using a fifth-order polynomial in age (different for men and women). Another approach is to use dummy variables for the various age/sex categories; this requires large sample sizes to permit good estimates for each age/sex category. If the population of interest includes only a subgroup of ages, for instance children under age 12 or persons aged 35–55, simpler models may be appropriate.

Size of R^2

In regression analyses of utilization data, the values of R^2 are usually on the order of $\leq$20% (27). This should not be surprising considering how difficult it would be to predict one's own utilization. The low values of R^2 indicate that we can not predict well for an individual; however, more often we are trying to predict the average cost for a group of individuals, and regression equations can often do this quite satisfactorily. Newhouse et al used theoretical and empirical arguments to estimate that the maximum possible R^2 is about 48% for outpatient expenditures and 15% for total costs (28).

One-Part Versus Two-Part Model

In analyzing utilization data, the investigator must decide whether to use a one-part or two-part model and on what scale to model the data.

A one-part model is fit to the data for all people, despite whether they used any services. The dependent variable could be cost or some transformation such as log (cost + 1). The one-part model is attractive because of its simplicity, but Figures 1 and 2 demonstrate that the data do not have the usually assumed distribution. It is also possible to model costs as having a Poisson, gamma, or negative-binomial distribution. All of these distributions allow for zeroes and for long right tails.

A conceptually attractive way to address the concentration of zero values is a two-part model (10), in which one equation predicts the probability that a person has any use and a second equation predicts the level of use (usually on the log scale) for users only. The expected level of use for an individual is then calculated by multiplying these two estimates together. (This framework has been extended to a four-part model, in which the probability of hospital use is estimated among all users and then the costs for users who were hospitalized and for those who were not are estimated separately.)

Two-part models are attractive because the data then tend to conform to the analytic assumptions (see Figure 2) and also because they provide insight into the utilization process. The decision to have any use at all is most likely made by the person and so is related primarily to personal characteristics, whereas the cost per user may be more related to characteristics of the health care system.

Analysis on the Log Scale

We have noted that utilization data are usually non-normal, right-skewed, and heteroscedastic, with variance that increases with the mean. These features do not necessarily cause problems. If the data set is very large, OLS regression on the untransformed data (including the zeros) will provide unbiased estimates of the regression parameters. The standard errors of regression coefficients may be too small, giving overly significant hypothesis tests; however, in the large data sets often available, significant effects are usually so strong that doubling or tripling the standard error would have little affect on the conclusions (12).

Utilization data have been transformed in various ways to improve their distribution, including logarithmic and square root transformations. It is difficult to interpret the coefficients of a square root equation, and both transformations cause problems if it is necessary to report results on the original dollar scale, because the estimates will be biased.

Analysis of utilization data on the log scale is attractive for the statistical reasons mentioned above. There are also conceptual considerations about whether the relationships being examined are most likely to be additive or multiplicative. For example, suppose costs for a healthy man are \$200 per year, for a healthy woman are \$300, and for a sick man are \$2000. Would we expect the cost for a sick woman to be $\$300 - \$200 = \$100$ more than a sick man (therefore \$2100), or would she cost proportionately more ($300/200 \times 2000 = \$3000$)? If the latter (multiplicative) model is felt to be more realistic, then the log scale is appropriate, because the regression coefficients can be exponentiated to provide estimated ratios of costs (or percentage increases in costs for females compared with males).

There are two ways to analyze the data on the log scale. The first is to analyze the logarithm of cost (for users only or for all subjects after adding some constant) by using OLS. The second is to choose an analytic model that incorporates the logarithmic scale. The generalized linear model permits such an analysis (25). The user chooses a link function that connects the mean and the covariates (such as $\log \mu = \alpha + \beta \mathrm{X}$), and a variance function (such as $\sigma^2 = \mu^2$) and fits a model that has these properties. This link and variance function have been found appropriate for utilization data (3; DK Blough & SD Ramsey, submitted for publication). The gamma distribution has this form. The lognormal model also has this form, although it can not be fit as a generalized linear model. Many statistical software packages, such as SAS, S+, and STATA, can be used to fit these models.

Other Models

Several authors (11, 23, 32) have advocated the use of the Cox proportional hazards model in cost analysis. Cox regression is semiparametric in that the underlying distribution is arbitrary and unspecified, which is attractive because the distribution of costs is skewed and difficult to parameterize. This approach to analyzing costs is valid if the data are not censored (see section on censoring, below). Unfortunately the regression coefficients in the Cox model pertain to the hazard ratio, which makes them quite difficult to interpret in the context of medical costs. Rather than referring to a difference in means or a ratio of costs, the regression coefficients refer to the relationship of the cumulative distributions. For example, if FEMALE was a binary variable in the regression equation and the coefficient of FEMALE was 1.2, this would mean that the proportion of females with costs above "*x*" would be equal to the proportion of males with costs above "*x*," raised to the $e^{1.2}$ power (for all values of *x*). We do not use the Cox model in this article.

Example

We analyzed a random 50% of the annualized outpatient data ($n = 3143$) to illustrate some of these points. The variable of interest is annualized outpatient charges or its logarithm. We fit the dependent variable as a quadratic in age, different for males and females. (This was done only for simplicity of the example, because Figure 2 suggests a more complex relationship with age and sex.) We also included the dummy variable IPA (versus HMO), MONTHS (number of months the person was followed), FAMSIZE (number of people in family), and NCHRONIC (number of chronic conditions).

ONE-PART MODEL The top three lines of Table 1 are regression coefficients and significance levels from three one-part models. (Age and sex are not shown). The first model used cost in dollars as the dependent variable with OLS and achieved $R^2 = 0.07$. The second used OLS to analyze log (cost + 1), with $R^2 = 0.18$. (These values, from models on different scales, are not comparable.) The third fit a gamma distribution to the data by using a log link. The first line shows that enrollees in the IPA averaged \$217 more than those in the HMO, that annualized costs increased \$3.80 for each month the person had been enrolled, that costs decreased \$17 for each additional family member (not significant), and that costs increased \$89 for each chronic disease.

Lines 2 and 3 summarize analyses on the log scale and can be interpreted by exponentiating the regression coefficient. For IPA, $e^{0.52} = 1.7$, which means that, on average, costs in the IPA were 70% higher than in the HMO. The statistical significance of the coefficients in the two models on the log scale is

Table 1 Selected coefficients from regression models*

	IPA	Months	Famsize	NCHRONIC
One-part models				
\$	217*	3.8*	−17.0	89.3*
Log-\$	0.519*	0.058*	−0.029	0.346*
Log-gamma	0.609*	0.011*	−0.042*	0.214*
Two-part models				
Part 1				
Logistic	0.118	0.083*	0.028	0.351*
Part 2				
\$	258*	−1.32	−24.8*	83.1*
Log \$	0.528*	0.001	−0.069*	0.178*
Gamma	0.604*	−.004	−0.047*	0.167*

*, $P < 0.05$.

similar to that on the dollar scale, except for FAMSIZE. Signs of the coefficients were the same in all three models.

TWO-PART MODEL, PART 1 The first part of the two-part model (the probability of having any use) is usually modeled by using logistic or probit regression. If the number of months the subject is observed is not the same for all subjects, it is important to control for length of enrollment (MONTHS) in the regression, perhaps by using length of enrollment or its logarithm as a covariate. The fourth line of Table 1 shows that odds of having any use were significantly related to MONTHS and NCHRONIC, but not to IPA or family size. Note that, in this situation, having utilization is not a "rare" event, and thus the odds ratio cannot be interpreted as a relative risk.

TWO-PART MODEL, PART 2 The bottom part of Table 1 shows the regression equations for part two of the two-part model (for users only). The first model uses OLS to estimate cost in dollars. IPA, NCHRONIC, and FAMSIZE are highly significant, and people who used services in an IPA spend about $258 per year more than those in an HMO. The final two analyses used a lognormal model and a gamma distribution with a log link, respectively. In both models MONTHS was not statistically significant, but the other three variables were. Exponentiating the coefficient for IPA, we find that use is higher in the IPAs than in the HMO by a factor of 1.65 or 1.73 in the two models.

GOODNESS OF FIT The fit of the models to the data can be examined somewhat by examining the "deviance" for the log normal and log gamma models, which is equivalent to the residual sum of squares in least squares and has a somewhat different definition for the gamma model (25). The deviances for the one-part model were 13,156 for the log normal and 33,197 for the gamma model. For part 2 of the two-part model, the deviances were 2936 for log normal and 2997 for the gamma. Because better models have smaller deviances, the gamma distribution fits the data better than the log normal distribution in the one-part model.

HYPOTHESIS TESTING AND PARAMETER ESTIMATION The one-part models in Table 1 all show that the IPA has significantly higher costs. For a two-part model, two hypotheses can be tested. The logistic regression equation showed that the relative odds of having any use were not significantly different in the IPA and HMO. Results for the second part of the model from Table 1 all show that cost for the users is considerably higher for IPA enrollees. Thus, both the one-part and the two-part models found that people in the IPA cost more than those in the HMO. The two-part model provides the additional insight that the enrollees in the two groups were equally likely to use services and that the difference was caused by higher costs in the IPA by those who have services, which is most likely a system effect. This finding for IPAs and HMOs has been noted elsewhere (8).

ESTIMATING COST FOR AN INDIVIDUAL For the one-part dollar-scale model, the estimated cost for an individual is simply his regression estimate. For the one-part models involving log transformations, however, the regression equation predicts log dollars rather than dollars. Exponentiation of a person's estimated log cost provides an estimate of the median cost (geometric mean cost), rather than the arithmetic mean cost, for people with the same set of covariates. To understand this relationship, consider a cost data set with 5 observations: \$1, \$10, \$100, \$1000, and \$10,000. Mean cost is \$2,222 and median cost is \$100. Log cost (base ten) would have the values 0, 1, 2, 3, and 4. The mean log cost is 2.0; $10^2 = \$100$, which is considerably below the mean cost. This will be true whenever the original data have a long right tail. However, if the log transformation succeeded in making the transformed data symmetric, then the mean of the logged data is approximately equal to its median, and the median of the log is the log of the median. For example, the mean (and median) log cost is 2, and $10^2 = 100$, which is the median cost.

Because the mean and median cost are usually quite different, a factor is needed to correct for the retransformation bias. There are two commonly used factors. The first assumes that the data actually have a log normal distribution and is calculated as $\exp[(1 - R^2)s^2_{\mathrm{LOG\ COST}}/2]$. The second is a nonparametric approach called the Duan Smear, which is calculated as $[\sum \exp(\mathrm{e_i})]/N$, where $\mathrm{e_i}$ is the ith residual of the regression on the log scale. For the examples in Table 1, for the one-part model the lognormal factor was 10.6, and the Duan factor was 5.0. For the two-part models, the lognormal factor was 1.91 and the Duan factor was 1.96. The two factors agreed better for the 2-part model because the lognormal assumption was more nearly met. The equation from the gamma model must be retransformed to the log scale, but does not have a retransformation bias because it was modeled on the dollar scale.

For the two-part model, a person's estimated cost is his probability of having any use multiplied by the expected cost conditional on being a user. The probability of having any use is estimated from the logistic regression equation (Table 2). The expected cost, given some use, is estimated by exponentiating the part 2 estimate and multiplying it by the appropriate retransformation factor.

ACCURACY OF PREDICTION The regression equations of Table 1 were estimated from a random 50% of the data. We then used these regression equations to estimate cost for the remaining subset. Table 2 shows two summary measures

Table 2 Root mean square error and mean absolute error in training and validation sample for different models

Model	Dist'n	Bias term	RMSE	MAE
One part	Normal		808	383
	Lognormal	Lognormal	2735	1233
		Duan	1395	599
	Gamma		819	386
Two part	Normal		808	380
	Lognormal	Lognormal	814	375
		Duan	814	378
	Gamma		813	381

of fit for the various models, root mean square error (RMSE), and mean absolute error. To calculate RMSE, we calculated each person's predicted cost, subtracted the observed cost, squared the difference, took the mean, and then took the square root. For the mean absolute error (MAE) we calculated the absolute difference between each person's observed and predicted cost and took its mean. Table 2 shows that RMSE is high for the one-part log models, using either retransformation factor. RMSE is fairly similar for the other six models, and there is no strong difference between the one-part and two-part models. The relative results are similar for the MAE criterion.

Other authors have compared a variety of statistical models for cost data (23; G Gifford, W Manning, M Finch, K Edwards, D Knutson & V Weslowski, submitted for publication; CW Madden, B Mackay, S Skillman, M Ciol & P Diehr, submitted for publication). Lipscomb et al (23) analyzed models built on data from 500,000 Medicare patients and assessed the model by using RMSE, MAE, and also a log score that measures how well the predicted distribution fits the observed distribution. They recommended against using the untransformed data. A study utilizing 10,890 disabled persons recommended models based on untransformed or square root–transformed data rather than log-transformed data (G Gifford, W Manning, M Finch, K Edwards, D Knutson & V Weslowski, submitted for publication). Others (CW Madden, B Mackay, S Skillman, M Ciol & P Diehr, submitted for publication) found little difference in prediction between the one-part and two-part models, using the same data from disabled persons and also data from 40,000 Washington State employees. Because the types of populations studied, the predictor variables, the number of observations, the statistical model, and the measures of "goodness" differed among these studies, there are many possible explanations for discrepancies in the findings. There is no current consensus about the best model.

RECOMMENDATIONS The authors have worked with utilization data for many years and have used a variety of analytic techniques to address a range of questions. The method we recommend depends on the primary goal of the analysis: hypothesis testing to improve understanding of the system, hypothesis testing to explore the net effect of covariates on costs, or estimation of a person's future utilization.

When the goal is understanding the system, a two-part model seems best because it permits the investigator to distinguish factors that affect the propensity to use any services from factors that affect volume of utilization once the person has entered the system. The easiest two-part model uses the lognormal distribution and is solved by using OLS. In the example above, we found that the IPA and HMO did not differ on the proportion of subjects using services, but that the cost per user was about twice as high in the IPA as in HMOs.

For understanding the effect of individual covariates on total costs, a one-part model is most useful because it generates a single regression coefficient for each variable and so can be interpreted easily. The one-part model on the dollar scale is easy to interpret but assumes an additive model. The gamma distribution might be preferred because it is a multiplicative model.

If the goal is prediction of future costs, we recommend a one-part model, with untransformed cost as the dependent variable, estimated by using least squares. This model works about as well as the two-part models and does not require retransformation. It may seem puzzling that the models that are most similar to the true distribution of the data do not give better predictions, but we and others have found this to be true in a variety of situations, including the example in Table 2. Further research and better algorithms may change these recommendations in the future.

DEATH AND CENSORING

Patients who die shortly after entering a study often use the fewest resources, whereas those who remain alive for a time but eventually die are among the most expensive. This suggests a potentially peculiar relationship in which the sickest subjects have either very high or very low costs—a U-shaped relationship that should be addressed during modeling. Depending on the goals of the analysis, a person who dies may not present a problem, because all of his utilization is known. Calculation of annualized rates can be a problem, however, because there is usually high utilization in the last year of life.

Censoring is an issue in estimating the average lifetime cost for treating a particular disease (or similarly the cost until cure, or the cost in a 12-month period). Most often, the complete costs for some subjects are not fully observed because the subjects are lost to follow-up or because they are still alive (or not cured or discharged or have not been enrolled for 12 months) at the time of data analysis. In the terminology of survival analysis, the lifetime costs for such subjects are censored. The estimated cost will be too low if one analyzes all the available cost data. Alternatively, calculating costs for only the uncensored subjects is likely to give too much emphasis to subjects who died early, because people who survive for a longer time are more likely to be censored.

Some investigators have applied standard survival analysis methods such as Kaplan-Meier estimators, log rank tests, and Cox regression to the problem of cost evaluation under censoring, by analyzing censored costs as though they were censored survival times (11, 16, 18, 30). Unfortunately, this strategy is generally invalid. The main problem is the requirement in standard survival analysis that the time of death and the corresponding time of censoring must be independent. When standard survival analysis methods are applied to censored costs, the time of death corresponds to the total cost at the time of death, and the time of censoring corresponds to the total cost at the time of censoring; we observe the former if the subject is uncensored and observe the latter otherwise. Unfortunately, the total cost at the time of death is not independent of the total cost at the time of censoring, even if the time of death and time of censoring are themselves independent. A subject who has high costs per month will usually have high total costs at both the time of death and time of censoring, whereas a subject with low costs per month will tend to have low total costs at both times. Thus, the total cost at the time of death tends to be positively correlated with the total cost at the time of censoring. This likely violation of the independent censoring assumption implies that censored costs should not be analyzed by standard survival methods.

To minimize the bias induced by censoring, Lin et al (22) propose partitioning the entire time period of interest into a number of small intervals and then estimating the average lifetime cost either by the sum of the Kaplan-Meier estimator for the probability of dying in each interval multiplied by the sample mean of the total costs for the deaths in that interval or by the sum of the Kaplan-Meier estimator for the probability of being alive at the start of each interval multiplied by an appropriate estimator for the average cost over the interval conditional on surviving to the start of the interval. This approach has since been addressed by others (15; RD Etzioni, EF Feuer, SD Sullivan, DY Lin, C Hu & SD Ramsey, submitted for publication).

Another approach is a modification of an estimate for quality-adjusted survival time (35). The Zhao-Tsiatis estimator applied to costs would be the sample proportion of those lifetime costs, which are known to be greater than *x* dollars, with contributions weighted inversely by the probabilities of inclusion. These two methods were originally developed to estimate the lifetime cost for a single group of patients, but can be easily extended for comparing two groups of patients. Regression techniques, which would allow one to study simultaneously various risk factors on the medical cost, are not currently available.

CONCLUSION

We have presented the traditional methods of analyzing health care utilization data and pointed out some areas in which new statistical methods are being applied. We do not expect major analytic improvements to result from applying different regression models to the data. It is more likely that new research in the area of censored cost data will yield improved understanding of the problem and that software will be developed to permit covariates to be incorporated in the analysis of such data.

Literature Cited

1. Ash A, Porell F, Gruenberg L, Sawitz E, Beiser A. 1989. Adjusting Medicare capitation payments using prior hospitalization. *Health Care Financ. Rev.* 10(4):17–29
2. Berlin MF, Faber BP, Berlin LM, Budzynski MR. 1997. RVU costing applications. *Health Care Financ. Manage.* 51:73–76
3. Blough DK, Madden CW, Hornbrook MC. 1999. Modeling risk using generalized linear models. *J. Health Econ.* 18:153–71
4. Cohen J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Erlbaum. 567 pp.
5. Cohen SB. 1997. *Sample Design of the 1996 Medical Expenditure Panel Survey Household Component. MEPS Methodol. Rep. No. 2. AHCPR Pub. No. 97-0027.* Rockville, MD: Agency Health Policy Res.
6. Diehr P. 1980. *Sample Size Calculation and Optimal Follow-Up Time in Health Services Research Using Utilization Rates. Tech. Rep. No. 42.* Seattle: Dep. Biostat., Univ. Washington
7. Diehr P, Madden C, Martin DP, Patrick DL, Mayers M, et al. 1993. Who enrolled in a state program for the uninsured: Was there adverse selection? *Med. Care* 31:1093–105
8. Diehr P, Martin D, Price K, Friedlander L, Richardson W, Riedel D. 1984. Use of ambulatory care services in three provider plans: interactions between patient characteristics and plans. *Am. J. Public Health* 74:47–51
9. Duan N. 1983. Smearing estimate: a nonparametric retransformation method. *J. Am. Stat. Assoc.* 78:605–10
10. Duan N, Manning WG, Morris CN, Newhouse JP. 1983. A comparison of alternative models for the demand for medical care. *J. Bus. Econ. Stat.* 1:115–26
11. Dudley RA, Harrell FE, Smith LE, Mark DB, Califf RM, et al. 1993. Comparison of analytic models for estimating the effect of clinical factors on the cost of coronary artery bypass graft surgery. *J. Clin. Epidemiol.* 46:261–71
12. Ellis RP, Ash A. 1995/1996. Refinements to the diagnostic cost group (DCG) model. *Inquiry* 32:418–29
13. Ellis RP, Pope G, Iezzoni LI, Ayanian JZ, Bates DW, et al. 1996. Diagnosis-based risk adjustment for Medicare capitation payments. *Health Care Financ. Rev.*
14. Epstein AM, Cumella E. 1988. Capitation payment: using predictors of medical utilization to adjust rates. *Health Care Financ. Rev.* 10(1):51–70
15. Etzioni R, Urban N, Baker M. 1996. Estimating the costs attributable to a disease with application to ovarian cancer. *J. Clin. Epidemiol.* 49:95–103
16. Fenn P, McGuire A, Phillips V, Backhouse M, Jones D. 1995. The analysis of censored treatment cost data in economic evaluation. *Med. Care* 33:851–63

17. Gatsonis C, Normand SL, Liu C, Morris C. 1993. Geographic variation of procedure utilization: a hierarchical model approach. *Med. Care* 31:YS54–59
18. Hiatt RA, Quesenberry CP, Selby JV, Fireman BH, Knight A. 1990. The cost of acquired immunodeficiency syndrome in Northern California: the experience of a large prepaid health plan. *Arch. Intern. Med.* 150:833–38
19. Iezzoni LI, ed. 1997. *Risk Adjustment for Measuring Health Care Outcomes*. Ann Arbor, MI/Chicago: Health Admin. 610 pp.
20. Kronick R, Dreyfus T, Lee L, Zhou Z. 1996. Diagnostic risk adjustment for Medicaid: the Disability Payment System. *Health Care Financ. Rev.* 17:7–33
21. Liang KY, Zeger S. 1993. Regression analysis for correlated data. *Annu. Rev. Public Health* 14:43–68
22. Lin DY, Feuer EJ, Etzioni R, Way Y. 1997. Estimating medical costs from incomplete follow-up data. *Biometrics* 53:419–34
23. Lipscomb J, Ancukiewica M, Parmigiani G, Hasselblad V, Samsa G, Matchar DB. 1998. Predicting the cost of illness: a comparison of alternative models applied to stroke. *Med. Decis. Making* 18 (Suppl.): S39–56
24. Martin D, Diehr P, Cheadle A, Madden C, Patrick D, Skillman S. 1997. Health care utilization for the "newly insured": results from the Washington Basic Health Plan. *Inquiry* 34:129–142
25. McCullagh P, Nelder J. 1983. Generalized linear models. New York: Chapman & Hall. 261 pp.
26. McGlynn EA, Danberg C, Kerr EA, Brook RH. 1999. *Health Information Systems: Design Issues and Analytic Applications, RAND MR-967-HF*. Santa Monica, CA: RAND.
27. Newhouse JP. 1982. Is competition the answer? *J. Health Econ.* 1:109–15
28. Newhouse JP, Manning WG, Keeler EB, Sloss EM. 1989. Adjusting capitation rates using objective health measures and prior utilization. *Health Care Financ. Rev.* 10:41–54
29. O'Brien BJ, Drummond MF, Labelle RJ, Willan A. 1994. In search of power and significance: issues in the design and analysis of stochastic cost-effectiveness studies in health care. *Med. Care* 32:150–63
30. Quesenberry CP, Fireman B, Hiatt RA, Selby JV. 1989. A survival analysis of hospitalization among patients with acquired immunodeficiency syndrome. *Am. J. Public Health* 79:1643–47
31. Salem-Schatz S, Moore G, Rucker M, Pearson SD. 1994. The case for case-mix adjustment in practice profiling: when good apples look bad. *J. Am. Med. Assoc.* 272(11):871–74
32. Siegel C, Alexander MJ, Lin S, Laska E. 1986. An alternative to DRGs: a clinically meaningful and cost-reducing approach. *Med. Care* 24(5):407–17
33. Starfield B, Weiner J, Mumford L, Steinwachs D. 1991. Ambulatory care groups: a categorization of diagnoses for research and management. *Health Serv. Res.* 26:53–74
34. Weiner JP, Starfield BH, Steinwachs DM, Mumford LM. 1996. Development and application of a population-oriented measure of ambulatory care case-mix. *Med. Care* 29:452–72
35. Zhao H, Tsiatis AA. 1997. A consistent estimator for the quality adjusted survival time. *Biometrika* 84:339–48

Annual Review of Public Health
Volume 20, 1999

CONTENTS

PUBLIC HEALTH IN THE TWENTIETH CENTURY: Advances and Challenges, *Jonathan E. Fielding* xiii

Personal Reflections on Occupational Health in the Twentieth Century: Spiraling to the Future, *Mark R. Cullen* 1

A Future for Epidemiology?, *S. Schwartz, E. Susser, M. Susser* 15

Lessons in Environmental Health in the Twentieth Century, *Michael Gochfeld, Bernard D. Goldstein* 35

US Health Care: a Look Ahead to 2025, *Eli Ginzberg* 55

What Have We Learned About Public Health in the Twentieth Century: A Glimpse Through Health Promotion's Rearview Mirror, *L. W. Green* 67

Understanding Changing Risk Factor Associations with Increasing Age in Adults, *G. A. Kaplan, M. N. Haan, R. B. Wallace* 89

Advances in Clinical Trials in the Twentieth Century, *Lloyd D. Fisher* 109

Methods for Analyzing Health Care Utilization and Costs, *P. Diehr, D. Yanez, A. Ash, M. Hornbrook, D. Y. Lin* 125

Time-Dependent Covariates in the Cox Proportional Hazards Regression Model, *Lloyd D. Fisher, D. Y. Lin* 145

Lessons from 12 Years of Comparative Risk Projects, *Ken Jones, Heidi Klein* 159

Unexplained Increases in Cancer Incidence in the United States from 1975 to 1994: Possible Sentinel Health Indicators?, *Gregg E. Dinse, David M. Umbach, Annie J. Sasco, David G. Hoel, Devra L. Davis* 173

Eradication of Vaccine-Preventable Diseases, *A. Hinman* 211

Immunization Registries in the United States: Implications for the Practice of Public Health in a Changing Health Care System, *David Wood, Kristin N. Saarlas, Moira Inkelas, Bela T. Matyas* 231

Teen Pregnancy Prevention: Do Any Programs Work?, *Josefina J. Card* 257

The Social Environment and Health: A Discussion of the Epidemiologic Literature, *I. H. Yen, S. L. Syme* 287

Health Status Assessment Methods for Adults: Past Accomplishments and Future Challenges, *Colleen A. McHorney* 309

Patient Outcomes Research Teams: Contribution to Outcomes and Effectiveness Research, *Deborah Freund, Judith Lave, Carolyn Clancy, Gillian Hawker, Victor Hasselblad, Robert Keller, Ellen Schneiter, James Wright* 337

Pharmacy Benefit Management Companies: Dimensions of Performance, *Helene L. Lipton, David H. Kreling, Ted Collins, Karen C. Hertz* 361

The Key to the Door: Medicaid's Role in Improving Health Care for Women and Children, *Diane Rowland, Alina Salganicoff, Patricia Seliger Keenan* 403