New models for old questions: generalized linear models for cost prediction

John L. Moran MBBS FRACP FJFICM,[1] Patricia J. Solomon PhD BSc,[2] Aaron R. Peisach MBBS FJFICM[3] and Jeffrey Martin BApSc[4]

[1]Senior Consultant, [3]Director, Department of Intensive Care Medicine, The Queen Elizabeth Hospital, Woodville, SA Australia
[2]Associate Professor, School of Mathematical Sciences, The University of Adelaide, Adelaide, SA, Australia
[4]Business Manager, Critical Care Business Management, The Queen Elizabeth Hospital, Woodville, SA, Australia

Keywords
costs, logarithmic transformation, generalized linear models, inverse Gaussian distribution, residual analysis

Correspondence
Dr John L. Moran
Department of Intensive Care Medicine
The Queen Elizabeth Hospital
28 Woodville Road
Woodville
SA 5011
Australia
E-mail: john.moran@nwahs.sa.gov.au

Accepted for publication: 23 January 2006

doi:10.1111/j.1365-2753.2006.00711.x

Abstract

Background Generalized linear models (GLMs) have recently been introduced into cost data analysis. GLMs, transformations of the linear regression model, are characterized by a particular response distribution from one of the exponential family of distributions and monotonic link function which relates the response mean to a scale on which additive model effects operate.

Objectives This study compared GLMs and ordinary least squares regression (OLS) in predicting individual patient costs in adult intensive care units (ICUs) and sought to define the utility of the inverse Gaussian distribution family within GLMs.

Methods A prospective 'ground-up' utilization costing study was performed in three adult university associated ICUs, enrolling consecutive ICU admissions over a 6-month period in 1991. ICU utilization, patient demographic and ICU admission day data were recorded by dedicated data collectors. Model performance was assessed by prediction error [mean absolute error (MAE), root mean squared error (RMSE)] and residual analysis.

Results The cohort, 1098 patients surviving ICU, was of mean (SD) age 56 (19.5) years and 41% female. Patient costs per ICU episode (1991 A$) were A$6311 (9689), with range A$106 to A$95602. Prediction error for mean costs was minimal (MAE 4780; RMSE 8965) with OLS using heteroscedastic retransformation of log costs and GLM with Gaussian family and log link (MAE 4798; RMSE 8907). Residual analysis suggested optimal overall performance for the above two models and a GLM with inverse Gaussian family and log link.

Conclusions Traditional cost models of OLS with (log) cost transformation may be supplemented by appropriately specified GLM which more closely model the error structure.

Introduction

Medical cost data are usually right skewed with variability increasing as the mean costs increases. The traditional model for cost prediction has been multivariable ordinary least squares regression (OLS) [1–3], with or without initial transformation, usually logarithmic, of the dependent cost variable [4]. As previously noted by Chhikara & Folks [5], the use of transformations suggested by the data still leaves the problem of interpretation of the results of analysis. Analysis on transformed scales does not 'provide inferences about population mean costs which are of primary interest' [6]. Thus, 'simple' logarithmic transformation has attendant problems in terms of both the appropriate back transformation into the original scale (i.e. in this study, Australian dollars [A$]) [7] and the interpretation of regression coefficients [8]. Recently, a new class of predictive models, generalized linear models (GLMs), have been introduced into the analysis of cost data [9–11]. GLMs are empirical transforms of the classical linear (Gaussian) regression model and are distinguished from OLS by particular model, rather than data, transformations: specifically, a response distribution of one of the exponential family of distributions (normal, Poisson, gamma, binomial, inverse Gaussian) and a (monotonic) link function (identity, logarithmic, square root, logistic, power) which relates the mean of the response to a scale on which the model effects combine additively [11]. It has been suggested that health care expenditure and use data frequently have a log-normal or gamma distribution and the studies using GLM for cost analysis have focused on the gamma response distribution [6,10]. However, the shape of the inverse Gaussian distribution, with a high initial peak and long right tail [5], may recommend its use for cost data.

The purpose of this paper was to compare the performance of OLS, various GLMs [specific combinations of distribution (fam-

ily) and link] in the analysis of individual patient costs derived from a 'ground-up' ICU utilization study and to answer the question: do GLMs, in particular a GLM using the inverse Gaussian distribution response distribution, have particular advantage when analysing medical cost data? Performance was adjudged using established indices [mean absolute error (MAE), root mean squared error (RMSE) and various coefficients of determination (R^2)] and graphical residual analysis [2,12].

Methods

Data sources and settings

Cost data for ICU patient stay, including all related management activity, but excluding costs associated with provision of services external to the ICU, was generated from a 3-month study (1991) in three South Australian adult ICUs; an in-detail analysis of this data has recently been reported [13]; where separate predictive models for survivors and non-survivors were presented. For all patients (n = 1333) [13] recorded total (1991) mean (SD) patient costs per ICU episode as A$6801 (10 311) with ICU length of stay 3.9 (6.1) days; using standard inflation adjustments, mean calendar year 2002 costs were A$9343. The computed (2002) occupied ICU bed-day costs at A$ 2395 were, as noted by Rechner & Lipman [14] quite similar to two recent studies: A$2670 for calendar year 2003 [14], from Australia, and approximately A$2400 for financial year 2000–01 [15], from the UK.

Data collection [13]

In each ICU, dedicated unit data collectors recorded daily activity and utilization. The specific utilization elements were: (i) drugs – data on actual drug usage, including parenterally administered fluids, were collected daily; (ii) procedural – medical and surgical supplies, all medical and surgical supplies were identified and recorded, by procedure or by individual item; (iii) pathology costs – all pathology tests consumed were recorded by individual patient and were costed using the current Commonwealth Government of Australia Benefits Schedule reimbursement rates; (iv) radiology costs – were recorded by individual patient and were costed using procedure costs developed by the South Australian Government Health Commission; (v) physiotherapy costs – each physiotherapy intervention was recorded by individual patient and costed using a standard unit of time; (vi) nursing staff costs – nursing salary and wage costs were derived using actual minutes of nursing time for each ICU patient day (time spent on educational activities was excluded), standard nursing practice was 1–1 nurse patient ratio; (vii) medical staff costs – medical salary costs were allocated to patients on the basis of days of ICU stay (time spent on educational activities was excluded), all medical staff were 'full-time'; (viii) overhead costs – overhead costs attributable to the operation of each ICU were derived using the Yale Diagnostic Related Group (DRG) costing methodology [16], and allocated to patients on the basis of ICU length of stay; and (ix) other costs – these were the residual costs reported in the ICU cost centre that remained unallocated to patients (such as, administration, repairs and maintenance, orderlies salaries and wages, linen and domestic supplies) and were allocated to patients on the basis of ICU length of stay. Re-admissions were included in the study and each stay was costed individually. Total costs (1991 A$) were computed as the sum of various cost fractions: (i) medication and procedural, (ii) nursing, physiotherapy and medical, (iii) radiology and pathology, and (iv) overhead and other; individual (patient) day costs were not available for analysis.

Additional patient data recorded included: (i) demographics – age, gender, ethnicity, co-morbidities consistent with the APACHE III algorithm [17]; (ii) ICU stay variables – patient source, admission diagnosis and principal physiological system dysfunction on admission, ventilatory status, cardiorespiratory (heart and respiratory rate, systolic and diastolic blood pressure), arterial blood gas (pH, PaO_2, $PaCO_2$) and biochemical variables such that an APACHE III score could be computed, ICU length of stay and outcome; and (iii) hospital stay variables – treating hospital, DRG, hospital length of stay and outcome. Categorical variables were score as 0/1, 0/1/2 as indicated. For the purposes of this analysis: (i) only ICU survivors were considered; (ii) potential predictor variables were drawn from demographic and ICU admission day data only; and (iii) two extreme (cost) outliers (ICU costs > A$ 100 000) and a single case with incomplete first day data were not considered.

Statistical analysis

Variables were reported as mean (SD) unless otherwise indicated; Stata® statistical software (Version 9.0 2005; Stata Corp, College Station, TX) was used. Probability plots (P-P) were initially used to compare the cost distribution with hypothesized distributions (normal, lognormal, gamma and inverse Gaussian). If $x_1, x_2, \ldots x_n$ is the ordered sample (size n) from a distribution with location and scale parameters α and β and F is the cumulative distribution function, the P-P plots $Z_i = F([X_i - \hat{u}]\hat{\sigma})$ against p_i, where $\hat{u}$ and $\hat{\sigma}$ are estimators of location and scale, respectively, and p_i are plotting positions [18].

Multivariable models to predict total costs were as follows: OLS; OLS with log transformation of costs and back-transformations of log-costs as: (i) simple exponential, (ii) 'naïve', that is the exponential of (predicted costs + 0.5*(RMSE)^2, where RMSE = square root of the mean square error of the OLS equation, (iii) Duan's smearing estimate [7] and (iv) heteroscedastic retransformation [10,19]; (v) GLM with Gaussian family and log link; (vi) GLM with gamma family and log link; and (vii) GLM with inverse Gaussian family and log link [12].

Variable selection from a full model used the Akaike information criterion ($AIC = -2(L) + 2(c + p + 1)$), where L is the log-likelihood, c is the number of model covariates and p is the number of model-specific ancillary parameters [20]. Specific attention was directed to both the question of model selection with correlated variables, and the potential effect of multi-colinearity (variance inflation factor [VIF] < 10 and condition number [CN] < 15). Non-linearity of covariate effect was investigated by using (parametric) fractional polynomials [21] and all first order interactions were explored. Model performance was variously assessed:

(i) Quantitative predictive indices
   a. MAE as mean of absolute difference between observed and predicted cost.
   b. RMSE as predicted cost minus observed, square of the difference, mean of the squared difference and square root of this value.

c. Correlation (Pearson, *rho*) of cost and predicted cost (22) with 95% bootstrap (BCa [23]) confidence intervals using 1000 repetitions.

d. Squared correlation (R^2), ensuring that scalar values of R^2 were compared on the same scale (24).

e. A 'pseudo-R^2' statistic from the GLM literature, the Ben-Akiva and Lerman adjusted likelihood-ratio index = $(1-(L(M_\beta-k)/L(M_\alpha))$, where M_β is the log-likelihood of model with intercept and predictors, k is the number of model parameters and M_α is the log-likelihood of model with intercept only. (12)

f. Lin's concordance correlation coefficient (*rho_c*, to be distinguished from Pearson's correlation coefficient) of cost and predicted cost, with 95% BCa CIs. As noted by commentators, correlation (*rho*) implies data for two variables Y_1 and Y_2 lying on a line $Y_1 = \alpha + \beta Y_2$ and large values of *rho* may occur with $\alpha \neq 0$ and/or $\beta \neq 1$ (that is, in the absence of 'perfect' agreement between Y_1 and Y_2). *rho_c* assesses the agreement between two paired sets of measurements by measuring the variation from the 45° line of identity [25].

g. For the OLS models, formal tests for heteroscedasticity (non-constant variance [26]) were performed; Breusch-Pagan/Cook-Weisberg, Szroeter and a likelihood ratio test for group-wise heteroscedasticity.

(ii) Graphical analysis using Anscombe residuals [12]

a. Residual plots versus fitted values, looking for even distribution of the residuals about the y = 0 line.

b. Standardized normal probability plots (P-P plots, focusing on centre of the distribution) and inverse normal quantile plots (Q-Q plots, emphasizing the tails of the distribution) of the residuals, looking for close approximation to the 45° line of identity [18].

c. Plots of residuals to assess residual heteroscedasticity [12,27]; heteroscedasticity was adjudged by the degree of slope (away form the horizontal) of the lowess (locally weighted scatter plot smoothing [28]) plot line relating the SD of the residuals to the mean values of grouped fitted values and grouped APACHE III scores looking for lack of trend.

d. Plots in a. to c. above were compared with those using deviance residuals.

Results

The cohort consisted of 1098 patients of mean (SD) age 56 (19.5) years and 41% were female. Further patient details are shown in Table 1. Total costs (1991 A$) were A$6311 (9689) with a range of A$106 to A$9 5602. The distribution showed marked kurtosis and skewness ($P = 0.0001$) and log transformation did not yield a normal distribution (Shapiro–Wilk W-test, $P = 0.0001$), albeit the kurtosis was modified ($P = 0.44$). Figure 1 shows: (i) in the upper panel, a probability (P-P) plot [21] of gamma and inverse Gaussian distributions generated from the total cost data, in particular, for the (two parameter) gamma distribution, the shape parameter (*alpha*) = 0.953 and the scale parameter (*beta*) = 6604; and for the inverse Gaussian distribution, mean (*mu*) = 6311 and *lambda* = 2677, where variance is $mu^3\ lambda^{-1}$ and (ii) in the lower panel, quantile-quantile (Q-Q) plots of the above two generated distributions against total costs. Total costs were better approximated (clustering of data points about the 45° line of identity) by the inverse Gaussian distribution; no routine transformation of costs [29] yielded a normal distribution.

Table 1 Patient demographics: mean (SD) or absolute numbers as indicated

Variable	
n	1098
Age (years)	56 (19.5)
APACHE III score	51 (22.5)
Hospital (patient number)	
1	415
2	257
3	426
Gender (female/male; *n* =)	447/651
Ventilated (*n* =)	552
Chronic dialysis (*n* =)	10
COPD (*n* =)	14
Hepatic failure (*n* =)	6
Metastatic carcinoma (*n* =)	26
ICU length of stay (days)	2 (0.5–67)*
Hospital length of stay (days)	16 (0.5–248)*

COPD, chronic obstructive pulmonary disease; ICU, intensive care unit. *Median (range).

Model covariates and performance indices are seen in Table 2. Consistency of covariate selection for APACHE III score, ventilation and hospital source was demonstrated across all models, with chronic obstructive pulmonary disease and chronic dialysis being the next most frequent selections. No significant interactions were demonstrated and continuous variables (APACHE III score and age) demonstrated consistent linear effects. Mean predicted costs, MAE and RMSE varied considerably across models (Table 2). Of note was the severe under-prediction of mean costs by simple exponentiation in the OLS – log costs model and modest over prediction of mean costs by GLM – inverse Gaussian family and log link. The range of total costs was considerable, A$106 to A$95 602; only three models had predicted costs > A$35 000; OLS – log costs with heteroscedastic retransformation, GLM – Gaussian family and log link, and GLM – inverse Gaussian family and log link. MAE and RMSE were minimal using OLS – log costs with heteroscedastic retransformation, and GLM – Gaussian family and log link. Correlation (observed vs. fitted costs) and R^2 were best with the GLM – Gaussian family and log link, and OLS – log costs with back transformation. Lin's concordance correlation coefficient (observed vs. fitted costs) suggested best performance with GLM – Gaussian family and log link, GLM – inverse Gaussian family and log link, and OLS: log cost with heteroscedastic retransformation. Considerable variation in concordance was observed between the various back transformations of the OLS log cost model.

Overall, model performance (systematic departure from model assumptions) was assessed by inspection of plots of residuals against fitted values (shown in Fig. 2) and standardized normal probability (shown in Fig. 3) and inverse normal quantile plots of residuals. Symmetrical distribution (residuals vs. fitted values) and normality of residuals (probability and quantile plots), suggesting optimal model performance, was observed for OLS – log costs,

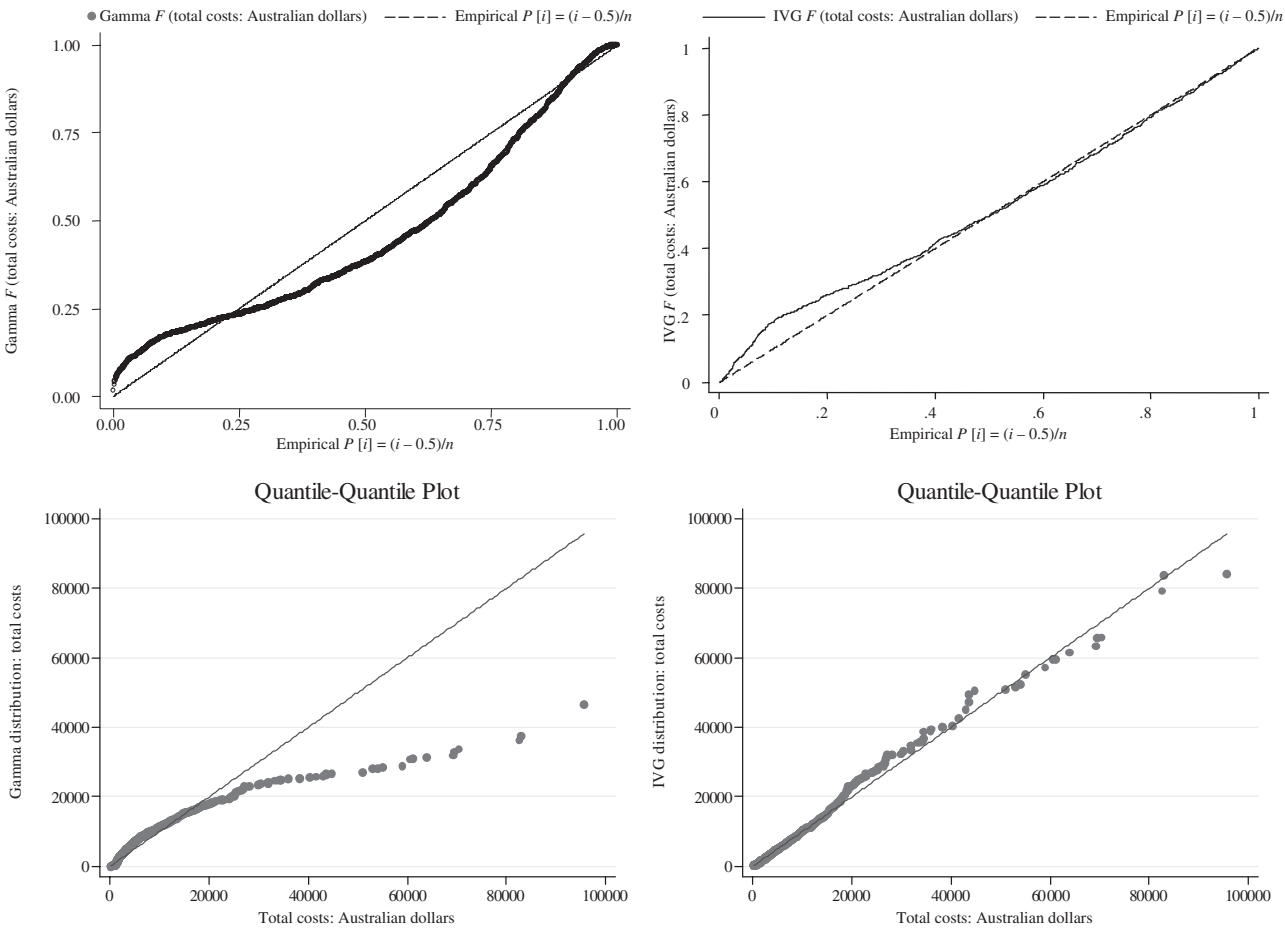

Figure 1 Probability and quantile-quantile plots for gamma and inverse Gaussian cost distributions. Upper panel: probability plot (P-P) of two parameter gamma (left) and inverse Gaussian (right) distributions against costs. Vertical axis (cumulative) probability, 0–1; horizontal axis, Hazen plotting position (=(i – 0.5)/n, where i = rank and n = count [18]. Lower panel: ordered quantile plots of distributions (gamma, left and inverse Gaussian, right) generated from total costs (vertical axis) against total costs (horizontal axis). IVG, inverse Gaussian.

and GLM – inverse Gaussian family and log link. The only models to reasonably satisfy homoscedasticity (constant variance assumption) were GLM – Gamma family and log link, and OLS – log costs, although all models appeared suspect (Fig. 4). This being said, tests of heteroscedasticity identified significant overall ($P = 0.001$) and covariate specific [APACHE III score ($P = 0.001$), ventilation status ($P = 0.001$)] heteroscedasticity for both OLS and OLS – log costs. No differential diagnostic sensitivity in the plots between Anscombe and deviance residuals was noted.

Discussion

The models considered above addressed the estimation of mean or total costs using particular covariate sets (conditional mean modelling [10]); formally, the estimation of $E(y|x)$. Although the dependent variable (y) was positively skewed, estimation of median costs was not considered, as this would have been less relevant to ICU administrative concerns, which focus on total costs = average costs × number of patients. The relevance of the actual costs has been detailed above (*Data sources and settings*).

Distributions and transformations

The traditional model for skewed health data is one of logarithmic transformation of the dependent variable [2,30–33]. Such a transformation usually induces symmetry rather than normality into the cost variable; if the variance-mean relationship is a power (square) function, logarithmic transformation serves to stabilize variance (homoscedasticity). OLS with a logged dependent variable ($\log(y)$) is contingent upon a linear relation of mean $\log(y)$ to the covariates and the constancy of variance, not necessarily normality. This being said, inference is on the log-dollar scale [34]. Logarithmic transformation results in comparison of geometric means and inference in comparing such means cannot be equated with a test of arithmetic means unless log-scale variances (between groups) are equal [30]. Back transformation to the original scale of the dependent variable (in this case, Australian dollars) is not simply a matter of exponentiation. As seen from Table 2, the concordance (*rho_c*) of total costs with predicted costs for the OLS log costs model is dependent upon the method of back transformation, with *rho_c* varying from 0.146 with simple expo-

Table 2 Total and predicted costs (A$) and model performance indices

	Covariates	Mean	SD	MAE	RMSE	Corr (95% CI)	*rho_c* (95% CI)	R^2	BAL
Total costs		6311	9689						
OLS	APIII, metca age, vent, copd	6311	3313	4995	9101	0.342 (0.284–0.413)	0.209 (0.159–0.256)	0.117	0.006
OLS: log, exp	Hosp, AP3, age, vent copd, metca, hfail	3936	2131	4242	9420	0.369 (0.283–0.290)	0.146 (0.11–0.197)	0.136	0.106
OLS: log, naïve	Hosp, AP3, age, vent copd, metca, hfail	5902	3195	4753	9018	0.369 (0.283–0.290)	0.219 (0.167–0.295)	0.136	0.106
OLS: log, Duan	Hosp, AP3, age, vent copd, metca, hfail	6298	3410	4914	9002	0.369 (0.283–0.290)	0.231 (0.176–0.309)	0.136	0.106
OLS: log, het	Hosp, AP3, age, vent copd, metca, hfail	6037	3753	4780	8965	0.379 (0.295–0.499)	0.255 (0.189–0.363)	0.144	0.106
GLM: gausslog	Hosp, AP3, age, vent copd, metca, hfail	6121	4102	4798	8907	0.396 (0.304–0.517)	0.283 (0.202–0.402)	0.155	0.009
GLM: gamlog	Hosp, AP3, vent	6368	3540	4990	9077	0.349 (0.283–0.454)	0.225 (0.171–0.295)	0.122	0.014
GLM: ivglog	Hosp, AP3, vent, cdial	6805	4467	5198	9136	0.353 (0.280–0.468)	0.268 (0.202–0.369)	0.124	0.0004

SD, standard deviation; MAE, mean absolute error; RMSE, root mean squared error; Corr, Pearson correlation with 95% bootstrap (BCa) CI; *rho_c*, Lin's concordance correlation coefficient with 95% bootstrap (BCa) CI; R^2, coefficient of determination; BAL, Ben-Akiva and Lerman adjusted likelihood ratio index; Hosp, hospital source; AP3, APACHE III score; age, age in years; Vent, ventilation on ICU admission day; copd, history of chronic obstructive pulmonary disease; metca, evidence of metastatic carcinoma; hfail, hepatic failure; cdial, chronic dialysis; OLS, ordinary least squares regression; OLS, log, exp, ordinary least squares regression using log transformed costs and exponential back transformation; OLS: log, naïve, ordinary least squares regression using log transformed costs and naive back transformation; OLS: log, Duan, ordinary least squares regression using log transformed costs and Duan's smearing back transformation; OLS: log, het, ordinary least squares regression using log transformed costs and heterscedastic back transformation; GLM: gausslog, generalized linear model with Gaussian family and log link; GLM: gamlog, generalized linear model with gamma family and log link; GLM: ivglog, generalized linear model with inverse Gaussian family and log link.

nentiation, 0.219 with 'naïve' transformation, 0.231 with Duan's smearing estimate and 0.255 for heteroscedastic retransformation. For OLS with normally distributed homoscedastic errors, the recommended transformation is: exponential (fitted values + 0.5*$(\text{RMSE})^2$). For non-normally distributed, homoscedastic errors, the smearing estimate ($\hat{\phi}$) has superior properties and is given by: $\hat{\phi} = \frac{1}{N}\sum_{i=1}^{n}\exp(\hat{\varepsilon}_i)$, where ε are the absolute residuals from the OLS regression (33); the predicted costs are then calculated as $E(y) = \hat{\phi}.\exp(X\beta)$, where $X\beta$ is the OLS linear predictor. In the current study, $\hat{\phi} = 1.60$. For normally distributed heteroscedastic residuals: exponential {fitted values + 0.5* [log scale variance function, $v(x)$]}, where $v(x)$ is the variance of the log-scale, obtained by regressing the squared residuals on the covariates [19]. Thus a seemingly 'simple' log transformation entails a rather complex model dependent re-transformation process to recover costs in the original scale, without incurring bias.

Similarly, the interpretation of regression coefficients with log transform of the dependent variable is not facile: for the homoscedastic normal regression, the effect on the (untransformed) dependent variable is in terms of a percentage change = $100 \times (\exp(\beta) - 1)$, where β is the (independent) variable regression coefficient, continuous or categorical [8,35]. For heteroscedastic regression, where the covariate (x_j) also appears in the variance model ($\sigma^2 = \exp(\gamma' x_j)$), covariate effect is somewhat more complex, as developed by Zhou *et al*. [35], with different interpretations of unit change of dependent variable for categorical and continuous covariates.

The gamma distribution, most useful with positive responses (≥0) having a constant coefficient of variation, has also been suggested as an appropriate distribution with which to model costs [2,9,10], but the current total cost distribution was poorly approximated by this distribution (Fig. 1). In a recent empirical investigation of costs generated from a randomized trial, the gamma distribution was found to be the most appropriate, based primarily upon analysis of residuals; no initial approximation of the cost distribution to the exponential family of distributions was provided [6]. Such was not the case in the current study, where the total cost distribution was poorly approximated by the gamma distribution (Fig. 1). The inverse Gaussian distribution, with a high initial peak with rapid drop-off and long right tail, would appear to adequately reflect cost and length of stay distributions, although little has been published on this [5]. A previous paper, applying the inverse Gaussian distribution to length of stay, used the method of sample quantiles (agreement of fitted and observed distributions at specific quantiles) [36], but did not model the length of stay. This being said, regression models appropriate for cost data are not necessarily optimal for length of stay prediction [37].

Generalized linear models

The generalized linear model, introduced by Nelder & Wedderburn [38] and first implemented in the statistical software GLIM, synthesizes the general techniques used to analyse continuous and discrete data into a unified conceptual framework [39]. Explanatory features are combined additively (see *Introduction*) as in classical linear models; the properties of the response variable are matched by the particular distribution (any of the exponential family of distributions, including the gamma and inverse Gaussian); the variance is a function of the mean ($\text{var}(y|x = \sigma^2 v(x))$),

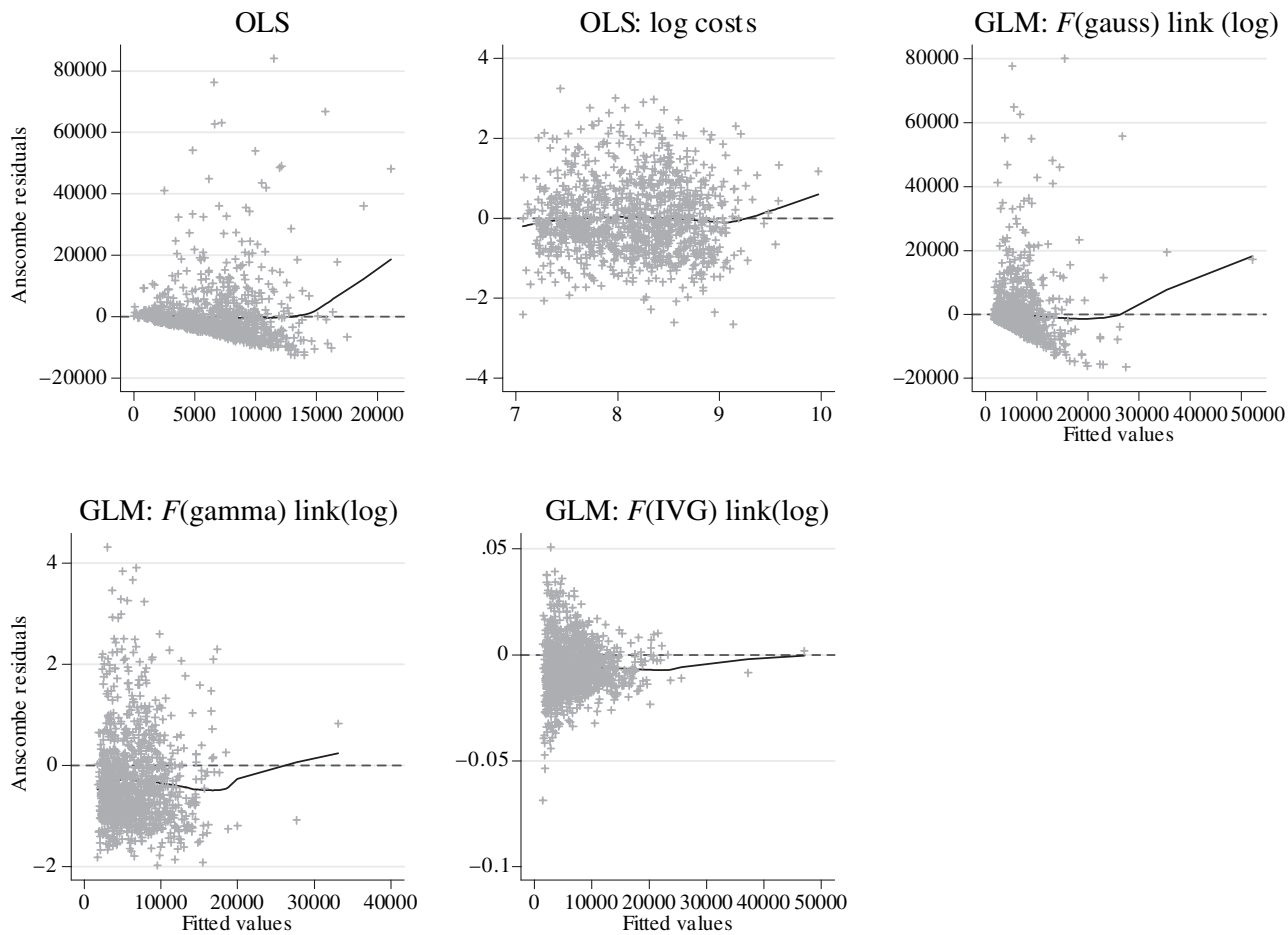

Figure 2 Plots of Anscombe residuals versus fitted values for various models. Plots of Anscombe residuals (vertical axis) against fitted values (horizontal axis) from regression models. Upper panel (left to right): least squares regression (OLS), OLS with log transformed costs, generalized linear model (GLM) Gaussian family and log link. Lower Panel (left to right): GLM gamma family and log link, GLM inverse Gaussian (IVG) family and log link.

except for the normal distribution, where the mean and variance are independent; and the link function determines the appropriate scale [40]. For example, in OLS with a (log) transformation (g), the expectation (E) is $E(g(Y_i)) = \alpha + x\beta$; for the GLM, the form of the expectation is $g(E(Y)) = \alpha + x\beta$. That is, the GLM log-links the predictor ($x\beta$) rather than the response and parameters are equal to the logs of arithmetic means (continuous variables) and their ratios (categorical variables) [41]; thus, parameters can be interpreted directly in a manner similar to odds ratios [34]. GLM are fitted by either maximum likelihood or iteratively re-weighted least squares and a key parameter is the deviance = $2\log\lambda$, where $\log\lambda$ = likelihood (full or 'saturated' model) – likelihood (null or intercept only model). For the normal distribution model, the deviance is the residual sum of squares and hence the notion of R^2 [=1 – (residual sum of squares/total sum of squares)] may be interpreted as the familiar 'per cent variance explained'. Although there are 'pseudo-R^2' statistics for the GLM, the deviance for non-normal distributions is different from the residual sum of squares and the scalar values of these various statistics are not monotone transformations, as would apply to the normal linear model. Thus, the squared correlation (R^2) of models showed modest correlation ($rho = 0.52$, $P = 0.1$) with the Ben-Akiva and Lerman adjusted likelihood ratio index (Table 2), but poor concordance ($rho\_c = 0.07$, $P = 0.15$).

Model performance

Overall, predictive performance was low, as adjudged by R^2, but similar to that of Becker *et al*. [1], who reported $R^2 = 0.13$ for a multivariable regression equation predicting costs after cardiac surgery and also limited the covariate recording period to ≤3 days post-operatively. There was no apparent advantage, in terms of R^2, of a 'full' model (17 covariates, data not shown), although the total patient number would have been sufficient [42]. Covariate selection was not constrained to be constant between models and variation of the model covariate sets occurred, similar to that reported by Dudley *et al*. [43] Formal data trimming was not initially undertaken [44] and there may have been a tendency in the OLS and GLM – inverse Gaussian family and log link models to overfitting, as evidenced by a relatively low RMSE and high MAE [10]. Across both quantitative indices and graphical assessment of model performance, OLS – log costs with heteroscedastic retrans-

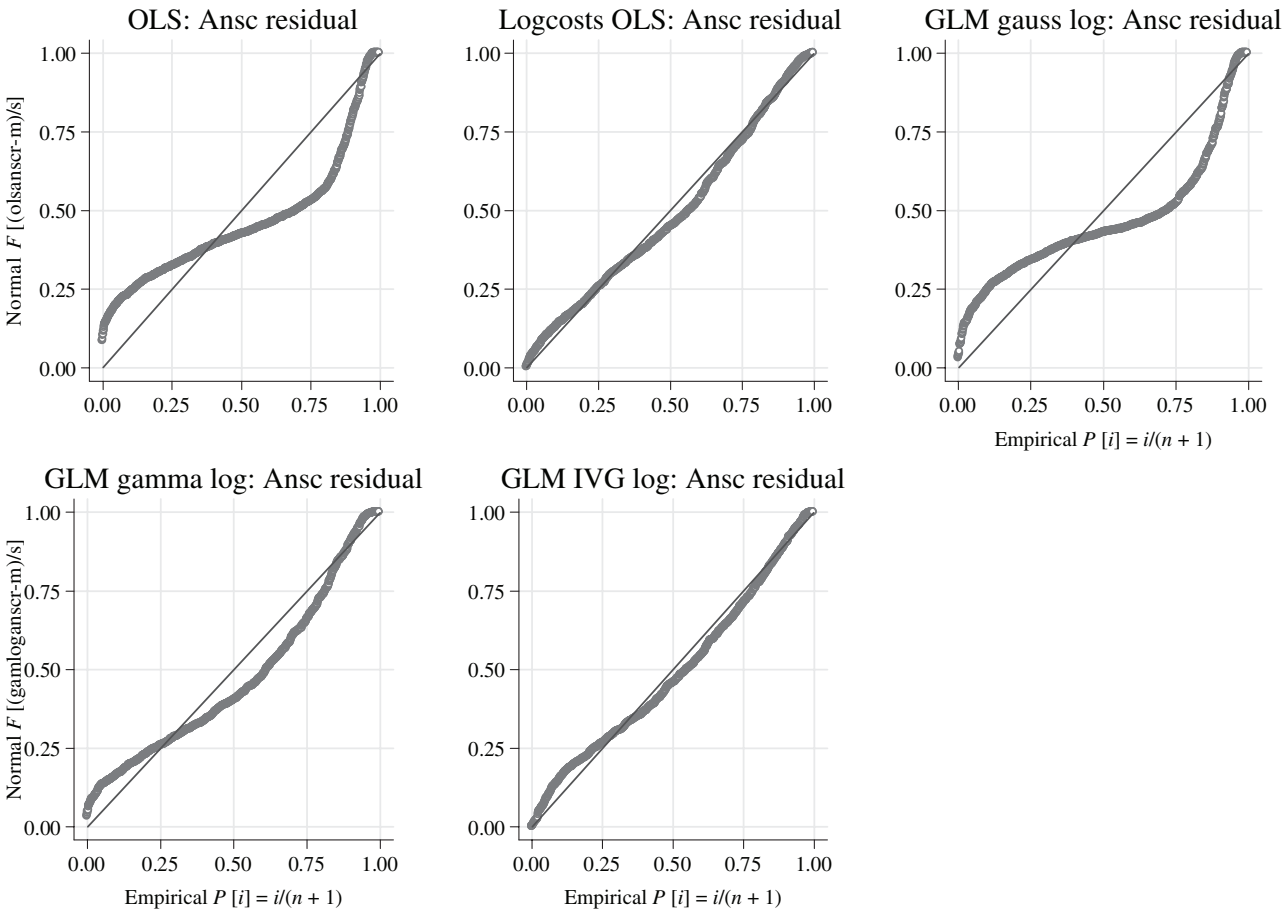

Figure 3 Standardized normal probability plot of Anscombe (Ansc) residuals (P-norm plot). Standardized normal probability plot ('P-norm' plot) of regression models. Vertical axis: cumulative probability related to normal distribution $\Phi\{(x_i - \hat{u})/\hat{\sigma}\}$), where $\hat{u}$ is mean of data and $\hat{\sigma}$ is standard deviation; horizontal axis: plotting positions (Weibull, $p_i = i/(n + 1)$). Upper panel (left to right): least squares regression (OLS), OLS with log transformed costs, generalized linear model (GLM) Gaussian family and log link. Lower Panel (left to right): GLM gamma family and log link, GLM inverse Gaussian (IVG) family and log link.

formation, and GLM – inverse Gaussian family and log link seemed the preferred models. That the GLM model(s) had comparative performance compared with OLS – log costs is of obvious advantage, in that re-transformation is avoided and $E(y|x)$ or $\ln(E(y|x))$ is 'directly' available [10]. In terms of selection between GLMs, an assessment of the power function of the variance mean ($= \mu$) relation has been proposed (9,10), using regression of the log of squared residuals $\log(y_i - \hat{y}_i)^2$, where y_i = observed costs and $\hat{y}_i$ = fitted or predicted values) against the log of the fitted values in the raw scale: $\ln(y_i - \hat{y}_i)^2 = \lambda_0 + \lambda_1 \ln(\hat{y}_i) + v_i$, the scalar quantity of the coefficient (λ_1) of the logged fitted values indicating the degree of this power relationship. For the GLM gamma family, $\lambda = 2$ (that is, variance = μ^2) and for the GLM inverse Gaussian family, $\lambda = 3$ (variance = μ^3). In the current data set, λ was calculated as 2.1, suggesting initial model preference for the gamma distribution; this was also reflected in model AIC values (normalized for n, lower values being preferred), comparing across GLMs (Table 2): 21.04, 19.24 and 26.28 (Gaussian, gamma and inverse Gaussian family GLM respectively).

Heteroscedasticity

The primary concern in this study was the prediction of total costs from ICU admission day data; that is, pragmatic rather than explanatory [45]. Thus, unlike other studies [4,10,19], the effect of, for example, patient groupings (into hospitals) and covariate heteroscedasticity on precision of the β coefficients and the appropriate compensation for this via robust or bootstrapped variance estimates [19], was not a focus of attention, although this would be an issue in assessing the relative importance of various covariates to cost determination. This being said, all models (including the 'full' model, data not shown) demonstrated heteroscedasticity to some degree, with the GLM – gamma family and log link exhibiting least tendency (Fig. 4). A number of factors undoubtedly contributed to this: the skewness of the cost data, patient groupings and the non-normality of the two continuous predictors, APACHE III score and age. Standard transformations and quantile ($n = 4$) categorization of the latter two covariates did not resolve this heteroscedasticity.

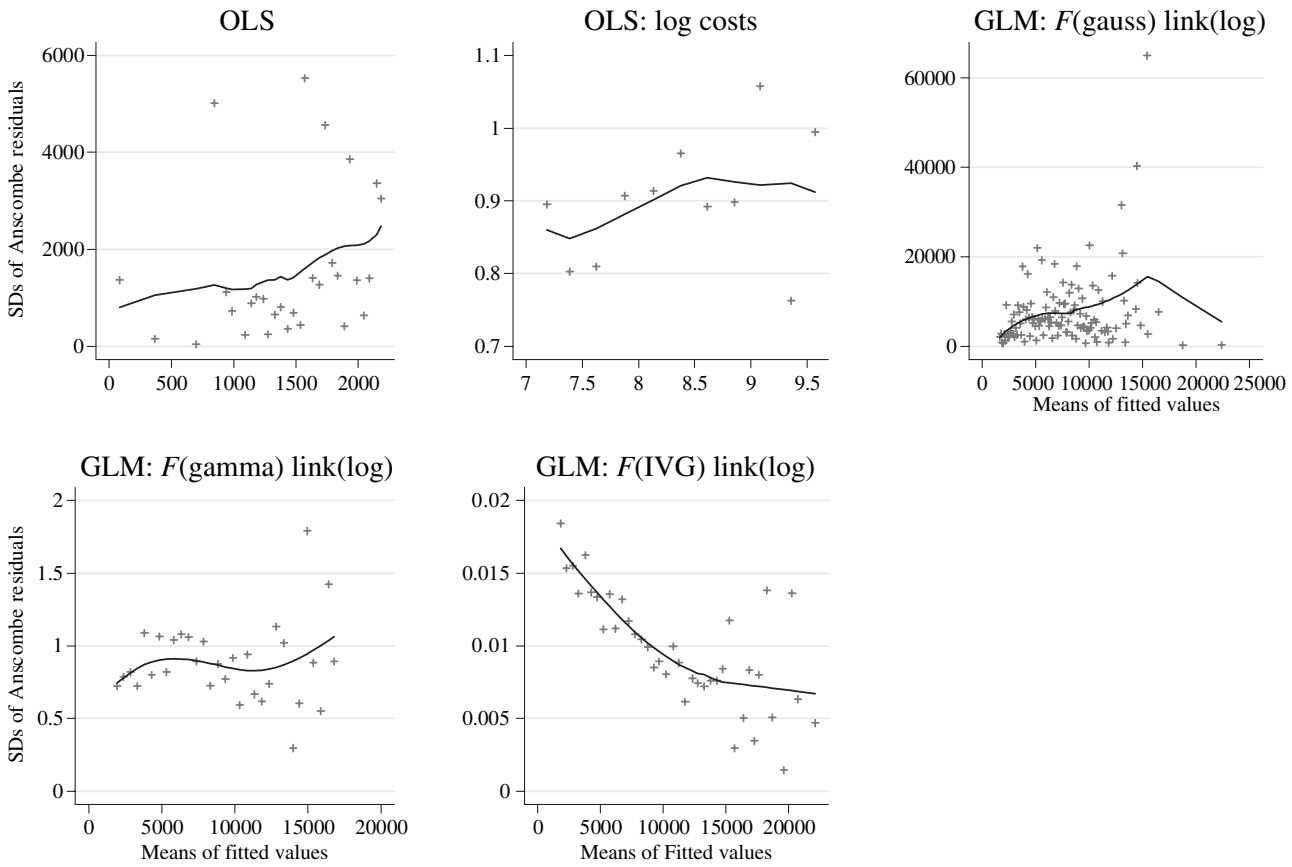

Figure 4 Lowess plots of SDs of Anscombe residuals versus means of grouped fitted values. Heteroscedasticity diagnostic plots using 'Lowess' smoothing [28]. Vertical axis: standard deviation of Anscombe residuals. Horizontal axis: means of grouped fitted values of regression models. Upper panel (left to right): least squares regression (OLS), OLS with log transformed costs, generalized linear model (GLM) Gaussian family and log link. Lower Panel (left to right): GLM gamma family and log link, GLM inverse Gaussian (IVG) family and log link.

Conclusions

Generalized linear models offer an alternative to the standard OLS model for cost prediction. OLS with log transformation of the dependent cost variable must appropriately formulate the problem of back transformation to avoid predictive bias. GLM using the inverse Gaussian response distribution may be of advantage in the analysis of cost data. The relative paucity of studies using GLMs in health cost studies may reflect the known lag-time of transfer of statistical methodology to the medical literature [46].

References

1. Becker, R. B., Zimmerman, J. E., Knaus, W. A., Wagner, D. P., Seneff, M. G., Draper, E. A., Higgins, T. L., Estafanous, F. G. & Loop, F. D. (1995) The use of APACHE III to evaluate ICU length of stay, resource use, and mortality after coronary artery by-pass surgery. *Journal of Cardiovascular Surgery*, 36, 1–11.
2. Diehr, P., Yanez, D., Ash, A., Hornbrook, M. & Lin, D. Y. (1999) Methods for analyzing health care utilization and costs. *Annual Review of Public Health*, 20, 125–144.
3. Sznajder, M., Leleu, G., Buonamico, G., Auvert, B., Aegerter, P., Merliere, Y., Dutheil, M., Guidet, B. & Le Gall, J. R. (1998) Estimation of direct cost and resource allocation in intensive care: correlation with Omega system. *Intensive Care Medicine*, 24, 582–589.
4. Manning, W. G. (1998) The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics*, 17, 283–295.
5. Chhikara, R. S. & J. L. Folks. (1989) The Inverse Gaussian Distribution: Theory, Methodology, and Applications. New York: Marcel Dekker, Inc.
6. Barber, J. A. & Thompson, J. C. (2004) Multiple regression of cost data: use of generalised linear models. *Journal of Health Services Research and Policy*, 9, 197–204.
7. Duan, N. (1983) Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association*, 78, 605–610.
8. Cole, T. J. (2000) Sympercents: symmetric percentage differences on the 100 log (e) scale simplify the presentation of log transformed data. *Statistics in Medicine*, 19, 3109–3125.
9. Blough, D. K., Madden, C. W. & Hornbrook, M. C. (1999) Modelling risk using generalized linear models. *Journal of Health Economics*, 18, 153–171.
10. Manning, W. G. & Mullahy, J. (2001) Estimating log models: to transform or not to transform? *Journal of Health Economics*, 20, 461–494.
11. Myers, R. H. & Montgomery, D. C. (1997) A tutorial on generalized linear models. *Journal of Quality Technology*, 29, 274–291.

12. Hardin, J. & J. Hilbe. (2001) Generalized Linear Models and Extensions. College Station, TX: Stata Press.
13. Moran, J. L., Peisach, A. R., Solomon, P. J. & Martin, J. (2004) Cost calculation and prediction in adult intensive care: a ground-up utilisation study. *Anaesthesia and Intensive Care*, 32, 787–797.
14. Rechner, I. J. & Lipman, J. (2005) The costs of caring for patients in a tertiary referral Australian Intensive Care Unit. *Anaesthesia and Intensive Care*, 33, 477–482.
15. Jacobs, P., Rapoport, J. & Edbrooke, D. (2004) Economies of scale in British intensive care units and combined intensive care/high dependency units. *Intensive Care Medicine*, 30, 660–664.
16. Fetter, R. B., Shin, Y., Freeman, J. L., Averill, R. F. & Thompson, J. D. (1980) Case mix definition by diagnosis-related groups. *Medical Care*, 18, 1–53.
17. Knaus, W. A., Wagner, D. P., Draper, E. A., *et al.* (1991) The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest*, 100, 1619–1636.
18. Gan, F. F., Koehler, K. J. & Thompson, J. C. (1991) Probability plots and distribution curves for assessing the fit of probability models. *American Statistician*, 45, 14–21.
19. Kilian, R., Matschinger, H., Loeffler, W., Roick, C. & Angermeyer, M. C. (2002) A comparison of methods to handle skew distributed cost variables in the analysis of the resource consumption in schizophrenia treatment. *Journal of Mental Health Policy and Economics*, 5, 21–31.
20. Lindsey, J. K. & Jones, B. (1998) Choosing among generalized linear models applied to medical data. *Statistics in Medicine*, 17, 59–68.
21. Royston, P., Ambler, G. & Sauerbrei, W. (1999) The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology*, 28, 964–974.
22. Zheng, B. & Agresti, A. (2000) Summarizing the predictive power of a generalized linear model. *Statistics in Medicine*, 19, 1771–1781.
23. Carpenter, J. & Bithell, J. (2000) Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, 19, 1141–1164.
24. Scott, A. & Wild, C. (1991) Transformations and R^2. *American Statistician*, 45, 127–129.
25. Zheng, B. (2000) Summarizing the goodness of fit of generalized linear models for longitudinal data. *Statistics in Medicine*, 19, 1265–1275.
26. Greene, W. H. (2003) Heteroscedasticity. In Econometric Analysis, 4th edn (ed. W. H. Greene), pp. 215–249. Upper Saddle River, NJ: Prentice Hall, Inc.
27. Dobson, A. J. (2001) An Introduction to Generalized Linear Models, 2nd edn, Boca Raton, CA, Chapman and Hall/CRC.
28. Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829–836.
29. Buchner, D. M. & Findley, T. W. (1990) Research in physical medicine and rehabilitation. VIII. Preliminary data analysis. *American Journal of Physical Medicine and Rehabilitation*, 69, 154–169.
30. Briggs, A. & Gray, A. (1998) The distribution of health care costs and their statistical analysis for economic evaluation. *Journal of Health Services Research and Policy*, 3, 233–245.
31. Coyle, D. (1996) Statistical analysis in pharmacoeconomic studies. A review of current issues and standards. *Pharmacoeconomics*, 9, 506–516.
32. Rascati, K. L., Smith, M. J. & Neilands, T. (2001) Dealing with skewed data: an example using asthma-related costs of medicaid clients. *Clinical Therapeutics*, 23, 481–498.
33. Rutten-van Molken, M. P., van Doorslaer, E. K. & van Vliet, R. C. (1994) Statistical analysis of cost outcomes in a randomized controlled clinical trial. *Health Economics*, 3, 333–345.
34. Blough, D. K. & Ramsey, S. D. (2000) Using generalized linear models to assess medical care costs. *Health Services and Outcomes Research Methodology*, 1, 185–202.
35. Zhou, X. H., Stroupe, K. T. & Tierney, W. M. (2001) Regression analysis of health care charges with heteroscedasticity. *Applied Statistics*, 50, 303–312.
36. Whitmore, G. A. (1975) The inverse Gaussian distribution as a model of hospital stay. *Health Services Research*, 10, 297–302.
37. Austin, P. C., Ghali, W. A. & Tu, J. V. (2003) A comparison of several regression models for analysing cost of CABG surgery. *Statistics in Medicine*, 22, 2799–2815.
38. Nelder, J. A. & Wedderburn, R. W. M. (1972) Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135, 370–384.
39. Breslow, N. E. (1996) Generalized linear models: checking assumptions and strengthening conclusions. *Statistica Applicata*, 8, 23–41.
40. Lane, P. W. (2002) Generalized linear models in soil science. *European Journal of Soil Science*, 53, 241–251.
41. Moran, J. L., Solomon, P., Ay Yeung, K. W., Pannall, P. R. & John, G. & Eliseo, A. (2002) Phosphate metabolism in intensive care patients with acute respiratory failure. *Critical Care and Resuscitation*, 4, 93–103.
42. Green, S. B. (1991) How many subjects does it take to do a regression analysis. *Multivariate Behavioural Research*, 26, 499–510.
43. Dudley, R. A., Harrell, F. E. Jr, Smith, L. R., Mark, D. B., Califf, R. M., Pryor, D. B., Glower, D., Lipscomb, J. & Hlatky, M. (1993) Comparison of analytic models for estimating the effect of clinical factors on the cost of coronary artery bypass graft surgery. *Journal of Clinical Epidemiology*, 46, 261–271.
44. Lee, A. H., Xiao, J., Vemuri, S. R. & Zhao, Y. (1998) A discordancy test approach to identify outliers of length of hospital stay. *Statistics in Medicine*, 17, 2199–2206.
45. Schwartz, D. & Lellouch, J. (1967) Explanatory and pragmatic attitudes in therapeutical trials. *Journal of Chronic Diseases*, 20, 637–648.
46. Altman, D. G. & Goodman, S. N. (1994) Transfer of technology from statistical journals to the biomedical literature. Past trends and future predictions. *Journal of the American Medical Association*, 272, 129–132.

