Published in final edited form as:

Nurs Res. 2010 ; 59(4): 301–307. doi:10.1097/NNR.0b013e3181e507f1.

An Application of Longitudinal Analysis with Skewed Outcomes

Andres Azuero, PhD, MBA [Assistant Professor; Biostatistician],

School of Nursing, University of Alabama at Birmingham, Birmingham, Alabama

Maria Pisu, PhD [Assistant Professor],

School of Medicine, Division of Preventive Medicine, University of Alabama at Birmingham, Birmingham, Alabama

Patrick McNees, PhD, FAAN [Professor; Associate Dean for Research],

School of Nursing, School of Medicine, Division of Preventive Medicine, School of Health Professions, University of Alabama at Birmingham, Birmingham, Alabama

Jeffrey Burkhardt, PhD [Professor],

School of Health Professions University of Alabama at Birmingham Birmingham, Alabama

Rachel Benz, BSN, RN [Research Assistant], and

School of Nursing University of Alabama at Birmingham Birmingham, Alabama

Karen Meneses, PhD, RN, FAAN [Professor; Associate Dean for Research]

School of Nursing University of Alabama at Birmingham Birmingham, Alabama

Abstract

Background—Longitudinal designs are indispensable to the study of change in outcomes over time, and have an important role in health, social, and behavioral sciences. However, these designs present statistical challenges particularly related to accounting for the variance and covariance of the repeated measurements on the same participants, and to modeling outcomes that are not normally distributed.

Objectives—To introduce a general methodology for longitudinal designs to address these statistical challenges and to present an example of an analysis conducted with data collected in a randomized clinical trial. In this example, the outcome of interest—monthly health-related out-of-pocket expenses incurred by breast cancer survivors—had a skewed distribution.

Methods—Common statistical approaches are reviewed for longitudinal analysis using linear and generalized linear mixed models and discussed methods are applied to analyze monthly health-related out-of-pocket expenses.

Discussion—While standard statistical software is available to conduct longitudinal analyses, training is necessary to understand and take advantage of the various options available for model fitting. However, knowledge of the basics of the methodology allows assimilation and incorporation into practice of evidence from the numerous studies that use these designs.

Correspondence: Andres Azuero, PhD, NB 1019F, 1530 3rd Ave. S., Birmingham, AL 35294-1210, Phone: (205) 996-9441, Fax: (205) 996-7177, andreo@uab.edu.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

biostatistics; longitudinal studies; linear mixed models; clinical research design; cancer survivors; out-of-pocket expenses

Longitudinal studies are used to measure an outcome repeatedly over time for the same participants. Such designs can be either experimental or observational, and are indispensable in studying the change in an outcome over time. Because the study of change is fundamental to almost every discipline, there has been steady growth in the number of studies using longitudinal designs (Fitzmaurice, Davidian, Verbeke, & Molenberghs, 2008). Common objectives of such analyses include testing the development over time of mean differences between groups, estimating the average treatment effect over time, determining a suitable model that describes the relationship of an outcome with time and other explanatory covariates, and estimating the correlation patterns among the repeated measurements on the same individuals.

For continuous outcomes (i.e., dependent variables in a model measured on a continuous scale), the most common approaches, such as repeated measures analysis of variance (ANOVA), profile analysis with multivariate ANOVA (MANOVA), and linear mixed models (LMM), require that the outcome variable be approximately normally distributed at each time point, (i.e., a histogram of the outcome at each time point should resemble the bell-shaped form of the normal distribution). Thus, in the presence of an outcome that is not normally distributed (e.g., health care cost data), the validity of statistical inferences is threatened if one of these common approaches is used to analyze such outcome.

The purpose of this presentation is to (a) review common statistical approaches for longitudinal analysis using linear and generalized linear mixed models; and (b) present an example of longitudinal analysis of data from a clinical trial with breast cancer survivors where the outcome of interest, monthly health-related out-of-pocket (OOP) expenses, had a skewed distribution.

Method

Classical Approaches to Longitudinal Models for Continuous Outcomes

The classical approaches to analyze continuous outcomes include repeated measures ANOVA and profile analysis with MANOVA (Tabachnick & Fidell, 2007). For inferences to be valid, both approaches require that the outcome variable be approximately normally distributed at each time point.

Linear Mixed-Effects Models for Longitudinal Data

Because of its flexibility, the LMM is probably the most widely used method for analyzing longitudinal data where the outcome is assumed to be normally distributed. These models are fitted through likelihood-based methods instead of ordinary least squares. Instead of matrix algebra alone, as in ANOVA and MANOVA, solving the equations for a mixed model requires the use of complex computer routines that maximize a *likelihood function* (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006). The likelihood function is the product of the probabilities for each data point under an assumed distribution for the outcome. The larger the likelihood, the better the model fit, under the assumed distribution.

Model fit is assessed using likelihood ratio tests comparing the likelihood function values between full and reduced (i.e., with fewer parameters) models. The name *mixed model* comes from the fact that a model may contain both fixed and random effect parameters. Fixed effects are assumed to be population constants, and random effects are assumed to be random

observations coming from an underlying distribution (often assumed to be normal). Simple examples of fixed effects are the treatment effect of an intervention, a gender effect, or the slope for age as a predictor for an outcome. Simple examples of random effects are classrooms in which students are nested, practices in which patients are nested, and, in the case of longitudinal analyses, individuals in which the repeated observations are nested.

Longitudinal models under the LMM framework include models with random effects, often referred to as subject-specific models, and marginal or population-averaged models that do not include random effects. Both model types allow conducting inferences on the fixed effects and on the variance-covariance structure of the repeated observations on the same participants, but only a model with random effects allows individual-specific prediction.

In this presentation the focus is on the simpler marginal models (without random effects). A common fitting technique for marginal models is the generalized estimating equation (GEE) approach. However, GEE is only one among a number of fitting techniques. Brown and Prescott (2006) provide a comprehensive treatment of both marginal and random effects models with a focus on clinical applications.

A Basic Longitudinal Model

For a marginal longitudinal model comparing two groups (e.g., treatment vs. control), at least three basic fixed-effect parameters need to be fitted, before adding any other covariates of interest to the model. A basic longitudinal model needs parameters for a time effect, a group effect, and a time by group interaction. For a reader familiar with common regression or ANCOVA models, it may look like the following equation:

$$y_{ij} = \beta_0 + \beta_1 \text{time} + \beta_2 \text{group} + \beta_3 \text{time} * \text{group} + e_{ij}$$

where y_{ij} is the outcome observation number i on individual j ; β_0 is the intercept; β_1 , β_2 , and β_3 are the time, group, and time by group interaction parameters, respectively; and e_{ij} is the error term.

The time effect, which can be modeled as either a categorical or a continuous variable, is used to determine whether the group means change significantly over time, but in the same direction. The simplest way of modeling the time effect is by assuming a linear trend for the trajectory of the means over time. Under this assumption, the time variable is continuous, and the coefficient is a slope. This assumption can be inspected visually by plotting the trajectory of the outcome means by group over each of the time points. If the trajectories do not follow approximately straight lines, then a categorical variable for the time effect can be fitted instead. The inclusion of a categorical variable for the time effect in the model allows estimating outcome means separately at each time point, regardless of their trajectory. The downside is that there are more time parameters to estimate ($k-1$, for k time points). The group effect, which is a categorical variable, is used to determine if there is constant separation between the groups over time. The time by group interaction is used to assess whether separation between the groups develops over time, and it is usually, but not necessarily, the parameter of interest. A graphical representation of situations where the time, group, and time by group interaction parameters are used is shown in Figure 1, under the assumption of a linear trend for the trajectory of the means over time.

A simple and often sufficient variance-covariance structure that can be fitted to the basic model is referred to as *compound-symmetric* structure. Assuming that the number of repeated measures i is 3; that the correlation, ρ , among any two measurements is constant; and that the

variance of the measurements, σ^2 , is also constant over time, a compound-symmetric variance-covariance matrix for the three observations on an individual j would look like the following:

$$V_j = \sigma^2 \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

Another common variance-covariance structure used in longitudinal analysis is referred to as an *autoregressive* structure. In this structure, the correlation among repeated measurements decreases with increased separation in time. Assuming now four repeated measurements in time, an autoregressive variance-covariance matrix for the four observations on an individual j would look like the following:

$$V_j = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

In this case ρ is referred to as the autoregressive parameter, and it is equal to the estimated correlation between any two measurements adjacent in time. Note that ρ^2 and ρ^3 would be the correlation among measurements separated by one and two time points, respectively. Since ρ ranges between 0 and 1, $\rho \geq \rho^2 \geq \rho^3$, and hence the correlation decreases as separation in time increases.

There are many more variance-covariance structures available for model fitting in standard software packages such as SAS (SAS Institute Inc., Cary, NC), SPSS (SPSS Inc., Chicago, IL), or R (The R Foundation for Statistical Computing, Wien, Austria). Likelihood-based measurements, such as Akaike's Information Criteria (AIC), can be used to determine what structure is the best fit for a model, given a constant set of fixed effect parameters. A model with the same fixed effects can be fitted with different variance-covariance structures, and the structure that results in the smaller AIC is preferred. However, parsimony has to be considered as well, so a model with the least number of variance-covariance parameters to estimate is preferred. An advantage of both compound symmetric and autoregressive structures is that only two parameters need to be estimated. The main issue of fitting an inadequate variance-covariance structure is that the variability of the measurements is not accounted for appropriately and thus confidence intervals and p-values may be calculated incorrectly.

Normalizing Transformations for Nonnormal Continuous Outcomes

An approach to the issue of having a continuous nonnormal outcome, such as with highly skewed data, is to apply a *normalizing* transformation to the raw data, and conduct the analyses with common techniques for normal data. A problem of this approach is that inferences are made on the means of the transformed data, whereas the interest is in the mean of the raw data. Once the analysis has been conducted, back-transforming to the original scale of the data is not possible. A common normalizing transformation is the natural logarithm. However, the mean of log-transformed observations is different from the logarithm of the mean of the original observations, so once the data are transformed, and the mean calculated, back-transformation to the original scale is not possible. Issues of transforming skewed outcomes, such as interpretability, lack of accuracy, and inefficiency (i.e., larger standard errors) are discussed by Manning (1998) and Manning and Mullahy (2001).

Generalized Linear Mixed-Effects Models for Nonnormal Longitudinal Data

For cross-sectional data that are not normally distributed, the generalized linear model (GdLM) provides a statistical framework that allows building regression-type models to assess relationships between explanatory variables and nonnormal outcomes such as binary responses, counts, or highly skewed continuous data. In GdLM models, it is not the mean outcome that is modeled (as in the LMM), but a function of the mean outcome. This function is referred to as the *link* function.

A typical example of a GdLM is logistic regression, which is often taught as a separate technique. If the outcome data are binary (e.g., presence or absence of disease, where an event is coded as 1 and a nonevent is coded as 0), the mean is simply the proportion of events p , which would lie between 1 and 0. The link function in this case is the *logit* transformation,

which is the natural logarithm of the odds: $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$. For count data (e.g., the number of days a patient is hospitalized), a common assumption is that the counts follow a Poisson distribution. For Poisson regression, the link function is the natural logarithm, so that the logarithm of the mean response is modeled. Each of the distributions that can be used for GdLM has a natural link function. The binomial, Poisson, beta, gamma, multinomial, and negative binomial distributions are examples of distributions available in standard software for GdLM fitting. As with the LMM, GdLM models are fitted using maximum likelihood methods, but assuming underlying distributions other than the normal.

The generalized linear mixed model (GLMM) is an extension of the GdLM to allow modeling longitudinal nonnormal outcomes. A basic longitudinal model fitted with a GLMM would look like the following equation:

$$\text{Link}(\text{mean}) = \beta_0 + \beta_1 \text{time} + \beta_2 \text{group} + \beta_3 \text{time} * \text{group}$$

The same discussion as in the LMM section on the basic parameters applies. In these models, however, the covariance structure is estimated in terms of the link scale.

Because of their complexity, the mathematical and statistical details for the GLMM are beyond the scope of this article. Instead, an example is presented in the following section. Readers interested in the statistical details can consult the books by Brown and Prescott (2006); McCulloch, Searle, and Neuhaus (2008); Littell et al. (2006); the documentation for the GLIMMIX procedure in SAS software; and the documentation for the GENLIN and GEE algorithms in SPSS software. As with the LMM, random-effects and marginal GLMM models can be constructed, and model fit is assessed through likelihood-based statistics. However, the fitting algorithms are even more complex than those for the LMM. The following example is focused on the simpler marginal models.

Application

The Breast Cancer Education Intervention (BCEI) was a 6-month randomized trial to examine the effect of a nurse-led psychoeducational support intervention on perceived quality of life in early-stage breast cancer survivors. The BCEI research findings of the main outcome variable, perceived quality of life, were reported elsewhere by Meneses et al. (2007). Briefly, 256 women participated in the BCEI, of whom 129 were randomized to the intervention group and 132 to the wait control. Data were collected at 3 time points: baseline, 3 months, and 6 months. Research protocols received institutional review board approval.

The focus of this analysis is on the 132 participants in the control group to examine a secondary outcome, monthly health-related OOP expenses incurred by these study participants. The OOP expenses were estimated by asking participants for their directly incurred dollar expenses related to cancer care, treatment side effects management, counseling, and health maintenance. Monthly estimates were calculated for each time point. Of the 132 participants in the control group, 121 provided information on annual household income. Two income level groups were based on household incomes below ($n = 56$) or above ($n = 65$) \$50,000. Demographic and treatment characteristics of participants based on income level group are shown in Table 1. Descriptive statistics for monthly OOP expenses by income level group and time point are shown in Table 2.

The objectives of this analysis were to (a) assess whether there were differences in mean monthly OOP between participants having household incomes below or above \$50,000; (b) determine whether separation over time in mean monthly OOP expenses developed between these two income level groups during the 6-month observation period; and (c) determine if other covariates of interest (age, education, marital status) were associated with OOP expenses. The first step was to inspect the distribution of monthly OOP expenses at each time point, and the trajectory of the mean monthly OOP expenses over time. The histograms in Figure 2 revealed highly right-skewed distributions at each time point, so the assumption of OOP expenses following a normal distribution had to be discarded. From examining Table 2, it was determined that the trajectory of the means over time was not linear, so a time variable fitted as categorical would be more appropriate than fitted as continuous.

The Gamma distribution has been used by health economists to model skewed outcomes such as health cost data (Manning, Basu, & Mullahy, 2005; Nixon & Thompson, 2004). Different shapes of the Gamma distribution are shown in Figure 3. A link function that can be used for Gamma models is the natural logarithm. Thus, assuming a Gamma distribution for monthly OOP expenses, the natural logarithm of the mean monthly OOP expenses can be modeled using the marginal model shown in Table 3 (model 1). Because there were only three time points, the covariance structure can be set to a simple compound symmetric structure. With SAS software, in the Proc GLIMMIX programming statements used for running this model, the explanatory variables were defined as categorical or class variables. Then, the statistical package was used to automatically select the reference categories, producing an overall test of effect for each variable (Type III test of fixed effect), with an F statistic. The other longitudinal models considered were models with age, education, and marital status as explanatory variables of interest (models 2-4). Test results from these models are shown in Table 4.

The results for model 1 indicated that (a) there was no significant difference in mean monthly OOP expenses between the two household income groups over time; (b) mean monthly OOP expenses did not significantly change during the 6-month period of observation for the two groups combined; and (c) separation in mean OOP expenses between these two groups did not develop over the 6-month period of observation. However, the results from models 2 and 3 suggested that there were constant differences over time in mean OOP expenses by education level and age. Model 5 was used to calculate time-averaged effects of education and age. In Model 6, the normal instead of the Gamma distribution is used. In this case, education would not remain as a significant predictor of mean monthly OOP expenses. Model-calculated estimates of monthly OOP expenses adjusted for age (calculated with models 5 and 6 for comparison) are shown in Table 5. Using the Gamma model, back-transforming the mean outcome from the link scale was straightforward using exponentiation, the inverse function of the natural logarithm. With the exception of the graduate school category, all confidence intervals for age-adjusted mean OOP expenses by education category were wider with the normal model than with the Gamma model.

Conclusions

The most common methods for analyzing longitudinal data in the biomedical sciences are LMM and GLMM. The methodologies, albeit complex, are very powerful and provide several advantages compared to the classical approaches. For highly skewed data, an alternative approach would be to use the median as measure of centrality instead of the mean. Although standard software is readily available to conduct median or quantile regression in cross-sectional data, the approach is not currently available for longitudinal data. In fact, median regression for longitudinal data is a current topic of research in statistical science. Additionally, in the case of health costs, health economists prefer the mean as measure of centrality because with knowledge of the mean cost and the sample size, it is possible to estimate the total costs incurred by the cohort. This estimation is not possible with the median.

While researchers do not need to know all the intricacies of LMM and GLMM, it is important to be familiar with the approaches, as more studies are published using the methodology. Because of considerable time and effort required to learn how to implement such techniques, inclusion of a biostatistician to write the methods sections in research proposals and to conduct analyses on collected data is highly recommended.

Acknowledgments

This research study was supported by a grant from the National Institute of Nursing Research and the Office of Cancer Survivorship at the National Cancer Institute (5R01-NR005332-04), USA.

References

- Brown, H.; Prescott, R. Applied mixed models in medicine. John Wiley & Sons; Chichester, UK: 2006.
- Fitzmaurice, G.; Davidian, M.; Verbeke, G.; Molenberghs, G. Longitudinal data analysis. Chapman & Hall/CRC; Boca Raton, FL: 2008.
- Littell, R.; Milliken, G.; Stroup, W.; Wolfinger, R.; Schabenberger, O. SAS for mixed models. 2nd ed.. SAS Institute, Inc.; Cary, NC: 2006.
- Manning WG. The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics* 1998;17(3):283–295. [PubMed: 10180919]
- Manning WG, Mullahy J. Estimating log models: To transform or not to transform? *Journal of Health Economics* 2001;20(4):461–494. [PubMed: 11469231]
- Manning WG, Basu A, Mullahy J. Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics* 2005;24(3):465–488. [PubMed: 15811539]
- McCulloch, C.; Searle, S.; Neuhaus, J. Generalized, linear, and mixed models. 2nd ed.. John Wiley & Sons, Inc.; Hoboken, NJ: 2008.
- Meneses KD, McNeese P, Loerzel VW, Su X, Zhang Y, Hassey LA. Transition from treatment to survivorship: Effects of a psychoeducational intervention on quality of life in breast cancer survivors. *Oncology Nursing Forum* 2007;34(5):1007–1016. [PubMed: 17878129]
- Nixon R, Thompson G. Parametric modeling of cost data in medical studies. *Statistics in Medicine* 2004;23(8):1311–1331. [PubMed: 15083485]
- Tabachnick, B.; Fidell, S. Using multivariate statistics. 5th ed.. Pearson Education Inc.; Boston, MA: 2007.

Figure 1.

Illustration of the use of the three basic parameters: time, group, and time by group, in a longitudinal model comparing two groups over time. (A) The two groups begin at similar levels, but separation occurs over time. Significance is expected in the *time by group* interaction parameter alone. (B) Both groups change over time, but there is no separation between them. Significance is expected in the *time* parameter alone. (C) The groups do not change for the duration of the study, but separation is constant over time. Significance is expected in the *group* parameter alone. (D) Separation is present at the beginning of the study and remains over time. Group 1 changes over time, while Group 2 remains constant. Significance is expected for the *group* and *time by group* parameters.

Figure 2. Histograms of monthly health-related out of pocket expenses in dollars by time point (n = 121)

Note the highly right-skewed distribution of the data at each time point

Figure 3. Assorted shapes of the Gamma distribution

The form of the distribution's curve depends on *shape* and *scale* parameters. The mean and variance are also functions of these two parameters. In a Generalized Linear Model using the Gamma distribution, the shape of the curve would be estimated through maximum likelihood methods.

Table 1
Demographic and Treatment Characteristics According to Household Income Level ($n = 121$)

Characteristic	Household Income		<i>p</i>
	< \$50,000 (<i>n</i> = 56)	≥ \$50,000 (<i>n</i> = 65)	
Demographics			
Age, in years, Mean (<i>SD</i>)	57.5 (12.7)	50.4 (10.5)	< .01 <i>a</i>
Rural resident, <i>n</i> (%)	9 (16.1)	16 (24.6)	.24 <i>b</i>
Race, <i>n</i> (%)			.49 <i>c</i>
African American	7 (12.5)	5 (7.7)	
Caucasian	44 (78.6)	55 (84.6)	
Hispanic or Latina	4 (7.1)	2 (3.1)	
Other	1 (1.8)	3 (4.6)	
Education, <i>n</i> (%)			< .01 <i>c</i>
High School or less	23 (41.1)	9 (13.9)	
Trade school	5 (8.9)	2 (3.1)	
College	21 (37.5)	40 (61.5)	
Graduate school	7 (12.5)	14 (21.5)	
Marital status, <i>n</i> (%)			< .01 <i>c</i>
Never married	7 (12.5)	1 (1.5)	
Married	21 (37.5)	50 (76.9)	
Living with partner	4 (7.1)	2 (3.1)	
Divorced or widowed	24 (42.9)	12 (18.5)	
Employment status, <i>n</i> (%)			.61 <i>c</i>
Full-time	29 (51.8)	38 (58.5)	
Part-time	5 (8.9)	4 (6.2)	
Retired	17 (30.3)	16 (24.6)	
Student	1 (1.8)	1 (1.5)	
Homemaker	2 (3.6)	5 (7.7)	
Unemployed	2 (3.6)	0 (0.0)	
On disability	0 (0.0)	1 (1.5)	
Treatment			
Months since diagnosis, Mean (<i>SD</i>)	8.5 (3.6)	8.7 (2.7)	.75 <i>a</i>
Cancer stage, <i>n</i> (%)			.32 <i>b</i>
I	28 (50.9)	39 (60.0)	
II	27 (49.1)	26 (40.0)	
Surgery, <i>n</i> (%)			.39 <i>b</i>
Lumpectomy	30 (53.6)	38 (58.5)	
Mastectomy	22 (39.3)	19 (29.2)	
Bilateral mastectomy	4 (7.1)	8 (12.3)	

Characteristic	Household Income		<i>p</i>
	< \$50,000 (<i>n</i> = 56)	≥ \$50,000 (<i>n</i> = 65)	
Radiation therapy, <i>n</i> (%)	36 (64.3)	41 (63.1)	.81 ^{<i>b</i>}
Chemotherapy, <i>n</i> (%)	26 (46.4)	41 (63.1)	.07 ^{<i>b</i>}
Hormonal therapy, <i>n</i> (%)	43 (76.8)	51 (78.5)	.82 ^{<i>b</i>}

Notes.

^{*a*} t-test

^{*b*} Chi-square test

^{*c*} Fisher's exact test

Table 2
Descriptive Statistics for Monthly Health-Related Out-of-Pocket Expenses in Dollars by Time Point and Household Income Level (n = 121)

Time	n ^a	Household Income					
		<\$50,000 (n = 56)			≥\$50,000 (n = 65)		
		Mean	Median	SD	Mean	Median	SD
Diagnosis to baseline	53	270.2	102.2	605.3	306.2	255.6	244.2
Baseline to month 3	50	271.5	130.8	439.9	355.8	125.0	642.3
Month 3 to month 6	50	240.9	94.7	458.5	214.5	118.3	323.4

Notes.

^aNumber that reported expenses

Table 3
Longitudinal Model for Health-Related Out-of-Pocket Expenses Using Time and Household Income Level as Explanatory Variables (Model 1)

Explanatory variable	Coefficient	Comments
Intercept	β_0	
Time (categorical variable with month 6 as the reference)	β_1	Indicator for baseline
	β_2	Indicator for month 3
Income (categorical variable with income $\geq \$50,000$ as reference)	β_3	Indicator for income $> \$50,000$
	β_4	Indicator for month 0*Income $< \$50,000$
Time*Income interaction	β_5	Indicator for month 3*Income $< \$50,000$

Table 4
Results for Longitudinal Models Fitted. The Dependent Variable is Monthly Health-Related Out-of-Pocket Expenses, Measured at Three Time Points: Diagnosis to Baseline, Baseline to Month 3, Month 3 to Month 6 (n = 121)

Model Number	Variables in model	Type III Tests of Fixed Effects			
		Numerator DF	Denominator DF	F	p
1	Household Income	1	116.6	0.28	0.5997
	Month	2	223.8	1.22	0.2983
	Household Income*month	2	223.8	0.45	0.6357
2	Age	1	108.1	11.08	0.0012
	Month	2	215.1	0.29	0.7474
	Age*Month	2	216.1	0.1	0.9093
	Education	3	109.2	5.29	0.0019
3	Month	2	213.6	0.41	0.6672
	Education*Month	6	213.7	1.72	0.1184
	Marital Status	3	113.2	1.64	0.1849
4	Month	2	219.7	0.59	0.5527
	Marital Status*Month	6	217.7	0.61	0.7192
5	Age	1	97.16	9.09	0.0033
	Education	3	98.06	3.65	0.0152
	Age	1	107	4.32	0.0401
6	Education	3	108	2.51	0.0623

Notes. Models 1 to 5 were fitted using a generalized linear mixed model with the Gamma distribution and natural logarithm link function. Model 6 was fitted using a linear mixed model, which assumes normally distributed data.

Table 5
Time-Averaged and Age-Adjusted Estimates of Mean Monthly Health-Related Out-of-Pocket Expenses by Education Level (n = 121)

Gamma model (model 5)			Time-averaged age-adjusted out of pocket expenses means by education level					
Variable	Coefficient	Estimate	Using Gamma model (model 5)		Using normal distribution (model 6)			
Intercept	β_0	7.3546	Equation to calculate estimate		Estimate (\$)		95% CI	
Age	β_1	-0.02469						
Education								
≤High School	β_2	-0.972	$\exp(\beta_0 + \beta_1 \cdot age + \beta_2)$		158.1	108.6	230.2	178.0
Trade school	β_3	-0.3279	$\exp(\beta_0 + \beta_1 \cdot age + \beta_3)$		301.0	137.9	657.2	303.2
College	β_4	-0.4584	$\exp(\beta_0 + \beta_1 \cdot age + \beta_4)$		264.2	202.2	345.2	274.5
Graduate school	(Reference)		$\exp(\beta_0 + \beta_1 \cdot age)$		417.9	264.8	659.4	439.4
								584.3

Notes: average age (age) is equal to 53.4 years. In the Gamma model, the intra-subject correlation estimate for out of pocket expenses was equal to 0.21, and out of pocket expenses variance estimate for each age category was equal to the dispersion (estimated at 2.2) multiplied by the square of the respective mean (in the Gamma model, a difference in means implies a difference in variances, as the variance is a mathematical function of the mean).