Modeling Semicontinuous Longitudinal Expenditures: A Practical Guide

Valerie A. Smith iD *, Matthew L. Maciejewski* iD *, and Maren K. Olsen*

Objective. To compare different strategies for analyzing longitudinal expenditure data that have a point mass at $0. We provide guidance on parameter interpretation, research questions, and model selection.

Data Sources, Study Design, and Data Collection. One-part models, uncorrelated two-part models, correlated conditional two-part (CTP) models, and correlated marginalized two-part (MTP) models have been proposed for longitudinal expenditures that often exhibit a large proportion of zeros and a distribution of continuous, highly right-skewed positive values. Guidance on implementing and interpreting each of these model is illustrated with an example of longitudinal (2000–2003) specialty care expenditures of veterans with hypertension, drawn from Veterans Administration data.

Principal Findings. The four strategies answer different research questions, are appropriate for different structures of data, and provide different results. If there is a point mass at $0, then the MTP model may be most useful if the primary interest is in mean expenditures of the entire population. A CTP model may be most useful if the primary interest is in the level of expenditures conditional on them being incurred.

Conclusions. Researchers should consider which modeling strategy for longitudinal expenditure outcomes is both consistent with research aims and appropriate for the data at hand.

Key Words. Semicontinuous data, two-part models, zero-inflated, expenditures, costs

Health care expenditures are often characterized as semicontinuous, consisting of two components: a portion with zero expenditures and a portion with continuously distributed, right-skewed positive values. These two components have led to considerable research examining strategies for modeling cross-sectional expenditures (Duan et al. 1983; Diehr et al. 1999; Madden et al. 2000; Buntin and Zaslavsky 2004). One approach may be to use

standard "one-part" (single-component) regression models, without treatment for the zero-valued component. An alternative is a "two-part" model, which explicitly models both the zero and positive-valued components.

In a simulation study, Manning and Mullahy (2001) examined a range of outcome distributions with one-part models and suggested fitting a series of parametric distributions and conducting a modified Park test to assess fit. Motivated by an analysis of Medicare expenditures, Buntin and Zaslavsky (2004) compared modeling approaches with respect to calibration of predictions and several error metrics, providing the analyst with suggested descriptive analyses to assess model performance. Basu and Manning (2009) summarized these and other current estimation strategies and affirmed that there is not one universally accepted approach for estimating cross-sectional expenditures.

All issues related to cross-sectional expenditure estimation are relevant to longitudinal expenditures. There are additional considerations with longitudinal expenditures, and less research compares strategies for modeling longitudinal expenditures. As with any longitudinal outcome, the estimation must incorporate the correlation of repeated measurements (Basu and Manning 2009). Furthermore, the distribution of longitudinal expenditures and the proportion of zeros are dependent upon the timeframe under consideration (e.g., person-month vs. person-year). In some situations, longer time periods allow more opportunity for individuals to have accrued positive expenditures; therefore, the proportion of expenditures observed in any given time interval may rely upon the length of the interval.

In this study, we compare popular strategies for analyzing semicontinuous longitudinal expenditures. Our focus is on methods where the $0 are true values, unlike, for example, the double-hurdle model (Deaton and Irish 1984). We focus on implications of analytic decisions on what research aims can be answered, ease of implementation, and appropriateness of use. We discuss the

---

Address correspondence to Valerie A. Smith, Dr.P.H., Center for Health Services Research in Primary Care, Durham Veterans Affairs Medical Center, 411 W Chapel Hill St., Suite 600, Durham, NC 27701, and Department of Population Health Sciences, School of Medicine, Duke University Medical Center, Durham, NC; e-mail: valerie.a.smith@duke.edu. Matthew L. Maciejewski, Ph.D., is with the Center for Health Services Research in Primary Care, Durham Veterans Affairs Medical Center, Durham, NC, and the Department of Population Health Sciences, School of Medicine, Duke University Medical Center, Durham, NC. Maren K. Olsen, Ph.D., is with the Center for Health Services Research in Primary Care, Durham Veterans Affairs Medical Center, Durham, NC, and the Department of Biostatistics and Bioinformatics, School of Medicine, Duke University Medical Center, Durham, NC.

rationale and details of each model's specification, software available, model parameter interpretation, and trade-offs. Each strategy is demonstrated with longitudinal specialty expenditures from a cohort of veterans with hypertension.

The remainder of the study is laid out as follows. Section 2 summarizes the motivating example, while Section 3 describes each modeling strategy, explaining implementation details, interpretation, and benefits and draw-backs. Section 4 illustrates each approach applied to the motivating example, and Section 5 concludes with a comparison of approaches and points to areas for future research. Code implementing each model is provided in Appendix SA2.

ILLUSTRATIVE EXAMPLE: SPECIALTY CARE EXPENDITURES

In December 2001, the Veterans Health Administration (VA) increased specialty visit copayments from \$15 to \$50, creating a natural experiment to examine changes in expenditures. The original study (Maciejewski et al. 2012) examined whether increasing specialty visit copayments impacted the likelihood of seeing a specialist or the level of specialty expenditures among veterans who did. Annual inflation-adjusted VA specialty expenditures were constructed for each patient in each year (2000–2003), 2 years prior to, and 2 years after the copayment change.

Veterans were exempt from copayments if they were low income or had disability from their military service to merit a need-based exemption. Remaining veterans were required to pay copayments. To reduce nonequivalence of the groups, the original study conducted one-to-one propensity score matching, generating a sample of 1,693 veterans exempt from copayments and 1,693 veterans required to pay.

The yearly percentage of patients with zero expenditures ranged from 23 percent to 29 percent over the study period, yielding potential need to account for the zero-valued expenditures in the analysis (Figure S1 in Appendix SA2). The explanatory variables of interest included (1) an indicator of whether a veteran had copayments required; (2) year indicator effects for 2001, 2002, and 2003; and (3) an interaction of copayment status and each time effect.

MODELING STRATEGIES FOR LONGITUDINAL SEMICONTINUOUS EXPENDITURES

This section describes four modeling strategies, including generalized linear models (GLMs) with empirical "sandwich" variance estimation via generalized estimating equations (GEEs) (i.e., a one-part model via GEE); uncorrelated, two-part GLMs with empirical variance estimation via GEE; correlated, conditional two-part mixed-effects models; and correlated, marginalized two-part mixed-effects models. Conditional versus marginal can sometimes refer to subject-specific versus population-average estimates; we reserve conditional and marginal to refer to whether estimates are conditional upon incurring a positive outcome or refer to the entire population.

Model Descriptions

One-Part Generalized Linear Model via Generalized Estimating Equations. A one-part GLM fit via GEEs treats the observed expenditures as realizations of a single process, so the model does not differentiate between zero and positive-valued expenditures. The GLM utilizes a link function to accommodate non-normally distributed data, eliminating need to transform the data prior to modeling:

$$g(E(Y_{ij})) = \beta_0 + \beta_1 x_{1ij} + \ldots + \beta_p x_{pij}$$

where Y_{ij} is the semicontinuous outcome for individual i at time j, $g(\cdot)$ is the link function relating the outcome to the linear predictor, and $x_{1ij}, \ldots, x_{pij}$ are the covariates for individual i at time j. Commonly, a log link, where $g(E(Y_{ij})) = \log(E(Y_{ij}))$, is used, but other link functions can be used, such as a square root or identity transformation.

Similarly to GLMs fit with quasi-likelihood for cross-sectional expenditures, GLMs fit via GEEs do not require specification of a parametric distribution (Liang and Zeger 1986; Diggle et al. 2002). Rather, one specifies only the mean and variance. Often, the variance is given as a mean–variance relationship (e.g., variance proportional to the mean, $\text{Var}(E(Y_{ij})) = \rho E(Y_{ij})$, where ρ represents a proportionality constant), and a link function, as described above, provides the form of the mean model. When modeling longitudinal expenditures using GEEs, the working correlation structure of expenditures also needs to be specified.

In addition to a substantial proportion of zeroes, expenditure data often exhibit skewness, heteroscedasticity, and heavy tails (i.e., unusually large

expenditures compared to the mean), so one must consider the specification of the mean–variance relationship. When coupled with empirical "sandwich" standard error estimates, GEE methodology is asymptotically robust to such misspecification. The sample size required to achieve robustness in the empirical standard errors may be quite large, particularly when data are very skewed or contain a significant number of zeros (Smith et al. 2017). If used with small sample sizes, empirical standard errors may underestimate the true standard errors. Additional work has extended GEEs to incorporate a broader set of mean–variance relationships via Extended Estimating Equations (Basu and Rathouz 2005) or partially linear mean function and semiparametric covariance structure (Chen, Liu, and Lee 2016). For more information on determining and testing appropriate model specification and fit in the context of one-part GLMs, see Manning and Mullahy (2001) and Buntin and Zaslavsky (2004).

Modeling longitudinal expenditures with one-part GLMs fit via GEEs has several advantages. Standard one-part GLMs fit with GEEs are easily implemented in most statistical software packages, including SAS's PROC GENMOD using a REPEATED statement and PROC GEE (SAS Institute, Cary, NC), Stata's xtgee (StataCorp, College Station, TX), and R's gee or geepack (www.r-project.org). Additionally, one-part GLMs use a single-component model with a link function, modeling the mean on the original scale, so one can use these models to estimate population-average expenditures and corresponding confidence intervals in the original (e.g., dollar) scale. When a log link is used, the multiplicative effect of a covariate on the population-average mean is easily obtained on the original scale by exponentiating the corresponding parameter(s). For example, $\exp(\beta_k)$ represents the multiplicative increase on $E(Y_{ij})$ from a one-unit increase in x_{kij}. Using one-part GLMs for expenditures gives analysts the ability to provide easily interpretable answers to investigators (Table 1).

Because one-part GLMs focus inference on the overall marginal mean, predictions of mean expenditures at any combination of included covariates are easily estimated from the model. If interest lies in the additive difference in means, however, one must use the method of "recycled predictions," also known as "standardization" (Hernan and Robins 2018). Here, predicted values are obtained for the sample as though each individual was in the treatment group and again as though each was in the control group, and the mean difference of these predictions is obtained. Computation of standard errors or confidence intervals for the additive difference in means can be obtained via bootstrapping, using the method of recycled predictions on each bootstrapped

Table 1: Research Questions, Model Specification, and Parameter Interpretation of Common Models for Longitudinal Semicontinuous Expenditures

Example of Research Questions	Model Specification	Parameter Interpretation
One-part GLM What is the effect of being required to pay a copayment on overall mean VA specialty care expenditures in each year? What are the estimated or predicted overall mean expenditures for those with and without a copayment requirement in each year?	Eq 1.1. $g(v_{ij}) = \beta_0 + \beta_1\,YR2001_j + \beta_2\,YR2002_j + \beta_3\,YR2003_j$ $+ \beta_4 MUSTPAY_i + \beta_5 MUSTPAY_i \times YR2001_j$ $+ \beta_6 MUSTPAY_i \times YR2002_j + \beta_7 MUSTPAY_i \times YR2003_j$ where $g(\cdot)$ is a link function relating the overall mean, v_{ij} for individual i at time j to the linear predictor	Population-average effects on the overall mean
Uncorrelated two-part GLM What is the effect of being required to pay a copayment at each time point on the probability of incurring VA specialty care expenditures? What is the effect of being required to pay a copayment at each time point on the level of VA specialty care expenditures given that some level of specialty expenditure is incurred? What is the estimated probability of incurring VA specialty care expenditures for those with and without a copayment requirement in each year?	Eq. 2.1. $\mathrm{logit}\{\mathrm{Pr}(\mathrm{Cost}_{ij} > 0)\} = \alpha_0$ $+ \alpha_1\,YR2001_j + \alpha_2\,YR2002_j + \alpha_3\,YR2003_j$ $+ \alpha_4 MUSTPAY_i + \alpha_5 MUSTPAY_i \times YR2001_j$ $+ \alpha_6 MUSTPAY_i \times YR2002_j + \alpha_7 MUSTPAY_i \times YR2003_j$ Eq 2.2. $g(\theta_{ij}) = \gamma_0 + \gamma_1\,YR2001_j + \gamma_2\,YR2002_j + \gamma_3\,YR2003_j$ $+ \gamma_4 MUSTPAY_i + \gamma_5 MUSTPAY_i \times YR2001_j$ $+ \gamma_6 MUSTPAY_i \times YR2002_j + \gamma_7 MUSTPAY_i \times YR2003_j$, where $g(\cdot)$ is a link function relating the conditional mean, θ_{ij} for individual i at time j to the linear predictor	Population-average estimates of the probability of incurring expenditures (binary component) and population-average effects on the conditional mean of expenditures among those with positive expenditures (continuous component)

Table 1. *Continued*

Example of Research Questions	*Model Specification*	*Parameter Interpretation*
What are the estimated or predicted conditional mean expenditures among those who incur expenditures for those with and without a copayment requirement in each year? Correlated conditional two-part model In each year, what is the probability of incurring VA specialty care expenditures for an individual required to pay a copayment as compared to an individual not required to pay a copayment? Conditional upon having a specialty visit, is there a difference in specialty log expenditures in each year for an individual required to pay a copayment as compared to an individual not required to pay?	Eq 3.1. $\text{logit}\{\Pr(\text{Cost}_{ij} > 0)\} = \alpha_0 + \alpha_1\,YR2001_j + \alpha_2\,YR2002_j$ $+ \alpha_3\,YR2003_j + \alpha_4 MUSTPAY_i + \alpha_5 MUSTPAY_i \times YR2001_j$ $+ \alpha_6 MUSTPAY_i \times YR2002_j + \alpha_7 MUSTPAY_i \times YR2003_j$ $+ b_{1i}$ Eq 3.2. $\mu_{ij} = E[\ln(\text{Cost}_{ij}|\text{Cost}_{ij} > 0)] = \delta_0$ $+ \delta_1\,YR2001_j + \delta_2\,YR2002_j + \delta_3\,YR2003_j$ $+ \delta_4 MUSTPAY_i + \delta_5 MUSTPAY_i \times YR2001_j$ $+ \delta_6 MUSTPAY_i \times YR2002_j + \delta_7 MUSTPAY_i' \times YR2003_j$ $+ W_i b_{2i},$ where μ_{ij} is the conditional mean of the log of the positive values for individual i at time j, and W_i is the $n_i \times 2$ matrix of **1** and **YEAR**, representing a random intercept and slope	Subject-specific estimates of the probability of incurring expenditures (binary component) and the conditional mean of log expenditures among those with positive expenditures (continuous component)
Correlated marginalized two-part models What is the effect of being required to pay a copayment at each time point on the probability of incurring VA specialty care expenditures?	Eq 4.1 $\text{logit}\{\Pr(\text{Cost}_{ij} > 0)\} = \alpha_0 + \alpha_1\,YR2001_j$ $+ \alpha_2\,YR2002_j + \alpha_3\,YR2003_j$ $+ \alpha_4 MUSTPAY_i + \alpha_5 MUSTPAY_i \times YR2001_j$	Parameters in the first part of the model take a subject-specific interpretation. Most parameters in the second part of the model take

Continued

Table 1. *Continued*

Example of Research Questions	Model Specification	Parameter Interpretation
What is the effect of being required to pay a copayment at each time point on overall mean VA specialty care expenditures? What is the estimated probability of incurring VA specialty care expenditures for those with and without a copayment requirement in each year? What are the estimated or predicted overall mean expenditures for those with and without a copayment requirement in each year?	$+ \alpha_6 MUSTPAY_i \times YR2002_j$ $+ \alpha_7 MUSTPAY_i \times YR2003_j + b_{1i} + b_{2i} t_{ij}$ Eq 4.2 $g(v_{ij}) = \beta_0 + \beta_1 YR2001_j$ $+ \beta_2 YR2002_j + \beta_3 YR2003_j + \beta_4 MUSTPAY_i$ $+ \beta_5 MUSTPAY_i \times YR2001_j + \beta_6 MUSTPAY_i \times YR2002_j$ $+ \beta_7 MUSTPAY_i \times YR2003_j + b_{3i} + b_{4i} t_{ij},$ where $g(\cdot)$ is a link function relating the overall mean, v_{ij} for individual i at time j to the linear predictor, and t_{ij} takes values of 0, 1, 2, or 3 for years 2000, 2001, 2002, and 2003, respectively	dual interpretations as both subject-specific and population average. Any parameters corresponding to terms also included as random effects are solely interpreted as subject-specific estimates.

sample (Basu and Rathouz 2005), or using the "margins" command in Stata (Williams 2012) or the margins package in R (R Development Core Team 2017; Leeper and Arnold 2017), both of which compute standard errors computationally (e.g., delta method). Approaches such as recycled predictions could similarly be extended to compute difference-in-difference estimates.

In many situations, it may be unsuitable to model the zero and positive expenditures as a single component. Smith et al. (2017) showed via simulation studies that one-part models produced significantly negatively biased covariate effect estimates and high type I error rates when fit to data containing 20 percent or more zeros. In such cases, two-part models should be considered.

Uncorrelated Two-Part GLMs. Two-part models hold appeal when analysts are interested in the probability of incurring a positive expenditure or the data contain enough zeros to bias results from a one-part model. In this case, it may be preferable to conceptualize longitudinal expenditures as semicontinuous.

Two-part GLMs fit with GEEs appear a natural extension of two-part models for cross-sectional data. They utilize two different models to describe the "binary part" and "continuous part" of the semicontinuous expenditure data, and each part separately accounts for correlation among the repeated measures:

$$\text{logit}(\Pr(Y_{ij} > 0)) = \alpha_0 + \alpha_1 x_{1ij} + \cdots + \alpha_p x_{pij}$$

$$g(E(Y_{ij}|Y_{ij} > 0)) = \gamma_0 + \gamma_1 x_{1ij} + \cdots + \gamma_p x_{pij},$$

where Y_{ij} and covariates are defined as above. While both components are written with the same covariates included, this is not required. These two independent models can each be fit using GEEs.

Because of the simplicity of independently modeling the two components, analysts may be tempted to consider this extension. The second component utilizes a link function similar to one-part GLMs, so one can estimate population-average expenditures and corresponding confidence intervals in the dollar scale. However, there are critical problems with this approach (see Su, Tom, and Farewell [2009] for detailed discussion).

Interpretation of the resulting estimates is not straightforward. The binary model provides population-average estimates of the probability of incurring positive expenditures *for the entire sample at all time points*. Specifically, $\exp(\alpha_k)$ is interpreted as the odds ratio for incurring a positive expenditure associated with a one-unit increase in the kth covariate. However, the continuous model only provides estimates of the level of expenditures *among*

the subset of individuals who incurred expenses at each time point rather than the entire sample; thus, the target population changes over time depending on the subset with positive expenditures. If a log link was used, $\exp(\gamma_k)$ would represent the multiplicative increase in expenditures at a given time point, conditional on incurring expenditures at that time point, associated with a one-unit increase in the kth covariate. Estimates represent different samples at different time points, so they lack clear interpretation for motivating policy decisions. Further, conditioning on positive values complicates the ability to make causal inferences from the second component, yielding causal effect interpretations erroneous or misleading; see Angrist (2001) for a more detailed discussion.

The two components of the model are often correlated over time, such that the probability of incurring any expense is associated with level of expenditures over time. Failure to account for this correlation leads to informative cluster sizes in the second component and biased results (Su, Tom, and Farewell 2009). Therefore, we cannot in general recommend use of uncorrelated two-part GLMs; it is valid if independence between the two components can be justified.

Correlated Two-Part Models—Conditional Specification. The correlated conditional two-part (CTP) random-effects model allows for estimation of correlation between the binary and continuous parts of the longitudinal expenditures by specifying a joint random-effects distribution, including correlations/covariance between the random effects (Olsen and Schafer 2001; Tooze, Grunwald, and Jones 2002; Liu et al. 2010). The joint random-effects distribution quantifies how the two parts of the outcome are related over time if, for example, receipt of specialist care leads to identification of previously undiagnosed conditions that require follow-up and treatment:

$$\text{logit}(\Pr(Y_{ij} > 0)) = \alpha_0 + \alpha_1 x_{1ij} + \ldots + \alpha_p x_{pij} + b_{1i}$$

$$E(\log(Y_{ij}|Y_{ij} > 0)) = \delta_0 + \delta_1 x_{1ij} + \cdots + \delta_p x_{pij} + b_{2i}$$

where b_{1i} and b_{2i} are random intercepts assumed to jointly follow a multivariate normal distribution as $\begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$. Correlation is allowed between the two components via $\boldsymbol{\Sigma}$, and while only random intercepts are incorporated above, additional random effects could be included.

A generalized linear mixed model with logit link is generally specified for the binary component. Random effects are included to capture the correlation structure of interest (e.g., random intercepts and coefficient for time). This is a subject-specific model, so parameter estimates are conditional upon the

random effects (Table 1). See Hedeker and Gibbons (2006), Fitzmaurice, Laird, and Ware (2012), and Diggle et al. (2002) for discussion of interpreting subject-specific estimates. In the continuous component, there are several modeling options depending upon the distributional characteristics of the positive expenditures. The lognormal is shown above. Liu et al. (2010, 2016) recommend generalized gamma mixed-effects models, which include the lognormal, gamma, inverse gamma, and Weibull distributions as special cases. Using model estimates to convert predictions from the log-dollar scale back to the dollar scale is not computationally feasible for a conditional model. However, one can use results from these models to obtain the average partial elasticity of the mean cost (for continuous covariates) or the log ratio of expected cost (for binary covariates) following Liu et al. (2010).

Intra-individual variation in expenditures over time is also modeled via random effects, often a random intercept and random coefficient for time. The variance components estimates represent the covariance between the random intercepts and coefficients for the probability of any expenditures and positive expenditures. It is useful to test whether or not these are equal to 0, indicating that the two parts of the model are separable (Olsen and Schafer 2001; Manning, Basu, and Mullahy 2005). Additionally, the analyst can use them to interpret how the two parts of the model are related. For example, a positive estimate for the covariance between the binary component's random intercept and the continuous component's random slope indicates that probability of any expenditure is positively related to the amount of expenditures over time.

Parameter estimation in the correlated CTP model is computationally challenging and interpretation is difficult, particularly for complex random-effects specifications (Table 1). This model can be fit via maximum-likelihood estimation in several software packages, including Mplus (www.statmodel.com, referred to as a "two-part semicontinuous growth model") and PROC NLMIXED in SAS (Cary, NC). For an illustration of model estimation in PROC NLMIXED under the generalized gamma distribution, see Liu et al. (2010). It is also possible to fit this model using Bayesian methodology (Cooper et al. 2007).

Correlated Two-Part Models—Marginalized Specification. The fourth approach is the marginalized two-part (MTP) model that provides a blend between the marginal interpretations of the one-part GLMs and structure of the correlated CTP model (Smith et al. 2015). Like other two-part models, the first part of the MTP model is a probability of use component that is typically fit using a GLM with logit link and random effects. The continuous component is no

longer conditional on having incurred expenditures in a given time interval. In the MTP model, the continuous part of the model represents the marginal mean by incorporating zero and positive values, along with random effects to allow for intra-individual variability in the marginal mean over time. Similar to the correlated two-part model, the random effects from the two components of the MTP model are assumed to be jointly normally distributed, allowing for dependence between the two parts of the model:

$$\text{logit}(\Pr(Y_{ij} > 0)) = \alpha_0 + \alpha_1 x_{1ij} + \cdots + \alpha_p x_{pij} + b_{1i}$$

$$\log(E(Y_{ij})) = \beta_0 + \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + b_{2i}$$

where b_{1i} and b_{2i} are random intercepts that are assumed to jointly follow a multivariate normal distribution as $\begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} \sim MVN(\mathbf{0}, \mathbf{\Sigma})$.

Like the correlated CTP model, fitting the two components simultaneously can be computationally challenging. Fitting the model has been proposed in a Bayesian framework using SAS PROC MCMC, although it has also been implemented using maximum likelihood (Burgette et al. 2017). Noninformative priors are suggested, and this framework allows incorporation of complicated random effect structures, such as fully correlated random intercepts and slopes in both components of the model (see Smith et al. 2015).

Similar to the correlated CTP model, parameter estimates in the binary component are subject-specific estimates. Specifically, $\exp(\alpha_k)$ represents the subject-specific odds ratio for incurring positive expenditures associated with a one-unit increase in the kth covariate. Parameter estimates in the second component represent effects on the overall mean, and those corresponding to covariates not included as random effects are both subject-specific and population average (Table 1). If a random intercept and slope for time were included, for example, the intercept and estimated coefficient for time would be subject-specific. Any covariates representing treatment or treatment by time interactions, on the other hand, would have both subject-specific and population-average interpretations. Therefore, $\exp(\beta_k)$ represents the multiplicative effect on the overall mean expenditures of the entire population associated with a one-unit increase in the kth covariate. These quantities, and any value calculated from the parameters, can be obtained easily along with corresponding credible intervals or highest posterior density intervals [HPDI] (a Bayesian analog to confidence intervals) via statements in PROC MCMC. Specifically, additive mean differences from the MTP model, on the dollar scale, can be obtained directly from model output, along with 95 percent HPDIs, without

need for postmodeling computations. Additionally, population-average estimates on the dollar scale are easily computed and represent the overall mean of both the zero and positive values (Table 1).

The most notable difference between the correlated CTP model and the correlated MTP model is the interpretation of the parameter estimates in the second part of the model (Table 1). In the correlated CTP model, these represent effects on the conditional mean of the subsample who incurred positive expenditures in any given time interval. Under the MTP model, parameter estimates in the second part represent effects on the overall marginal mean of the entire sample.

One can similarly estimate variance components from the MTP model and assess the level of correlation between the two components of the model. In the MTP model, these covariances represent different underlying quantities than they do in the conditional model because the second component includes the zero values. With the MTP model, the covariances represent relationships between the binary part and the overall unconditional mean rather than between the binary part and the mean of the positive values as in the conditional model.

DATA ILLUSTRATION

One-Part Generalized Linear Model via Generalized Estimating Equations

To illustrate, we fit a one-part GLM via GEEs to our example of veterans' longitudinal specialty expenditures. Following specification testing (Manning and Mullahy 2001), a log link with a proportional mean–variance relationship was determined to fit best. Model-estimated mean expenditures were $1,011 (95 percent CI: $916, $1,116) in 2000 for those not required to pay copayments and $797 (95 percent CI: $719, $883) for those required to pay. By 2003, mean expenditures increased to $1,200 (95 percent CI: $1,102, $1,306) for those not required to pay copayments, but remained steady at $798 (95 percent CI: $714, $892) for those who were (Table 2). Parameter estimates from this model are shown in Table 3, and multiplicative mean effects of copayment requirement each year are shown in Table 2. By looking at the parameter estimates corresponding to years 2001 through 2003 (β_1, β_2, β_3, in equation 1.1 from Table 1, respectively), we see that for those without a copayment, expenditures increased over time and were significantly higher in 2003 than in 2000. For those with a copayment, expenditures were lower than those without. To estimate the difference between groups at any given year, one can

Table 2:   Model-Estimated Effects of Copayment Requirement in Each Year

Year	Odds Ratio of Positive Expenditures for Those Required to Pay vs. Not (95% HPDI*)	Model-Estimated Mean Expenditures among Those Required to Pay Copayments (95% CI/HPDI*)	Model-Estimated Mean Expenditures among Those Not Required to Pay Copayments (95% CI/HPDI)	Multiplicative Effect on Overall Mean (95% CI/HPDI)	Additive Effect on the Overall Mean Expenditures for Those Required to Pay vs. Not (95% CI/HPDI)†
Estimates from one-part GLM fit with GEEs					
2000		$797 (719, 883)	$1,011 (916, 1,116)	0.79 (0.68, 0.91)	−$214
2001		$860 (781, 947)	$1,067 (977, 1,164)	0.81 (0.71, 0.92)	−$207
2002		$821 (740, 911)	$1,084 (981, 1,199)	0.76 (0.66, 0.88)	−$263
2003		$798 (714, 892)	$1,200 (1,102, 1,306)	0.67 (0.58, 0.77)	−$402
Estimates from correlated MTP model					
2000	0.35 (0.26, 0.44)	$887 (792, 998)	$1,250 (1,107, 1,404)	0.71 (0.61, 0.80)	−$363 (−513, −217)
2001	0.35 (0.27, 0.45)	$893 (806, 984)	$1,289 (1,173, 1,434)	0.69 (0.61, 0.78)	−$396 (−549, −266)
2002	0.30 (0.23, 0.39)	$821 (744, 903)	$1,418 (1,272, 1,569)	0.58 (0.51, 0.65)	−$597 (−745, 453)
2003	0.25 (0.18, 0.33)	$848 (758, 949)	$1,651 (1,482, 1,845)	0.51 (0.45, 0.58)	−$803 (−981, −634)

Table 2. *Continued*

Year	Odds Ratio of Positive Expenditures for those Required to Pay vs. Not (95% CI)	Conditional Model-Estimated Mean Expenditures among those with Positive Expenditures and Required to Pay Copayments (95% CI)	Model-Estimated Mean Expenditures among those with Positive Expenditures and Not Required to Pay Copayments (95% CI)	Multiplicative Effect on Conditional Mean (95% CI)	Additive Effect on the Conditional Mean Expenditures for those Required to Pay vs. Not[†]
Estimates from uncorrelated two-part GLMs fit with GEEs					
2000	0.52 (0.45, 0.60)	$1,150 (1,038, 1,275)	$1,257 (1,141, 1,385)	0.92 (0.79, 1.05)	−$107
2001	0.49 (0.42, 0.58)	$1,204 (1,096, 1,323)	$1,264 (1,163, 1,373)	0.95 (0.84, 1.08)	−$60
2002	0.44 (0.37, 0.52)	$1,120 (1,011, 1,240)	$1,257 (1,140, 1,386)	0.89 (0.77, 1.03)	−$137
2003	0.45 (0.39, 0.53)	$1,121 (1,001, 1,256)	$1,416 (1,301, 1,540)	0.79 (0.69, 0.91)	−$295

*HPDI = Highest posterior density interval, applicable to estimates from the correlated MTP model fit with a Bayesian MCMC algorithm.
[†]Note 95% CIs are not provided for the one-part GLM nor the uncorrelated two-part GLM because they are not provided as part of standard model output.

Table 3: Model-Estimated Parameter Estimates and Standard Errors from the One-Part GLM, Uncorrelated Two-Part GLMs, and Correlated Marginalized Two-Part Random Effects Model (REM)

Coefficient	One-Part GLM* Estimate of Effects on $\log E(Cost_{ij})$ (SE)	Uncorrelated Two-Part GLMs[†] Estimate of Effect on $PR(Cost_{ij} > 0)$ (SE)	Estimate of Effect on $\log E(Cost_{ij}\|Cost_{ij} > 0)$ (SE)	Correlated Marginalized Two-Part REM[‡] Estimate of Effect on $Pr(Cost_{ij} > 0\|\mathbf{b}_i)$ (SD)	Estimate of Effect on $\log E(Cost_{ij})$ (SD)
Intercept	6.92 (0.05)	1.24 (0.06)	7.14 (0.05)	2.13 (0.12)	6.24 (0.05)
Year2001	0.05 (0.05)	0.19 (0.06)	0.006 (0.05)	0.31 (0.11)	0.13 (0.05)
Year2002	0.07 (0.06)	0.42 (0.07)	0.0005 (0.06)	0.83 (0.14)	0.23 (0.05)
Year2003	0.17 (0.05)	0.22 (0.07)	0.12 (0.06)	0.65 (0.17)	0.28 (0.05)
MUSTPAY	−0.24 (0.07)	−0.65 (0.08)	−0.09 (0.07)	−1.06 (0.14)	−0.34 (0.07)
MUSTPAY*Year2001	0.02 (0.08)	−0.05 (0.09)	0.04 (0.08)	0.005 (0.14)	−0.02 (0.07)
MUSTPAY*Year2002	−0.04 (0.09)	−0.17 (0.10)	−0.03 (0.09)	−0.15 (0.16)	−0.20 (0.07)
MUSTPAY*Year2003	−0.17 (0.09)	−0.14 (0.09)	−0.14 (0.09)	−0.34 (0.18)	−0.32 (0.07)

*Fit via GEE with proportional mean–variance, log link, and unstructured correlation.

[†]Fit via GEE using a binomial variance and logit link for the first component and proportional mean–variance and log link for the second. Unstructured correlation was used for both components.

[‡]Fit via MCMC with noninformative prior distributions; note estimates from the marginalized model are posterior means and standard deviations.

exponentiate the sum of the coefficients for *MUSTPAY* and the interaction of *MUSTPAY* and the year of interest. For example, the estimated ratio of dollars spent by those with the copayment requirement in 2001 versus those without was estimated as $\exp(\beta_4 + \beta_5)$ from equation 1.1 in Table 1. Using model-estimated parameters, this becomes $\exp(-0.24 + 0.024) = 0.81$ with 95 percent CI (0.71, 0.92). These quantities are easily obtained with SAS ESTIMATE statements (see code in Appendix SA2) and are interpreted as those required to pay copayments incurred in 2001, on average, 0.81 times the specialty care expenditures of those not required to pay copayments.

If interest lies in additive differences, SAS ESTIMATE statements no longer provide these nonlinear combinations of parameters. Instead, the difference can be obtained by subtracting the estimated means and confidence intervals can be obtained via bootstrapping or the delta method (Efron 1982; Cox 1990).

Two-Part Generalized Linear Models via Generalized Estimating Equations

As with the one-part GLM, a log link with proportional mean–variance relationship fit the conditionally positive data well. The first part of the model was fit with a logit link and binomial variance. Both parts incorporated an unstructured covariance and used empirical standard errors (Table 3). Odds ratios for the probability of incurring positive expenditures and multiplicative effects on the conditional mean of the positive values are shown in Table 2.

From the first part of this model, the odds ratios suggest that the probability of incurring positive specialty expenditures was significantly lower for those required to make copayments compared to those not required. The effect of copayment on the conditional mean among those incurring expenditures appears less strong than the effect of copayment on the probability of incurring expenditures. The multiplicative effect on the conditional mean ranges from 0.92 in year 2000 to 0.79 in year 2003, which can be interpreted as, among those who incurred expenditures in year 2000, those required to pay copayments incurred on average 0.92 times the expenditures of those not required to pay copayments in 2000 and 0.79 times in 2003. Note the subsample incurring expenditures in year 2000 is likely not the same subsample incurring expenditures in 2003, which complicates interpretation of results. We also urge caution in interpreting these estimates as estimates from uncorrelated two-part longitudinal models can be biased (Su, Tom, and Farewell 2009).

Correlated Two-Part Models—Conditional Specification

For the CTP specification, a linear mixed-effects model was specified for the log-transformed positive expenditures. Unlogged positive specialty expenditures exhibited significant skewness (5.6) and kurtosis (75.3), whereas log-transformed positive specialty expenditures were nearly normally distributed (skewness = −0.1, kurtosis = 2.7). As shown in equations 3.1 and 3.2 in Table 1, intra-individual variation in outcomes over time was modeled via three random effects that were assumed to be jointly normally distributed. The binary component included a random intercept, and the continuous component included a random intercept and a random coefficient of time.

Results from the binary component (Table 4) show that a veteran required to pay specialty visit copayments was less likely to have a specialty visit. Trends over time were similar for veterans who were and were not required to pay specialty visit copayments; a veteran's probability of a specialty visit was higher in 2001 through 2003 than in 2000, as indicated by the significant time effects (i.e., α_1 through α_3 in equation 3.2). Among veterans who had positive specialty expenditures, those required to pay specialty visit copayments had lower specialty expenditures in 2000 than those exempt from copayments (i.e., δ_4 in equation 3.2) and compared to those not required to pay, their expenditures continued to decrease over time as indicated by significant interaction terms (i.e., δ_5 through δ_7 in equation 3.2). Although we can understand general comparisons and trends from this model, deriving policy-related estimates is not straightforward. One option is to calculate the log ratio of the outcome for specific levels of a covariate (e.g., predicted specialty expenditures between exempt and nonexempt veterans at each year) by taking the expectation with respect to the marginal distribution of the random effect, as shown in equation (9) in Liu et al. (2010). These log ratios incorporate both parts of the model to provide an overall estimate of how demand for specialty care (on the log scale) differs for co-pay-exempt and nonexempt patients. A log ratio of zero indicates equivalence between groups; negative values, as estimated in our example (Table 4), indicate those required to pay incurred lower expenditures than those not required. Confidence intervals can be generated via bootstrapping (Liu et al. 2010; Maciejewski et al. 2012).

Additionally, the joint random-effects specification allows us to quantify how the two parts of the model are related over time. The correlation between the random intercepts of the two parts was 0.7, indicating a strong, positive correlation between the probability of having a specialty visit and higher amounts of specialty expenditures in 2000. In contrast, the correlation

Table 4:   Model-Estimated Parameter Estimates, Standard Errors, and Log Ratios from the Correlated Conditional Two-Part REM*

Coefficient	Estimate of Effect on $\Pr(\text{Cost}_{ij} > 0\mid\mathbf{b}_i)$ (SE)	Estimate of Effect on $E[\ln(\text{Cost}_{ij}\mid \text{Cost}_{ij} > 0)]$ (SE)	Year	Log Ratio of Expected Specialty Expenditures for Those Required to Pay Copayments vs. Those Not Required (95% CI)
Intercept	2.01 (0.10)	6.04 (0.04)	2000	−0.30 (−0.41,−0.18)
Year2001	0.29 (0.11)	0.07 (0.04)	2001	−0.05 (−0.18,0.08)
Year2002	0.64 (0.11)	0.13 (0.04)	2002	−0.26 (−0.40,−0.13)
Year2003	0.33 (0.11)	0.23 (0.04)	2003	−0.35 (−0.50,−0.22)
MUSTPAY	−1.03 (0.13)	−0.17 (0.06)		
MUSTPAY*Year2001	−0.06 (0.14)	−0.04 (0.06)		
MUSTPAY*Year2002	−0.23 (0.14)	−0.23 (0.06)		
MUSTPAY*Year2003	−0.21 (0.14)	−0.31 (0.06)		

*Fit via ML on log expenditures with random intercept in binary part and random intercept and slope in conditionally positive continuous part.

between the random intercept in the binary part and the random slope in the continuous part was −0.01 and not statistically significant, suggesting no association between the probability of specialty care in 2000 and the change in amount of specialty expenditures over time.

Correlated Two-Part Models—Marginalized Specification

The MTP model is fit following a Bayesian approach with noninformative prior distributions for all parameters, which allows computational flexibility and may be more successful than maximum-likelihood estimation in incorporating correlated random slopes.

From the first component, the odds ratios suggest that the probability of incurring positive specialty expenditures was significantly lower for those with required copayments compared to those without (Table 2), consistent with results from the first components of the uncorrelated two-part GLMs and the correlated CTP model. Parameter estimates from both MTP components are shown in Table 3.

To calculate the multiplicative effect of copayment requirement on the overall mean, one exponentiates the sum of the coefficients for *MUSTPAY* and the interaction terms (e.g., $\exp(\beta_4 + \beta_5)$ in equation 4.2 for 2001). The effect of copayment on the overall mean suggests notably lower overall expenditures for those required to pay copayments, ranging 0.71 times the expenditures of those

not required to pay in year 2000 to 0.51 times in 2003. Examining additive differences in means, we see a similar pattern, with expenditure differences among those with required copayments ranging from $363 lower per year (95 percent highest posterior density interval [HPDI] −$513 to −$217) 2 years prior to the copayment increase to $803 lower per year (95 percent HPDI −$981 to −$634). These estimates are quite different from those generated from the one-part GLM, which is consistent with simulation results suggesting that one-part GLMs fit with GEEs provide negatively biased estimates in the presence of many zeros (Smith et al. 2017). Magnitude of estimates from the uncorrelated two-part GLMs or correlated CTP model cannot be compared as the target of inference in the second component is those with positive expenditures.

DISCUSSION

Estimation of health expenditures is challenging when there is a point mass at zero and a heavily skewed distribution of positive expenditures. In longitudinal applications, these challenges are compounded by the correlation over time between the probability of incurring positive expenditures and the level of expenditures. We have illustrated the relative strengths and limitations and the research questions addressed by four estimation strategies.

The decision of which modeling approach to employ should be driven first by the research question of interest, and secondly, by the data structure. To determine the appropriate approach for the research question, two main considerations should be answered. First, is there interest in what influences the probability of incurring expenditures? If so, a two-part model may be appropriate. Secondly, is the primary interest in overall mean expenditures of the entire population or is more interest in the level of expenditures conditional on them being incurred? If the former, the one-part GLM or the MTP model is preferable; if the latter, one should consider a CTP model. The uncorrelated two-part GLM should only be considered if the analyst feels confident of no correlation between the two components, often an untenable assumption. Even if such assumptions are met and interest is only in the level of expenditures conditional on them being incurred, care must also be taken with causal interpretations from the second component of this model due to its conditioning on positive values (Angrist 2001). Once the research questions are identified, the analyst must examine the data for its model fit. If there are a significant proportion of zeros (i.e., more than 10 percent), we recommend use of a two-part model (Smith et al. 2017).

Longitudinal cost modeling continues to be an area of research with many evolving specialized methods. We intend this study not to be a comprehensive review of all possible approaches to longitudinal expenditure modeling, nor do we intend it to formally examine the performance of each model under varying data scenarios. For a more comprehensive general overview of all current approaches, we refer readers to Neelon, O'Malley, and Smith (2016) and Farewell et al. (2016); for a more comprehensive examination of model performance under varying data structures, we refer readers to Smith et al. (2017). Instead, we hope this will be a helpful guide when grappling with the challenges inherent in analyzing semicontinuous, highly skewed expenditure outcomes.

ACKNOWLEDGMENTS

REFERENCES

Angrist, J. D. 2001. "Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice." *Journal of Business & Economic Statistics* 19 (1): 2–28.

Basu, A., and W. G. Manning. 2009. "Issues for the Next Generation of Health Care Cost Analyses." *Medical Care* 47 (7_Supplement 1): S109–14.

Basu, A., and P. J. Rathouz. 2005. "Estimating Marginal and Incremental Effects on Health Outcomes Using Flexible Link and Variance Function Models." *Biostatistics* 6 (1): 93–109.

Buntin, M. B., and A. M. Zaslavsky. 2004. "Too Much Ado about Two-Part Models and Transformation? Comparing Methods of Modeling Medicare Expenditures." *Journal of Health Economics* 23 (3): 525–42.

Burgette, J. M., J. S. Preisser, M. Weinberger, R. S. King, J. Y. Lee, and R. G. Rozier. 2017. "Enrollment in Early Head Start and Oral Health-Related Quality of Life." *Quality of Life Research* 28: 1–12.

Chen, E. J., Y. Liu, and F. H. Lee. 2016. "A Statistical Model for the Unconfined Compressive Strength of Deep-Mixed Columns." *Geotechnique* 66 (5): 351–65.

Cooper, N. J., P. C. Lambert, K. R. Abrams, and A. J. Sutton. 2007. "Predicting Costs Over Time Using Bayesian Markov Chain Monte Carlo Methods: An Application to Early Inflammatory Polyarthritis." *Health Economics* 16 (1): 37–56.

Cox, C. 1990. "Fieller's Theorem, the Likelihood, and the Delta Method." *Biometrics* 46: 709–18.

Deaton, A., and M. Irish. 1984. "Statistical Models for Zero Expenditures in Household Budgets." *Journal of Public Economics* 23 (1–2): 59–80.

Diehr, P., D. Yanez, A. Ash, M. Hornbrook, and D. Lin. 1999. "Methods for Analyzing Health Care Utilization and Costs." *Annual Review of Public Health* 20 (1): 125–44.

Diggle, P., P. Heagerty, K. Y. Liang, and S. L. Zeger. 2002. *Analysis of Longitudinal Data.* Oxford: Oxford University Press.

Duan, N., W. G. Manning, C. N. Morris, and J. P. Newhouse. 1983. "A Comparison of Alternative Models for the Demand for Medical Care." *Journal of Business & Economic Statistics* 1 (2): 115–26.

Efron, B. 1982. *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia: SIAM.

Farewell, V., D. L. Long, B. D. Tom, S. Yiu, and L. Su. 2016. "Two-Part and Related Regression Models for Longitudinal Data." *Annual Review of Statistics and Its Application* 4: 283–315.

Fitzmaurice, G. M., N. M. Laird, and J. H. Ware. 2012. *Applied Longitudinal Analysis.* Hoboken, NJ: John Wiley & Sons.

Hedeker, D. R., and R. D. Gibbons. 2006. *Longitudinal Data Analysis.* Hoboken, NJ: Wiley-Interscience.

Hernan, M. A., and J. M. Robins. 2018. *Causal Inference.* CRC Boca Raton, FL.

Liang, K. Y., and S. L. Zeger. 1986. "Longitudinal Data-Analysis Using Generalized Linear-Models." *Biometrika* 73 (1): 13–22.

Leeper, T. J., and J. Arnold. 2017. *Marginal Effects for Model Objects.* R package version 0.3.0.

Liu, L., R. L. Strawderman, M. E. Cowen, and Y. C. Shih. 2010. "A Flexible Two-Part Random Effects Model for Correlated Medical Costs." *Journal of Health Economics* 29 (1): 110–23.

Liu, L., R. L. Strawderman, B. A. Johnson, and J. M. O'Quigley. 2016. "Analyzing Repeated Measures Semi-Continuous Data, with Application to an Alcohol Dependence Study." *Statistical Methods in Medical Research* 25 (1): 133–52.

Maciejewski, M. L., C. F. Liu, A. L. Kavee, and M. K. Olsen. 2012. "How Price Responsive Is the Demand for Specialty Care?" *Health Economics* 21 (8): 902–12.

Madden, C. W., B. P. Mackay, S. M. Skillman, M. Ciol, and P. K. Diehr. 2000. "Risk Adjusting Capitation: Applications in Employed and Disabled Populations." *Health Care Management Science* 3 (2): 101–9.

Manning, W. G., A. Basu, and J. Mullahy. 2005. "Generalized Modeling Approaches to Risk Adjustment of Skewed Outcomes Data." *Journal of Health Economics* 24 (3): 465–88.

Manning, W. G., and J. Mullahy. 2001. "Estimating Log Models: To Transform or Not to Transform?" *Journal of Health Economics* 20 (4): 461–94.

Neelon, B., A. J. O'Malley, and V. A. Smith. 2016. "Modeling Zero-Modified Count and Semicontinuous Data in Health Services Research Part 1: Background and Overview." *Statistics in Medicine* 35 (27): 5070–93.

Olsen, M. K., and J. L. Schafer. 2001. "A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data." *Journal of the American Statistical Association* 96 (454): 730–45.

R Development Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Smith, V. A., B. Neelon, J. S. Preisser, and M. L. Maciejewski. 2015. "A Marginalized Two-Part Model for Longitudinal Semicontinuous Data." *Statistical Methods in Medical Research* 26: 1949–68.

Smith, V. A., B. Neelon, M. L. Maciejewski, and J. S. Preisser. 2017. "Two Parts Are Better Than One: Modeling Marginal Means of Semicontinuous Data." *Health Services and Outcomes Research Methodology* 17 (3–4): 198–218.

Su, L., B. D. M. Tom, and V. T. Farewell. 2009. "Bias in Two-Part Mixed Models for Longitudinal Semicontinuous Data." *Biostatistics* 10 (2): 374–89.

Tooze, J. A., G. K. Grunwald, and R. H. Jones. 2002. "Analysis of Repeated Measures Data with Clumping at Zero." *Statistical Methods in Medical Research* 11 (4): 341–55.

Williams, R. 2012. "Using the Margins Command to Estimate and Interpret Adjusted Predictions and Marginal Effects." *Stata Journal* 12 (2): 308.

Supporting Information

Additional supporting information may be found online in the supporting information tab for this article:

Appendix SA1: Author Matrix.
Appendix SA2: Example SAS Code for Each Modeling Strategy.
Figure S1: Specialty Care Expenditure Distributions by Year.