Research Article

Statistics in Medicine

Received 27 January 2013, Accepted 26 May 2014 Published online 29 June 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6239

A latent variable transformation model approach for exploring dysphagia

Anna C. Snavely,[a,b*†] David P. Harrington[c,d] and Yi Li[e]

Multiple outcomes are often collected in applications where the quantity of interest cannot be measured directly or is difficult or expensive to measure. In a head and neck cancer study conducted at Dana-Farber Cancer Institute, the investigators wanted to determine the effect of clinical and treatment factors on unobservable dysphagia through collected multiple outcomes of mixed types. Latent variable models are commonly adopted in this setting. These models stipulate that multiple collected outcomes are conditionally independent given the latent factor. Mixed types of outcomes (e.g., continuous vs. ordinal) and censored outcomes present statistical challenges, however, as a natural analog of the multivariate normal distribution does not exist for mixed data. Recently, Lin *et al.* proposed a semiparametric latent variable transformation model for mixed outcome data; however, it may not readily accommodate event time outcomes where censoring is present. In this paper, we extend the work of Lin *et al.* by proposing both semiparametric and parametric latent variable models that allow for the estimation of the latent factor in the presence of measurable outcomes of mixed types, including censored outcomes. Both approaches allow for a direct estimate of the treatment (or other covariate) effect on the unobserved latent variable, greatly enhancing the interpretability of the models. The semiparametric approach has the added advantage of allowing the relationship between the measurable outcomes and latent variables to be unspecified, rendering more robust inference. The parametric and semiparametric models can also be used together, providing a comprehensive modeling strategy for complicated latent variable problems. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: latent variables; semiparametric modeling; multiple outcomes; dysphagia

1. Introduction

Investigators often collect multiple outcomes in settings where the quantity of interest cannot be measured directly or is difficult or expensive to measure [1]. In this scenario, each of the measured outcomes provides information about the unobservable quantity of interest, with each outcome possibly capturing a different aspect of that quantity. Latent variable models are commonly adopted in this setting. In these models, the multiple outcomes are assumed to be conditionally independent given a latent factor, where the latent factor represents the quantity of interest in the model. Mixed types of outcomes (e.g., continuous vs. ordinal) are common in many applications [2] as are censored outcomes. These varying outcome types present statistical challenges as a natural extension of the multivariate normal distribution for mixed data does not exist. For example, in a head and neck cancer (HNC) study conducted at Dana-Farber Cancer Institute (DFCI), the investigators wanted to determine the effect of clinical and treatment factors on dysphagia (or difficulty in swallowing) [3]. This goal is challenging, however, because multiple measures can be used to describe dysphagia. In particular, simply capturing dysphagia through patient reports of swallowing difficulty may not provide a complete picture of the condition as studies have shown that patients' perceptions of swallowing do not always align well with what is observed on video swallow studies [4]. Also, as HNC patients often require feeding tubes, swallowing may not be tested enough for patients to know their own swallowing capabilities. Therefore, the investigators collected three objective

[a]*Department of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, U.S.A.*
[b]*UNC Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, U.S.A.*
[c]*Department of Biostatistics, Harvard School of Public Health, Boston, MA, U.S.A.*
[d]*Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, U.S.A.*
[e]*Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, U.S.A.*
[*]*Correspondence to: Anna C. Snavely, UNC Lineberger Comprehensive Cancer Center, Chapel Hill, NC, U.S.A.*
[†]*E-mail: anna_snavely@med.unc.edu*

surrogate outcome measures to capture aspects of dysphagia: duration of feeding tube usage, weight loss after treatment, and diet (liquid, soft, etc.). Feeding tube duration is measured from the end of treatment to the tube removal. However, as some patients still had a feeding tube at the time of last follow-up, this outcome is subject to censoring. The other two outcomes are of mixed types; weight loss was measured on a continuous scale, whereas diet was measured on an ordinal scale. Limited statistical tools for accommodating such complicated data have hampered proper analyses.

When the measurable outcomes are all continuous, the methods are well developed within the latent variable paradigm [5, 6]. Some methods also exist in the context of latent variable modeling when the outcomes are of mixed types; see, for example, [7–12]. However, a limitation of these models for mixed outcomes is that the relationship between the measurable outcomes and the Gaussian latent variable must be known *a priori*. As the latent variable is not observed, there is little guidance for the appropriate relationship. If the relationship is misspecified, using the common likelihood approaches leads to biased estimates for the parameters. Recently, Lin *et al*. [13] proposed a semiparametric latent variable transformation model to address some of these shortcomings. However, none of the approaches described earlier allow for survival or event time outcomes where censoring is present.

Extending the work of Lin *et al*. [13], we propose a class of semiparametric latent variable models that allows for the estimation of the latent factor in the presence of mixed outcomes types as well as censored outcomes. Our proposed method allows the relationship between the measurable outcomes and latent variable to be unspecified, rendering more robust inference, and allows for direct estimation of the treatment (or other covariate) effect on the unobserved latent variable. We further propose a class of parametric latent variable models that also handles outcomes of mixed types, including censored outcomes, and allows for direct estimation of the treatment effect on the unobserved latent variable. The semiparametric and parametric models are presented together in this paper in order to facilitate making comparisons between the two approaches and to highlight the idea that the semiparametric approach can be used to inform a parametric model. In this way, the modeling strategies can be used together, providing a comprehensive approach for complicated latent variable problems.

Section 2 discusses the semiparametric model, and Section 3 describes the parametric approach. Section 4 discusses model diagnostics. Simulation results are presented in Section 5, and the HNC study is analyzed in Section 6. We conclude with a discussion in Section 7.

2. Semiparametric latent variable transformation model

2.1. The model

Suppose there are n subjects, each with p distinct measurable outcomes. For simplicity, we will focus on the setting where there is a single outcome that is subject to censoring, though the extension to accommodate multiple censored outcomes is straightforward. Without loss of generality, we assume that the first measurable outcome is a continuous event time, denoted by T, which can be censored by a competing censoring variable, denoted by C. We further assume that T and C are independent and that C is independent of the covariates. Let $Y_{i1} = \min(T_i, C_i)$ and $\Delta_i = I(Y_{i1} = T_i)$, where $I(\cdot)$ is the indicator function. Then, for each individual i, we observe vectors of covariates $X_{i1}, \cdots, X_{ip}$ (e.g., age and gender) and Z_i (e.g., treatment), a failure indicator Δ_i, and a vector of measurable outcomes $Y_i = (Y_{i1}, \cdots, Y_{ip})^T$. The elements of Y_i are ordered such that the first p_1 elements are continuous (with the first element being the event time), and the remaining $p_2 = p - p_1$ elements are ordinal (including binary). The ordinal measurable outcomes are linked to underlying continuous variables as in [1, 14]. Specifically, let Y_{ij}^u be a continuous variable underlying Y_{ij}. Then, for the ordinal outcomes, for $Y_{ij} \in \{1, \cdots, d_j\}$, $Y_{ij} = \sum_{l=1}^{d_j} lI(c_j(l-1) < Y_{ij}^u \leqslant c_j(l))$, where d_j is the number of categories for the jth outcome and $c_j = (c_j(0), \cdots, c_j(d_j))^T$ are unknown thresholds satisfying $-\infty = c_j(0) < \cdots < c_j(d_j) = \infty$. As d_j can be close to ∞ as $n \to \infty$, this method could also accommodate count data. For the measurable outcomes that are already continuous, $Y_{ij} = Y_{ij}^u$.

Following Lin *et al*. [13], we relate the continuous or underlying continuous outcomes to the latent variable (e_i) of primary interest through a semiparametric linear transformation model:

$$\begin{aligned} H_1(T_i) &= X_{i1}^T\beta_1 + \alpha_1 e_i + \varepsilon_{i1}, \\ H_2(Y_{i2}^u) &= X_{i2}^T\beta_2 + \alpha_2 e_i + \varepsilon_{i2}, \\ &\vdots \\ H_p(Y_{ip}^u) &= X_{ip}^T\beta_p + \alpha_p e_i + \varepsilon_{ip}. \end{aligned} \tag{1}$$

H_1 is an unknown non-decreasing transformation function such that $H_1(0) = -\infty$ and $H_2, \cdots, H_p$ are unknown non-decreasing transformations that satisfy $H_j(-\infty) = -\infty$ and $H_j(\infty) = \infty$ for $j = 2, \cdots, p$. $\beta = (\beta_1^T, \cdots, \beta_p^T)^T$ is a vector of regression coefficients, $\alpha = (\alpha_1, \cdots, \alpha_p)^T$ are factor loadings, e_i is a latent variable for subject i, and $\varepsilon_i = (\varepsilon_{i1}, \cdots, \varepsilon_{ip})^T$ is a vector of independent errors distributed as $N(0, \text{diag}(\sigma_1^2, \cdots, \sigma_p^2))$.

Furthermore, additional structure for the latent variable is assumed:

$$e_i = Z_i^T \gamma + \epsilon_i, \tag{2}$$

where Z_i records treatment or other covariates, γ is a vector of unknown regression coefficients, and ϵ_i is the random error distributed as $N(0, \sigma_e^2)$. In most instances, γ is the primary parameter for inference because it relates covariates of interest, such as treatment, to the latent variable (outcome of interest). We assume that Z_i and ϵ_i are independent and that for identifiability, Z_i and X_{ij} do not contain constant terms, $\sigma_e^2 = 1$, and $\sigma_j^2 = 1$ for $j = 1, \cdots, p$ [13]. One of the factor loadings is also constrained to be positive because of the indeterminancy between the factor loadings and the scale of the latent variable [15]. Although related, our model is different from the ordinary random effect models. Random effects are mainly introduced to describe the unobserved heterogeneity and are usually covariate independent, whereas the latent variables, e_i, represent specific traits measured by covariates and hence are covariate dependent.

2.2. *Likelihood and estimating equations*

For each given $y_j \in \{1, \cdots, d_j\}$, $j = p_1 + 1, \cdots, p$ (the ordinal measurable outcomes), let $\tilde{H}_j(y_j) = H_j(c_j(y_j))$, where c_j is the unknown upper limit of Y_{ij}^u when $Y_{ij} = y_j$. Because both H_j and Y_{ij}^u are unknown, they cannot be identified separately. However, $H_j(c_j(1)), \cdots, H_j(c_j(d_j - 1))$ provide the distribution of the observed outcome Y_{ij} and can be estimated. In other words, for the discrete measurable outcomes, estimation of the transformation means estimation of the unknown transformed thresholds. Also, let $\tilde{H}_j = H_j$ for the continuous measurable outcomes (for ease of notation), $\mathbf{\Theta} = (\beta, \alpha, \gamma)$, and $\tilde{\mathbf{H}} = (\tilde{H}_1, \cdots, \tilde{H}_p)$.

As the error terms in models (1) and (2) are assumed to be normally distributed, the vector of transformed continuous outcomes follows a multivariate normal distribution. However, as not all of the continuous outcomes are observable, Lin *et al*. [13] showed that the likelihood for the observed data can be expressed as

$$L(\mathbf{\Theta}; \tilde{\mathbf{H}}) \propto |\Sigma_{22}|^{n/2} \prod_{i=1}^{n} \int_{\mathbf{x}^{[2]} \in \mathcal{H}_i^{[2]}} \exp\left[-\frac{1}{2} \left(\begin{pmatrix} \tilde{\mathbf{H}}_i^{[1]} \\ \mathbf{x}^{[2]} \end{pmatrix} - X_i\beta - \alpha\gamma^T Z_i \right)^T \Sigma_{22}^{-1} \left(\begin{pmatrix} \tilde{\mathbf{H}}_i^{[1]} \\ \mathbf{x}^{[2]} \end{pmatrix} - X_i\beta - \alpha\gamma^T Z_i \right) \right] d\mathbf{x}^{[2]}, \tag{3}$$

where

$$X_i = \text{diag}(X_{i1}^T, \cdots, X_{ip}^T),$$

$$\tilde{\mathbf{H}}_i^{[1]} = \left(\tilde{H}_1\left(Y_{i1}^u\right), \cdots, \tilde{H}_{p_1}\left(Y_{ip_1}^u\right) \right)^T,$$

$$\tilde{\mathbf{H}}_i^{[2]} = \left(\tilde{H}_{p_1+1}\left(Y_{i,p_1+1}^u\right), \cdots, \tilde{H}_p\left(Y_{ip}^u\right) \right)^T,$$

and

$$\mathcal{H}_i^{[2]} = \prod_{j=p_1+1}^{p} \left[\tilde{H}_j\left(Y_{ij}\right), \tilde{H}_j\left(Y_{ij}+1\right) \right].$$

This likelihood arises from the fact that on the basis of models (1) and (2),

$$\tilde{\mathbf{H}}_i \equiv \left(\tilde{\mathbf{H}}_i^{[1]^T}, \tilde{\mathbf{H}}_i^{[2]^T} \right)^T \sim N\left(X_i\beta + \alpha\gamma^T Z_i, \Sigma_{22} \right),$$

where $\Sigma_{22} = \alpha\alpha^T + I_{p\times p}$. Here, $\tilde{\mathbf{H}}_i^{[1]}$ is completely observed, whereas $\tilde{\mathbf{H}}_i^{[2]}$ is only known to fall in $\mathcal{H}_i^{[2]}$. The transformed event time is included in $\tilde{\mathbf{H}}_i^{[1]}$ when the time is not censored and is included in $\tilde{\mathbf{H}}_i^{[2]}$ when

the time is censored (because now the time is not completely observed). In the case of a censored event time, the bounds of integration are $([\tilde{H}_1(Y_{i1}), \infty])$. These bounds are incorporated in $\mathcal{H}_i^{[2]}$.

The likelihood (specifically the conditional likelihood given $\tilde{\mathbf{H}}$) in Equation (3) involves the unknown transformation functions. For the purposes of estimation, a two-stage approach is used where a series of estimating equations are utilized to first estimate the transformation functions. The parameter $\boldsymbol{\Theta}$ is then estimated by maximizing the pseudo-likelihood, which is Equation (3) with the transformations replaced by their estimated values.

Let $Y_i(t) = I(Y_{i1} \geqslant t)$ and $N_i(t) = \Delta_i I(Y_{i1} \leqslant t)$. Then $\tilde{H}_1$ for the event time outcome can be estimated using [16]

$$\sum_{i=1}^{n} \left[dN_i(t) - Y_i(t) d\Lambda\{\tilde{H}_1(t) - X_{i1}^T \beta_1 - \alpha_1 Z_i^T \gamma\} \right] = 0 \quad (t \geqslant 0), \tag{4}$$

where Λ is the cumulative hazard function for the transformed event time (i.e., the cumulative hazard for $N(0, \alpha_1^2 + 1)$). For computational purposes, the following simpler (but asymptotically equivalent) estimating equations can be used [16]:

$$\begin{pmatrix} 1 - \sum_{i=1}^{n} Y_i(t_1) \Lambda\{\tilde{H}_1(t_1) - X_{i1}^T \beta_1 - \alpha_1 \gamma^T Z_i\} \\ 1 - \sum_{i=1}^{n} Y_i(t_2) \lambda\{\tilde{H}_1(t_2-) - X_{i1}^T \beta_1 - \alpha_1 \gamma^T Z_i\} \Delta\tilde{H}_1(t_2) \\ \vdots \\ 1 - \sum_{i=1}^{n} Y_i(t_K) \lambda\{\tilde{H}_1(t_K-) - X_{i1}^T \beta_1 - \alpha_1 \gamma^T Z_i\} \Delta\tilde{H}_1(t_K) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{5}$$

where $t-$ is the time just before time t, $\Delta\tilde{H}_1(t) = \tilde{H}_1(t) - \tilde{H}_1(t-)$, and K is the number of observed events. The resulting estimate of $\tilde{H}_1(\cdot)$ will be a non-decreasing step function that jumps only at the K observed event times.

The functions $\tilde{H}_j(y_j), j = 2, \cdots, p$, can be estimated using

$$\sum_{i=1}^{n} \left[I(Y_{ij} \leqslant y_j) - \Phi\left(\frac{\tilde{H}_j(y_j) - (X_{ij}^T \beta_j + \alpha_j Z_i^T \gamma)}{\sqrt{\alpha_j^2 + 1}} \right) \right] = 0, \tag{6}$$

where Φ is the standard normal cumulative distribution function. This estimating Equation (6) arises from considering the 'marginal' probability of the event $Y_{ij} \leqslant y_j$ as described in Lin *et al.* [13].

The estimator $\hat{\tilde{H}}_j(\cdot)$ of $\tilde{H}_j(\cdot)$ is a non-decreasing step function with jumps only at the observed Y_{ij} for the continuous measurable outcomes. For the ordinal measurable outcomes, the transformed thresholds are estimated through (6). Thus, we have effectively reduced the problem of solving the infinite dimensional system of equations defined by (5) and (6) to that of solving a finite system of equations.

2.3. *Estimation algorithm*

To draw inference, we propose a two-stage estimation procedure. Specifically, given $\boldsymbol{\Theta}$, Equations (5) and (6) are used to estimate the transformation functions, $\tilde{H}_j(\cdot)$, (for $j = 1, \cdots, p$) denoted by $\tilde{H}(\boldsymbol{\Theta})$. The finite parameters, $\boldsymbol{\Theta}$, are then estimated through maximizing the pseudo-likelihood, which is the likelihood function $L(\boldsymbol{\Theta}, \tilde{\mathbf{H}}(\boldsymbol{\Theta}))$ with the transformations replaced by their estimated values. Iteration between estimating the transformations and maximizing the pseudo-likelihood continues until convergence.

For implementation, we have used the following steps:

Step 1: Choose initial values for β, α, and γ. Denote these estimates by $\hat{\beta}^{(0)}$, $\hat{\alpha}^{(0)}$, and $\hat{\gamma}^{(0)}$. Using an initial estimate of 1 for each of the parameters works well in practice. Picking initial values of 0 for all of the parameters does not work well.

Step 2: Use the estimating Equations (5) and (6) with β, α, and γ set equal to $\hat{\beta}^{(0)}$, $\hat{\alpha}^{(0)}$, and $\hat{\gamma}^{(0)}$ to obtain initial estimates of the transformation functions, $\hat{\tilde{H}}_j^{(0)}(\cdot)$.

Suppose that we have estimates of β, α, γ, and $\tilde{H}_j(\cdot)$ from the $(m-1)$th iteration; denote these estimates by $\hat{\beta}^{(m-1)}$, $\hat{\alpha}^{(m-1)}$, $\hat{\gamma}^{(m-1)}$, and $\hat{\tilde{H}}_j^{(m-1)}(\cdot)$.

Step 3: Maximize the likelihood (3) with respect to β, α, and γ, replacing $\tilde{H}_j(\cdot)$ with $\hat{\tilde{H}}_j^{(m-1)}(\cdot)$, to obtain new estimates: $\hat{\beta}^{(m)}$, $\hat{\alpha}^{(m)}$, and $\hat{\gamma}^{(m)}$. We used the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method. This is a quasi-Newton method that often performs well for optimization problems [17].
Step 4: Use the estimating Equations (5) and (6) with β, α, and γ set equal to $\hat{\beta}^{(m)}$, $\hat{\alpha}^{(m)}$, and $\hat{\gamma}^{(m)}$ to obtain new estimates of the transformation functions, $\hat{\tilde{H}}_j^{(m)}(\cdot)$.
Step 5: Repeat steps 3 and 4 until predetermined convergence criteria are met.

2.4. Bootstrap

As the parameter estimates come from maximizing the pseudo-likelihood, likelihood based standard errors cannot be used for inference because these estimates do not account for the additional variability that arises from estimating the transformation functions. Therefore, we use the traditional nonparametric bootstrap to estimate standard errors for the parameters [18].

3. Parametric latent variable transformation model

3.1. The model

We can use the same model structure proposed in (1) and (2) to formulate a parametric latent variable transformation model. For each continuous or underlying continuous measurable outcome, we assume a linear transformation. The linear transformations are then linked together through a shared latent variable, e_i. In the parametric approach, the transformation functions will be pre-specified rather than estimated from the data. The continuous or underlying continuous outcomes can be related to the latent variable (e_i) of primary interest through the following model:

$$\begin{aligned} H_1(T_i) &= X_{i1}^T\beta_1 + \alpha_1 e_i + \varepsilon_{i1}, \\ H_2(Y_{i2}^u) &= X_{i2}^T\beta_2 + \alpha_2 e_i + \varepsilon_{i2}, \\ &\vdots \\ H_p(Y_{ip}^u) &= X_{ip}^T\beta_p + \alpha_p e_i + \varepsilon_{ip}. \end{aligned} \tag{7}$$

As in the semiparametric case, $\beta = \left(\beta_1^T, \cdots, \beta_p^T\right)^T$ is a vector of regression coefficients, $\alpha = (\alpha_1, \cdots, \alpha_p)^T$ are factor loadings, e_i is a latent variable for subject i, and $\varepsilon_i = (\varepsilon_{i1}, \cdots, \varepsilon_{ip})^T$ is a vector of independent errors distributed as $N(0, \text{diag}(\sigma_1^2, \cdots, \sigma_p^2))$. Unlike the semiparametric case, the Hs in (7) are pre-specified monotone transformation functions. A typical choice for the event time would be a log transformation. The identity link may be an appropriate choice for many continuous measurable outcomes (i.e., assume normality). For the ordinal measurable outcomes, a transformation is not needed because we do not observe the underlying continuous variable. Transformed thresholds $(H_j(c_j(1)), \cdots, H_j(c_j(d_j - 1)))$ can be estimated through the likelihood, where the assumption is that the underlying continuous variable has been transformed to be normally distributed.

The latent variable is assumed to have additional structure:

$$e_i = Z_i^T\gamma + \epsilon_i, \tag{8}$$

where Z_i records covariates of interest such as treatment, γ is a vector of unknown regression coefficients, and ϵ_i is the random error distributed as $N\left(0, \sigma_e^2\right)$. The assumptions are the same as in the semiparametric model, except now, $\sigma_j^2 = 1$ for only $j = p_1 + 1, \cdots, p$ (ordinal measurable outcomes) for identifiability. Also, in the parametric setting, X_{ij} can contain constant terms for the event time and continuous measurable outcomes, but not for the ordinal measurable outcomes.

3.2. Likelihood specification and parameter estimation

Using the same notation as before, the likelihood for the observed data for the parametric setting based on models (7) and (8) can be expressed as

$$L(\boldsymbol{\Theta};\tilde{\mathbf{H}}) \propto |\Sigma_{22}|^{n/2} \prod_{i=1}^{n} \int_{\mathbf{x}^{[2]} \in \mathcal{H}_i^{[2]}} \exp\left[-\frac{1}{2} \left(\begin{pmatrix} \tilde{\mathbf{H}}_i^{[1]} \\ \mathbf{x}^{[2]} \end{pmatrix} - X_i\beta - \alpha\gamma^T Z_i \right)^T \Sigma_{22}^{-1} \left(\begin{pmatrix} \tilde{\mathbf{H}}_i^{[1]} \\ \mathbf{x}^{[2]} \end{pmatrix} - X_i\beta - \alpha\gamma^T Z_i \right) \right] d\mathbf{x}^{[2]}. \tag{9}$$

The parametric likelihood (9) differs from the semiparametric case in the form of Σ_{22}. In the parametric setting, $\tilde{\mathbf{H}}_i \equiv \left(\tilde{\mathbf{H}}_i^{[1]^T}, \tilde{\mathbf{H}}_i^{[2]^T} \right)^T \sim N(X_i\beta + \alpha\gamma^T Z_i, \Sigma_{22})$, where $\Sigma_{22} = \alpha\alpha^T + \Psi$, $\Psi = \text{diag}(\sigma_1^2, \cdots, \sigma_p^2)$), and $\sigma_j^2 = 1$ for $j = p_1 + 1, \cdots, p$ (ordinal measurable outcomes). Because the parametric model does not involve unspecified transformations, an iterative estimation procedure is not necessary. The model parameters most relevant for inference are β, α, and γ. However, the transformed thresholds associated with the ordinal measurable outcomes and the variance parameters associated with the continuous measurable outcomes also need to be estimated. All of these parameters can be estimated through maximizing the likelihood (9).

Many different maximization routines could be considered, as long as constrained maximization is supported (we use the BFGS method as in the semiparametric setting [17]). For each ordinal measurable outcome, the transformed thresholds must be constrained to be ordered, and one of the factor loadings (i.e., one of the α parameters) must be constrained to be positive [15]. Because this is a constrained optimization problem, inference based on likelihood theory could be incorrect. However, simulation results suggest that the model based standard errors arising from the likelihood are reliable, and therefore, inference based on these standard errors is reasonable. In other words, the constraints do not appear to cause boundary issues in this setting, and traditional likelihood theory can be used for inference.

4. Model diagnostics

The semiparametric latent variable transformation model presented in Section 2 can be used to provide guidance for appropriate transformation functions in the parametric model. The step functions arising from the semiparametric estimation procedure can be used to inform a parametric model. For example, the shape of an estimated transformation from the semiparametric approach might suggest using a log link in a parametric model.

Residuals can be used to consider model fit in the parametric setting. For each of the measurable outcomes, separate residuals can be obtained. In the continuous case, the residuals will be $H_j(y_{ij}) - \left(X_{ij}^T \hat{\beta}_j + \hat{\alpha}_j Z_i^T \hat{\gamma} \right)$. Similar to the linear regression setting, these residuals should be normally distributed and have mean 0. Plots of the residuals versus the fitted values and Q-Q plots of the residuals are particularly useful for establishing lack of fit [19]. Simulated data suggest that misspecified transformations should be apparent in these residual plots.

For the ordinal measurable outcomes, the underlying continuous variable can be predicted from the model by $X_{ij}^T \hat{\beta}_j + \hat{\alpha}_j Z_i^T \hat{\gamma}$. The estimated transformed thresholds can then be applied to the underlying continuous variables to obtain a predicted category for each individual, i. Model fit could then be assessed by seeing how well the predicted categories and observed categories match up in a cross-classified table. Association could be assessed using a Fisher exact test, but there is limited power for this kind of test, particularly in a small sample.

5. Simulations

We evaluated the performance of both the semiparametric and parametric methods through simulations. In order to mimic the motivating HNC data, we incorporated three measurable outcomes: an event time outcome, a continuous outcome, and an ordinal outcome with five categories. For the semiparametric settings, models (1) and (2) were assumed (with $p = 3$ and Z_i and X_i being a scalar). For the parametric settings, the following model was assumed:

$$H_1(T_i) = \beta_{01} + X_i\beta_1 + \alpha_1 e_i + \varepsilon_{i1},$$
$$H_2(Y_{i2}) = \beta_{02} + X_i\beta_2 + \alpha_2 e_i + \varepsilon_{i2},$$
$$H_3(Y_{i3}^u) = X_i\beta_3 + \alpha_3 e_i + \varepsilon_{i3},$$

and

$$e_i = Z_i\gamma + \epsilon_i.$$

For both the semiparametric simulations and the correctly specified parametric simulations, $H_1 = \log$ and $H_2 = H_3 =$ Identity. The underlying continuous variables were generated from the multivariate normal distribution, $N(X_i\beta + \alpha\gamma Z_i, \Sigma_{22})$, where $\Sigma_{22} = \alpha\alpha^T + I_{3\times3}$. In the parametric case, this means we assume that $\sigma_1 = 1$ and $\sigma_2 = 1$. Through this structure, the measurable outcomes are correlated, and this correlation is determined by the α parameters. The event time outcome was then created through an anti-log transformation. Censoring was modeled with an exponential random variable, with the parameter chosen to give a particular percentage of censoring. The continuous variable did not require further transformation, and the ordinal outcome was obtained by using the underlying continuous variable arising from the multivariate normal model and then applying the following thresholds: (-∞, -1, 0, 1, 2, ∞). We assumed that there was a single continuous X covariate common to all three measurable outcomes and a single binary Z covariate to represent treatment or some other binary covariate of interest. Specifically, $X_i \sim N(0, 1)$ and $Z_i \sim$ Bernoulli(0.50). True parameter values were selected to be $\beta_1 = 0.5$, $\beta_2 = 0.9$, $\beta_3 = 0.75$, $\alpha_1 = 0.5$, $\alpha_2 = 0.9$, $\alpha_3 = 0.75$, and $\gamma = 1$. For the parametric settings, $\beta_{01} = 0.5$ and $\beta_{02} = 0.5$ (note: intercept terms are not permitted in the semiparametric model).

In the parametric case, we are particularly interested in the impact on parameter estimates when the transformations are misspecified. Therefore, we also consider simulations for a misspecified parametric model where a log link is fit to the data, but a square root link should have been used. The event time data were generated through a square transformation in this scenario instead of an anti-log transformation. The other modeling assumptions are the same as in the correctly specified parametric settings.

In the semiparametric case, six different simulation settings were considered. For each setting, 250 simulations were carried out. Two different sample sizes were explored: $n = 100$ and $n = 200$, and three different censoring levels were considered: 0%, 7%, and 17%. The 7% was chosen to mimic the HNC data. For each setting, 100 bootstrap samples were used to determine the bootstrap standard errors. In the parametric case, we considered 50% censoring in addition to the three other censoring levels, giving eight different simulation settings. Simulation results for 7% censoring are presented in Tables I and II. Important simulation findings will be reviewed here, but complete simulation results can be found in the tables and figure included in the Supporting Information. Through the semiparametric simulations, we discovered that there is some numerical instability in the estimation procedure as seen by the fact that convergence of the algorithm is sensitive to the particular data set. Non-convergence is more frequent when the sample size is small and also for larger amounts of censoring. Because potential users should be aware of convergence issues, the percentage of simulations that did not converge for each simulation setting is presented in Table III for the semiparametric setting (note: lack of convergence persisted even with altered initial values in the estimation procedure). For the parametric setting, convergence does not appear to be a concern as demonstrated in Table IV. Of note, the results presented in Table I and Figure 1 are conditional on convergence.

5.1. Sample size

Sample size is an important factor in the performance of the proposed methods. Table I presents semiparametric simulation results for $n = 100$ and $n = 200$, both with 7% censoring. There appears to be some underestimation for the parameters associated with the continuous outcome (α_2 and β_2). Despite this fact, none of the parameters are significantly biased because the empirical confidence intervals do not exclude the truth. Also, point estimation can be improved by increasing the sample size. For both sample sizes included in the simulations, the point estimate for the γ parameter is well estimated. Inference, however, may be somewhat unreliable for a sample as small as 100. For example, numerical instability of the estimation procedure is more of a problem with a small sample size (21.7% of simulations failed to converge), and the 95% coverage probabilities based on the bootstrap standard errors tend to deviate somewhat from the nominal level. Even though the point estimate for γ seems reasonable for $n = 100$, the bootstrap standard error is overestimated, leading to a coverage probability that is too large. On the other hand, inference for $n = 200$ with 7% censoring appears to be reliable as demonstrated by 95% coverage probabilities that are close to the nominal level. The larger sample size of 200 also has the added advantage of better numerical stability, with only 10.4% of simulations failing to converge.

Simulation results suggest that the performance of the correctly specified parametric latent variable transformation model is quite good in a variety of settings, including smaller sample sizes. The bias for the β, α, and γ parameters is consistently small, and the coverage probabilities all tend to be close to the

Table I. Simulation results for 7% censoring.

	β_{01}	β_{02}	β_1	β_2	β_3	α_1	α_2	α_3	γ
True parameter values	0.5	0.5	0.5	0.9	0.75	0.5	0.9	0.75	1
Semiparametric with $n = 100$									
Mean			0.519	0.666	0.682	0.622	0.525	0.574	0.979
Bias			0.019	−0.234	−0.068	0.122	−0.375	−0.176	−0.021
Empirical SE			0.187	0.150	0.162	0.381	0.234	0.256	0.316
Bootstrap SE			0.152	0.151	0.164	0.289	0.210	0.226	0.411
95% CI coverage			0.938	0.928	0.959	0.851	0.877	0.892	0.995
Semiparametric with $n = 200$									
Mean			0.524	0.730	0.694	0.650	0.609	0.595	1.007
Bias			0.024	−0.170	−0.056	0.150	−0.291	−0.155	0.007
Empirical SE			0.120	0.112	0.111	0.324	0.181	0.187	0.294
Bootstrap SE			0.106	0.119	0.118	0.231	0.171	0.176	0.276
95% CI coverage			0.915	0.969	0.964	0.906	0.924	0.942	0.920
Parametric with $n = 100$									
Mean	0.511	0.504	0.502	0.910	0.830	0.496	0.900	0.872	1.014
Bias	0.011	0.004	0.002	0.010	0.080	−0.004	<0.001	0.122	0.014
Empirical SE	0.141	0.183	0.118	0.153	0.766	0.118	0.191	1.432	0.307
Model SE	0.141	0.182	0.114	0.134	0.192	0.123	0.168	0.267	0.297
95% CI coverage	0.952	0.948	0.956	0.919	0.956	0.960	0.927	0.980	0.948
Parametric with $n = 200$									
Mean	0.499	0.513	0.490	0.897	0.772	0.498	0.886	0.770	1.005
Bias	−0.001	0.013	−0.010	−0.003	0.022	−0.002	−0.014	0.020	0.005
Empirical SE	0.097	0.141	0.081	0.085	0.132	0.091	0.117	0.178	0.214
Model SE	0.100	0.130	0.080	0.095	0.125	0.087	0.119	0.167	0.208
95% CI coverage	0.948	0.928	0.952	0.968	0.936	0.940	0.952	0.948	0.944
Parametric with $n = 100$; event time link misspecified									
Mean	−0.650	0.504	0.435	0.899	0.986	0.421	0.875	1.033	1.039
Bias	−1.150	0.004	−0.065	−0.001	0.236	−0.079	−0.025	0.283	0.039
Empirical SE	0.260	0.191	0.211	0.137	1.300	0.284	0.272	1.644	0.397
Model SE	0.258	0.187	0.219	0.135	0.263	0.235	0.202	0.378	0.330
95% CI coverage	0.936	0.949	0.962	0.953	0.941	0.915	0.932	0.958	0.915
Parametric with $n = 200$; event time link misspecified									
Mean	−0.639	0.506	0.416	0.889	1.036	0.422	0.858	1.230	0.973
Bias	−1.139	0.006	−0.084	−0.011	0.286	−0.078	−0.042	0.480	−0.027
Empirical SE	0.176	0.138	0.157	0.095	1.719	0.205	0.294	2.952	0.366
Model SE	0.181	0.133	0.154	0.095	0.186	0.169	0.146	0.258	0.225
95% CI coverage	0.958	0.946	0.942	0.954	0.729	0.929	0.888	0.504	0.921

Table II. Simulation results for 7% censoring—estimates of secondary parameters for parametric models.

	$H_3(c_3(1))$	$H_3(c_3(2))$	$H_3(c_3(3))$	$H_3(c_3(4))$	σ_1	σ_2
Correct parametric model						
$n = 100$	−1.078	0.008	1.080	2.175	0.952	0.879
$n = 200$	−1.040	0.006	1.025	2.054	0.984	0.988
Parametric model with event time link misspecified						
$n = 100$	−1.267	0.015	1.257	2.570	1.910	0.771
$n = 200$	−1.395	0.039	1.415	2.842	1.814	0.732

True thresholds are −1, 0, 1, and 2. True standard deviations are 1.

Table III. Percentage of semiparametric simulations that failed to converge.

	0% censoring	7% censoring	17% censoring
$n = 100$	32.4	21.7	31.7
$n = 200$	2.8	10.4	31.7

Table IV. Percentage of correctly specified parametric simulations that failed to converge.

	0% censoring	7% censoring	17% censoring	50% censoring
$n = 100$	0.4	0.8	0.4	1.2
$n = 200$	0	0	0	0

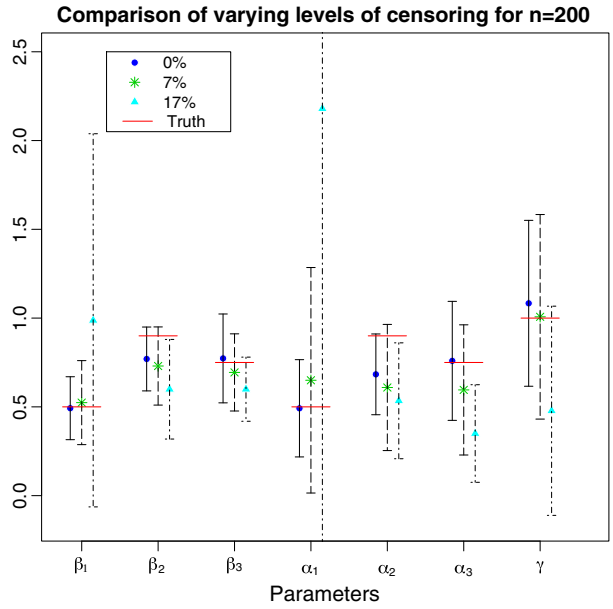

Figure 1. Plots of means and empirical 95% confidence intervals from semiparametric simulations with $n = 200$ and 0%, 7%, and 17% censoring.

nominal level. The model-based standard errors tend to be close to the empirical standard errors, suggesting that inference using these model based standard errors should be reliable. The noticeable exception to this is for the β_3 and α_3 parameters (associated with ordinal outcome). For these parameters, the model-based standard errors are often small relative to the empirical standard errors, and the bias is somewhat larger. These results are driven by one to two simulations with extreme values. When the simulations with these extreme values are removed, the results for the β_3 and α_3 parameters are in line with the results of the other parameters. Exceptionally large parameter estimates should be easily recognizable by an analyst and should be treated as a model that cannot be reliably fit. The benefits of an increased sample size are illustrated in Table II. The transformed thresholds tend to be better estimated in a larger sample, and the standard deviation estimates are improved in a larger sample. A larger sample is, therefore, preferable. However, results in Table I suggest that even with a smaller sample size, inference of the primary parameters (β, α, and γ) is quite reasonable using the parametric approach.

5.2. *Censoring*

Censoring is also an important factor for performance of the proposed methods. Figure 1 considers semiparametric simulation results for 0%, 7%, and 17% censoring for the larger sample size of 200 (complete simulation results for these settings can be found in the Supporting Information). It is clear from the figure that when the censoring reaches the moderate level of 17%, performance of the semiparametric method suffers. The standard errors of the parameters associated with the event time (α_1 and β_1) are huge, and there are parameters that are significantly biased on the basis of the empirical confidence intervals. Even the γ parameter is not well estimated with a larger amount of censoring. But, when there is no censoring, the performance of the semiparametric method is good. Only 2.8% of the simulations did not converge (which is similar to that in [13]), and the coverage probabilities are close to the nominal level. The parametric approach is not nearly as sensitive to the amount of censoring. In particular, a higher censoring percentage does not create estimation difficulties in the parametric setting but rather simply increases standard errors a bit.

5.3. Misspecified transformations

When the event time link is misspecified in the parametric approach, we see that the parameters associated with the event time (β_{01}, β_1, α_1, and σ_1) are biased. This is to be expected, however, because these parameters would now have a different interpretation. The other parameters are still well estimated (with the same caveat about the few simulations with extreme values as in the correctly specified case). In particular, γ still has a fairly small bias and good coverage probability. This suggests that misspecifying the event time link should not have a major impact on the inference for γ. When the sample size is decreased to 100, misspecification of the event time link leads to a bit more instability in the maximum likelihood procedure. The percentage of simulations that cannot be estimated through maximum likelihood, however, is still under 8%. When both the event time and continuous links are misspecified, results are not as promising. For example, we considered the case when the event time link is misspecified in the same way as before and the continuous measurable outcome is generated using an anti-log transformation, but the identity link is fit to the data. Results for this setting are not shown, but it is important to note that this amount of misspecification leads to a fairly unstable maximum likelihood procedure (maximum likelihood estimation failed for as many as 30% of simulations) and leads to substantial bias in all of the parameters.

5.4. Conclusions

Simulation results suggest that the semiparametric methodology can be useful when there is a larger sample size with a small amount of censoring. However, care should be taken when the percentage of censoring is high or when the sample size is small because of convergence issues. On the other hand, the parametric approach has good performance in a range of settings, including smaller sample sizes and larger amounts of censoring. Misspecification of transformations can be an issue in the parametric setting, however. Some misspecification may not be a problem (i.e., log instead of square root transformation for the event time), but major misspecification can lead to substantial bias in all of the parameters.

Furthermore, there is a trade-off between improved efficiency and avoiding misspecificiation when choosing between the semiparametric and parametric models. A correctly specified parametric model has better efficiency as seen by shorter confidence intervals (Table I). The parametric model is also substantially faster computationally, which is a huge practical advantage. However, a severely misspecified parametric model can lead to unreliable estimates. The semiparametric approach eliminates the need to pre-specify transformations, which helps to avoid the misspecification problem. However, with this added flexibility comes reduced efficiency, increased computational burden from the two-stage estimation procedure, and potential numerical instability. These trade-offs need to be carefully considered in practice. One potential way to utilize the advantages of both models is to use the two approaches together, using the semiparametric procedure to inform a parametric model.

Both the semiparametric and parametric procedures described in this paper have been implemented in R. Code is available upon request to the corresponding author.

6. Measuring dysphagia in head and neck cancer patients

We applied the proposed methods to a study of HNC patients carried out at DFCI [3]. Patients were identified for the study through a retrospective chart review and were eligible if they were diagnosed between 1998 and 2008 with an advanced-stage squamous cell carcinoma of the oropharynx, hypopharynx, larynx, or unknown primary and were treated with chemoradiotherapy (chemoRT) and neck dissection. Twenty-four months of follow-up after chemoRT without recurrence was also required for inclusion.

Squamous cell carcinoma of the head and neck represents about 5% of newly diagnosed cancers in adults in the USA. These patients tend to present with locally advanced disease and are treated aggressively with some combination of surgery, chemotherapy, and radiotherapy [20]. Intensive chemoRT regimens have been found to be effective in the management of HNC in terms of improving both progression-free and overall survival [21]. With aggressive chemoRT treatment, however, come side effects such as dysphagia, or difficulty swallowing, that have a negative impact on a patient's quality of life [22, 23]. The goal of the study, therefore, is to determine clinical and treatment factors associated with dysphagia in this group of HNC patients. This goal is challenging, however, because there is not one definitive way to measure or define dysphagia objectively. In order to best capture dysphagia, the use of multiple objective measures has been suggested [3, 4]. The investigators at Dana-Farber collected

information on several measurable outcomes that are often used to describe dysphagia: time from end of chemoRT to removal of the gastrostomy tube, weight loss after chemoRT, and diet (liquid, soft, etc.). In general, how long a gastrostomy tube is in place reflects the severity of dysphagia [24]. Weight loss and limitations of diet have also been identified to be associated with dysphagia in HNC patients [25–28]. Using the proposed methodology for the analysis will allow us to combine the multiple measurable outcomes through the latent variable structure and then explore factors associated with the latent variable (dysphagia) as desired by the investigators.

Eighty-eight patients were eligible for the study. Two patients were excluded from the analysis because they never had a gastrostomy tube and thus represent a different patient population. Sixty-six patients were then available with complete outcome information. We were able to impute weight information using a later weight measurement (measurement taken at some point after our baseline of 1 month post-chemoRT) for nine patients, giving us 75 patients available for analysis. For each patient i, let Y_{i1} be the observed time from end of chemoRT to removal of the gastrostomy tube in days (note this outcome is potentially censored), Y_{i2} be weight loss after chemoRT in kilograms, and Y_{i3} be diet (regular, soft, pureed, liquid, and no food; ordinals 1–5). For identifiability, α_3 will be constrained to be greater than 0. Using the proposed methodology, e_i characterizes the level of dysphagia for patient i with a larger e_i indicating worse dysphagia, and Z_i is treatment or some other clinical factor potentially associated with dysphagia. In this way, γ is the parameter of primary interest for inference.

Because of the small sample size of the HNC study ($n = 75$), we are unable to fit complex models with many covariates. We have therefore decided to focus on two models that are small but clinically interesting for both the semiparametric and parametric analyses. Model 1 will include T-stage (ordinal) as the Z covariate and sex as the X covariate, and model 2 will include treatment (induction vs. concurrent chemoRT) as the Z covariate and sex as the X covariate. T-stage is clinically relevant because T-stage has been shown to be associated with adverse swallowing outcomes previously [3, 23, 29]. Treatment is of interest to determine if patients treated with the more aggressive induction chemotherapy followed by chemoRT have worse dysphagia as compared with patients treated with primary concurrent chemoRT. Sex is included as the X covariate in both models because it is not a variable of primary interest, but we might want to control for it in the analysis.

We start by using the semiparametric approach, with results for models 1 and 2 displayed in Table V. For both models, none of the β parameters are significant as evidenced by 95% confidence intervals that cover 0. This suggests that sex is not associated with any of the transformed outcomes included in the model. Also, for both models, α_1 and α_3 are significant, but α_2 is not. The α parameters are factor loadings, so these findings indicate that time on the gastrostomy tube and diet are significantly associated with the latent variable (dysphagia). Worse dysphagia is associated with a longer time on the feeding tube and a more modified diet. Weight loss after chemoRT does not appear to be significantly related to dysphagia using the semiparametric approach. This is not a particularly surprising result as clinically we know that

Table V. Head and neck semiparametric data analysis results.

	Estimate	SE	95% CI
Model 1 (T-stage)			
β_1	0.154	0.413	(−0.655, 0.964)
β_2	−0.582	0.301	(−1.171, 0.008)
β_3	−0.220	0.466	(−1.135, 0.694)
α_1	0.577	0.288	(0.012, 1.141)
α_2	0.017	0.127	(−0.233, 0.266)
α_3	1.518	0.614	(0.315, 2.721)
γ	−0.027	0.356	(−0.724, 0.670)
Model 2 (treatment)			
β_1	0.167	0.423	(−0.663, 0.997)
β_2	−0.581	0.367	(−1.301, 0.138)
β_3	−0.156	0.670	(−1.469, 1.158)
α_1	0.528	0.146	(0.242, 0.815)
α_2	0.019	0.094	(−0.166, 0.204)
α_3	1.341	0.304	(0.746, 1.936)
γ	0.231	0.401	(−0.554, 1.016)

For both models, sex is the X covariate.

weight loss may not be a great measure of dysphagia. On the one hand, it makes sense that if a patient is having difficulty swallowing, then that patient is likely to eat less and lose more weight. However, when the feeding tube is being used, adequate nutrition can be obtained through the tube without the need to swallow. This would suggest that even if swallowing is difficult for a patient, it may not be seen through measuring the weight of that patient. Both semiparametric models 1 and 2 indicate that weight loss may not be a useful measure to capture dysphagia.

In semiparametric model 1, T-stage is included as the Z covariate. The γ parameter associated with T-stage in this model is not significant. This suggests that a higher T-stage (increased size of the primary tumor) is not associated with worse dysphagia. Similarly, from semiparametric model 2, there is no evidence of a significant association between treatment and dysphagia, meaning that there is not evidence that induction chemotherapy is associated with worse dysphagia. Although neither of the Z covariates considered were found to be significant, it is important to keep in mind the limitations of the proposed semiparametric methodology when interpreting these results. The sample size for the HNC data is 75, and there is 6.7% censoring. We know that inference may not be totally reliable in this setting. In particular, the standard error estimate for the γ parameter may be too large. Because of the small sample size, it is not clear how meaningful these semiparametric results would be in the clinical literature. However, the semiparametric approach could be helpful in informing a parametric model. We can use the estimated transformations from the semiparametric approach to decide on appropriate link functions for a parametric model. Figure 2 presents the estimated transformations obtained from model 1. The estimated transformations for model 2 are very similar. On the basis of this plot, it would seem reasonable to use either a log or square root transformation for the time on the gastrostomy tube. Using the identity link (i.e., assuming normality) for weight loss also seems appropriate.

We now consider the parametric approach, using the results from the semiparametric analysis to inform transformations. Inference using the parametric approach should be more reliable given the small sample size. Utilizing residual plots, we determined that using a log link for the time on the feeding tube and an identity link for weight loss after chemoRT was reasonable. Results for the parametric models 1 and 2 are presented in Table VI. As in the semiparametric case, for both models, none of the β parameters are significant. This further suggests that sex is not associated with any of the transformed outcomes. For parametric model 1, α_1 and α_3 are significant but α_2 is not. This is consistent with the semiparametric findings, suggesting that longer time on the feeding tube is associated with worse dysphagia and a more modified diet is associated with worse dysphagia, but weight loss after chemoRT is not. In parametric model 2, only α_1 is significant. However, the large parameter and standard error estimates for model 2 suggest that this model may not be very reliable. In parametric model 1, γ captures the relationship between T-stage and dysphagia. This parameter is significant, and results suggest a higher T-stage is associated with worse dysphagia. The γ parameter is not significant in parametric model 2, suggesting that having induction chemotherapy is not associated with worse dysphagia.

Through the parametric modeling procedure, we used residual plots to consider model fit. We considered fitting both parametric models assuming a log link and assuming a square root link for the time on

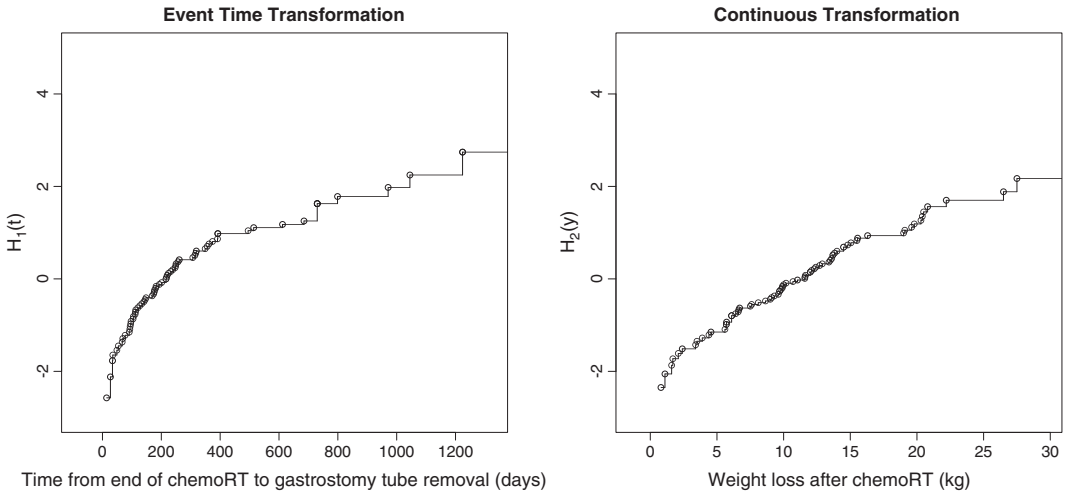

Figure 2. Estimated transformations for semiparametric model 1 (T-stage).

Table VI. Head and neck parametric data analysis results.

	Estimate	SE	95% CI
Model 1 (T-stage)			
β_1	0.221	0.309	(−0.384, 0.827)
β_2	−3.150	1.826	(−6.729, 0.429)
β_3	−0.133	0.450	(−1.015, 0.750)
α_1	0.910	0.077	(0.759, 1.062)
α_2	0.218	0.682	(−1.118, 1.555)
α_3	0.780	0.161	(0.464, 1.096)
γ	0.274	0.105	(0.069, 0.479)
Model 2 (treatment)			
β_1	0.201	1.932	(−3.586, 3.988)
β_2	−3.18	7.128	(−17.156, 10.787)
β_3	−1.145	48.638	(−96.475, 94.186)
α_1	0.590	0.106	(0.382, 0.797)
α_2	−0.417	1.795	(−3.934, 3.101)
α_3	13.120	21.923	(−29.849, 50.087)
γ	0.441	2.175	(−3.823, 4.704)

For both models, sex is the X covariate. Intercept terms were included for Y_{i1} and Y_{i2}.

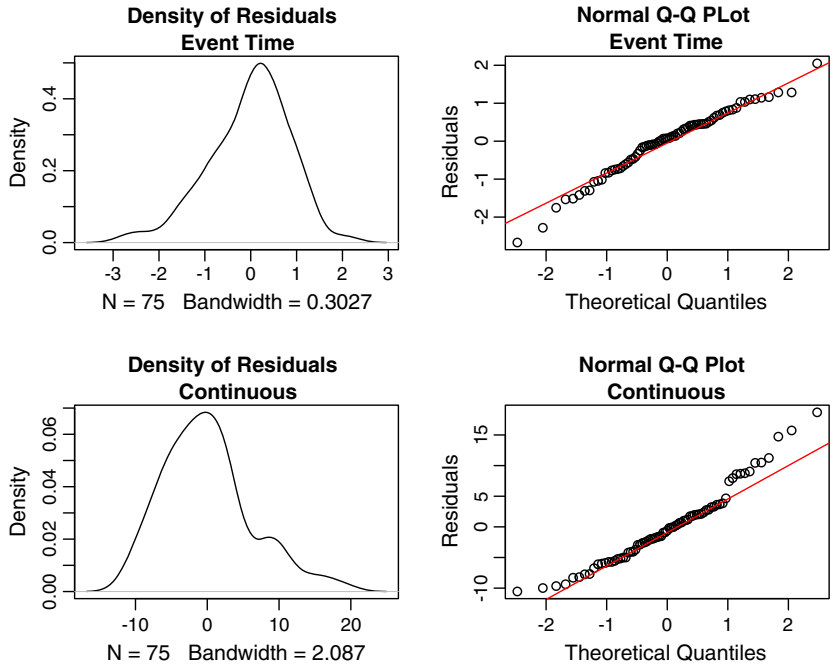

Figure 3. Plot looking at the normality of the residuals for parametric model 1 (T-stage).

the feeding tube. For model 2, the square root link could not be fit (failed maximum likelihood procedure). Residual plots for model 1 suggested that the log link was a better fit for this model, although this choice did not influence the γ parameter estimate much and did not alter any of the conclusions. Density and Q-Q plots for both the event time and continuous outcomes are shown in Figure 3 for parametric model 1. These plots suggest that model 1 fits quite well, as there are no major departures from normality.

This analysis illustrates how both the semiparametric and parametric approaches can be utilized to analyze real-world data. However, it is important to recognize that we are limited in the conclusions that can be drawn by the small sample size. We know that inference may be unreliable because of the small sample size in the semiparametric case. Model 2 in the parametric setting is also questionable owing to the large parameter and standard error estimates (particularly for the parameters associated with the ordinal outcome, diet). However, the semiparametric models and parametric model including T-stage are consistent in terms of the conclusions for the α and β parameters. Results from the analyses suggest that

time on the feeding tube and diet both are important measures of dysphagia, but that weight loss after treatment is not. Sex of the patient also does not seem to be particularly relevant. The parametric analysis further suggests that T-stage is associated with dysphagia. This is inconsistent with the semiparametric analysis, but the parametric approach is more reliable given the sample size constraints. Therefore, there is a suggestion that a higher T-stage is associated with worse dysphagia, although larger study would be helpful to further explore this relationship.

7. Discussion

We have proposed two classes of models for dealing with complicated multiple outcomes data. Both approaches allow for multiple outcomes of mixed types, including censored outcomes, to be incorporated into a latent variable framework and allow for estimation of a treatment (or other covariate) effect on the unobserved latent variable. The two approaches could be used as separate modeling techniques. Or, as we have demonstrated, the two methods could also be used together as a comprehensive modeling strategy with the semiparametric analysis being used to inform a parametric analysis.

The semiparametric approach has much appeal because it does not require pre-specifying a link between the measurable outcomes and the latent variable of interest. Simulations suggest that the proposed method has much utility when the sample size is large and the censoring proportion is small. More specifically, the method performs well for a sample size of 200 with 7% censoring or less. However, when the censoring percentage reaches 17%, simulations indicate inference may not reliable. Censoring adds additional incomplete information to an already complex framework, so the sensitivity to censoring is not surprising, especially when the sample size is small. Similarly, a sample size as small as 100 may be too small to completely trust inference made using the semiparametric approach. A larger sample size is likely required because of the large number of parameters that must be estimated, either in the transformation functions or through the likelihood. In the end, a small sample likely does not contain enough information to reliably estimate all of the pieces. Despite these limitations, a sample size of 200 or more is not unreasonable in many potential applications.

The performance of the parametric latent variable transformation model is quite good, even when you have a fairly small sample size or a large amount of censoring, unless you have substantial model misspecification. A limitation of the parametric approach is the need to pre-specify link functions. However, we have suggested the potential use of the semiparametric methodology to suggest appropriate transformations for a parametric model and have demonstrated the use of residuals to diagnose incorrect transformation functions. These tools should make the parametric approach a reliable means of analysis. In light of the strengths and limitations of each approach, we would recommend using the semiparametric approach when the sample size is large and the censoring percent is low. The semiparametric approach could also be particularly useful when, for example, the time-to-event outcome does not appear to have any straightforward transformation. The parametric method would be more appropriate when you have a smaller sample or there is a larger amount of censoring.

A potential limitation of both the parametric and semiparametric methods is the use of a threshold model for the ordinal measurable outcomes. The threshold model is practical because then we have all continuous outcomes that can be jointly modeled using the multivariate normal. However, this means that nominal measurable outcomes cannot be meaningfully incorporated in this approach. Also, in order to use the threshold approach, it must be plausible that there is some underlying continuous quantity that gives rise to the ordinal categories that are observed. In the HNC case, the ordinal outcome is diet, and it is plausible that there is some underlying biological quantity that determines the change from one food type to another. However, it is possible the threshold approach would not make sense in a different application. Choosing which covariates should be X covariates and which should be Z covariates is also a consideration. In order to look at the association between a covariate and the latent variable, the covariate must be included in Z. If you want to control for a variable but are less interested in its relationship with the latent variable, it can be included in X. This is a subject-matter, rather than statistical, decision.

In exploring dysphagia, specifically, we were able to use both approaches to analyze the data, but we are limited by the small sample size and retrospective design of the DFCI study. Even using the parametric approach, the small sample size is a limitation. Specifically, we must keep in mind that not only do the α, β, and γ parameters have to be estimated, but so do the transformed thresholds and standard deviation parameters. When the models can be fit using maximum likelihood, the performance is pretty good, even with a sample as small as 75 as in the DFCI data. However, we are limited by the sample size in the number of covariates that can be included. Also, there are instances when models of interest

simply cannot be fit because the maximum likelihood procedure fails or models that are not reliable (e.g., parametric model 2). In exploring the DFCI data, we did run into models that simply could not be fit and/or covariates that could not be included because of too few people in each category. For example, DFCI investigators were interested in whether type of radiation is related to dysphagia. However, there were simply not enough patients in each of the different radiation categories to be able to consider this variable in the model. Because the study was retrospective, we also did not have the advantage of the treatment assignment being randomized and were missing covariates such as smoking status and alcohol use that may be particularly relevant to head and neck data.

Despite these limitations, we were able to use a small data set to learn something about dysphagia through the latent variable approach. Results suggest that having a higher T-stage is associated with worse dysphagia. Also, we have proposed a comprehensive modeling strategy that can be useful in other settings and could be used to further investigate dysphagia once more data are collected.

Acknowledgements

The paper originates from the Harvard dissertation of the first author. We thank Dr Huazhen Lin for sharing an earlier draft of her manuscript [13], which motivated the proposed work. We also thank Dr Laura Goguen and Dr Claudia Chapuy for access to the DFCI head and neck data used in the analysis. Finally, we thank the Editor, Associate Editor and two referees for many insightful suggestions that greatly improved the quality of the paper.

References

1. Dunson DB. Bayesian dynamic modeling of latent trait distributions. *Biostatistics* 2006; **7**(4):551–568.
2. Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. *Biometrics* 1987; **43**(3):487–498.
3. Chapuy CI, Annino DJ, Snavely A, Li Y, Tishler RB, Norris CM, Haddad RI, Goguen LA. Swallowing function following postchemoradiotherapy neck dissection. *Otolaryngology—Head and Neck Surgery* 2011; **145**(3):428–434.
4. Caudell JJ, Schaner PE, Meredith RF, Locher JL, Nabell LM, Carroll WR, Magnuson JS, Spencer SA, Bonner JA. Factors associated with long-term dysphagia after definitive radiotherapy for locally advanced head-and-neck cancer. *International Journal of Radiation Oncology*Biology*Physics* 2009; **73**(2):410–415.
5. Sammel MD, Ryan LM. Latent variable models with fixed effects. *Biometrics* 1996; **52**(2):650–663.
6. Roy J, Lin X. Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics* 2000; **56**(4): 1047–1054.
7. Catalano PJ, Ryan LM. Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association* 1992; **87**(419):651–658.
8. Fitzmaurice GM, Laird NM. Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association* 1995; **90**(431):845–852.
9. Sammel MD, Ryan LM, Legler JM. Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society. Series B (Methodological)* 1997; **59**(3):667–678.
10. Regan MM, Catalano PJ. Likelihood models for clustered binary and continuous outcomes: application to developmental toxicology. *Biometrics* 1999; **55**(3):760–768.
11. Moustaki I, Knott M. Generalized latent trait models. *Psychometrika* 2000; **65**(3):391–411.
12. Huber P, Ronchetti E, Victoria-Feser M-P. Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 2004; **66**(4):893–908.
13. Lin H, Zhou L, Elashof RM, Li Y. Semiparametric latent variable transformation models for multiple mixed outcomes. *Statistica Sinica* 2013; **24**:833–854.
14. Muthén B. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* 1984; **49**(1):115–132.
15. Dunson DB. Dynamic latent trait models for multidimensional longitudinal data. *Journal of the American Statistical Association* 2003; **98**(463):555–563.
16. Chen K, Jin Z, Ying Z. Semiparametric analysis of transformation models with censored data. *Biometrika* 2002; **89**(3): 659–668.
17. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. *Numerical Recipes in C: The Art of Scientific Computing* Second. Cambridge University Press: Cambridge, 1992.
18. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. Chapman & Hall: New york, 1993.
19. Kutner MH, Nachtsheim CJ, Neter J, Li W (eds). *Applied Linear Statistical Models* 5th ed. McGraw-Hill: New york, 2004.
20. Posner MR, Hershock DM, Blajman CR, Mickiewicz E, Winquist E, Gorbounova V, Tjulandin S, Shin DM, Cullen K, Ervin TJ, Murphy BA, Raez LE, Cohen RB, Spaulding M, Tishler RB, Roth B, Viroglio RdelC, Venkatesan V, Romanov I, Agarwala S, Harter KW, Dugan M, Cmelak A, Markoe AM, Read PW, Steinbrenner L, Colevas AD, Norris CM, Haddad RI. Cisplatin and fluorouracil alone or with docetaxel in head and neck cancer. *New England Journal of Medicine* 2007; **357**(17):1705–1715.

21. Pignon JP, Bourhis J, Domenge C, Designé L. Chemotherapy added to locoregional treatment for head and neck squamous-cell carcinoma: three meta-analyses of updated individual data. *The Lancet* 2000; **355**(9208):949– 955.
22. Goguen LA, Posner MR, Norris CM, Tishler RB, Wirth LJ, Annino DJ, Gagne A, Sullivan CA, Sammartino DE, Haddad RI. Dysphagia after sequential chemoradiation therapy for advanced head and neck cancer. *Otolaryngology—Head and Neck Surgery* 2006; **134**(6):916– 922.
23. Nguyen NP, Frank C, Moltz CC, Vos P, Smith HJ, Nguyen PD, Martinez T, Karlsson U, Dutta S, Lemanski C, Nguyen LM, Sallah S. Analysis of factors influencing aspiration risk following chemoradiation for oropharyngeal cancer. *British Journal of Radiology* 2009; **82**(980):675– 680.
24. Cowen M, Simpson S, Vettese T. Survival estimates for patients with abnormal swallowing studies. *Journal of General Internal Medicine* 1997; **12**(2):88– 94.
25. Nguyen NP, Moltz CC, Frank C, Vos P, Smith HJ, Karlsson U, Dutta S, Midyett FA, Barloon J, Sallah S. Dysphagia following chemoradiation for locally advanced head and neck cancer. *Annals of Oncology* 2004; **15**(3):383–388.
26. Fessler T. Enteral nutrition for patients with head and neck cancer. *Today's Dietitian* 2008; **10**(6):46.
27. van den Berg MGA, Rütten H, Rasmussen-Conrad EL, Knuijt S, Takes RP, van Herpen CML, Wanten GJA, Kaanders JHAM, Merkx MAW. Nutritional status, food intake, and dysphagia in long-term survivors with head and neck cancer treated with chemoradiotherapy: a cross-sectional study. *Head & Neck* 2013; **36**:60– 65.
28. Denaro N, Merlano M, Russi E. Dysphagia in head and neck cancer patients: pretreatment evaluation, predictive factors, and assessment during radio-chemotherapy, recommendations. *Clinical and Experimental Otorhinolaryngology* 2013; **6**(3): 117–126.
29. Machtay M, Moughan J, Trotti A, Garden AS, Weber RS, Cooper JS, Forastiere A, Ang KK. Factors associated with severe late toxicity after concurrent chemoradiation for locally advanced head and neck cancer: an RTOG analysis. *Journal of Clinical Oncology* 2008; **26**(21):3582–3589.