Analyzing clustered continuous response variables with ordinal regression models

Yuqi Tian¹ ☑ | Bryan E. Shepherd¹ ☑ | Chun Li² ☑ | Donglin Zeng³ ☑ | Jonathan S. Schildcrout¹ ☑

¹Department of Biostatistics, Vanderbilt University, Nashville, Tennessee, USA
²Department of Population and Public Health Sciences, University of Southern California, Los Angeles, California, USA
³Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Correspondence
Jonathan S. Schildcrout, Department of Biostatistics, Vanderbilt University, Nashville, TN 37203, USA. Email: jonny.schild@vumc.org

Funding information
National Institute of Allergy and Infectious Diseases, Grant/Award Numbers: P30 AI110527, R01 AI093234; National Institutes of Health, Grant/Award Number: K23 AI120875; National Heart, Lung, and Blood Institute, Grant/Award Numbers: R01 HL072966, R01 HL094786

Abstract
Continuous response data are regularly transformed to meet regression modeling assumptions. However, approaches taken to identify the appropriate transformation can be ad hoc and can increase model uncertainty. Further, the resulting transformations often vary across studies leading to difficulties with synthesizing and interpreting results. When a continuous response variable is measured repeatedly within individuals or when continuous responses arise from clusters, analyses have the additional challenge caused by within-individual or within-cluster correlations. We extend a widely used ordinal regression model, the cumulative probability model (CPM), to fit clustered, continuous response data using generalized estimating equations for ordinal responses. With the proposed approach, estimates of marginal model parameters, cumulative distribution functions, expectations, and quantiles conditional on covariates can be obtained without pretransformation of the response data. While computational challenges arise with large numbers of distinct values of the continuous response variable, we propose feasible and computationally efficient approaches to fit CPMs under commonly used working correlation structures. We study finite sample operating characteristics of the estimators via simulation and illustrate their implementation with two data examples. One studies predictors of CD4:CD8 ratios in a cohort living with HIV, and the other investigates the association of a single nucleotide polymorphism and lung function decline in a cohort with early chronic obstructive pulmonary disease.

KEYWORDS
clustered data, cumulative probability model, generalized estimating equation, longitudinal data, ordinal regression model

1 | INTRODUCTION

Analyses of quantitative response variables are often challenged by distributions that do not follow standard parametric assumptions. While it is common in such settings to transform the response variables to satisfy model assumptions, such transformations are often ad hoc and model parameters can be difficult to interpret on their natural, untransformed scale.

For example, researchers are interested in studying factors associated with longitudinal measures of the CD4:CD8 ratio among people with HIV on antiretroviral therapy.
The CD4:CD8 ratio is the ratio of CD4 lymphocyte count (cells/mm$^3$) to CD8 lymphocyte count (cells/mm$^3$), and a low CD4:CD8 ratio has been associated with immune senescence, inflammation, and comorbidities for people with HIV (Castilho et al., 2016). The CD4:CD8 ratio tends to be right-skewed (Web Appendix A), so it is often transformed prior to analysis; however, there is no standard accepted transformation, and results may be sensitive to the choice of transformation. Researchers have analyzed the CD4:CD8 ratio with no transformation (Castilho et al., 2016), log-transformation (Sauter et al., 2016), square-root transformation (da Silva et al., 2018), fifth-root transformation (Gras et al., 2019), and various categorizations (Petoumenos et al., 2017).

As a second example, the Lung Health Study was a randomized clinical trial of smokers with early chronic obstructive pulmonary disease (COPD) to examine whether a smoking cessation intervention and inhaled bronchodilator use could slow lung function decline (Anthonisen et al., 1994). Lung function was measured with forced expiratory volume in the first second of exhalation (FEV1), and thresholds defined by FEV1 are commonly used to define the severity of chronic obstructive pulmonary disease (Global Initiative for Chronic Obstructive Lung Disease, 2021). An analysis that fits a binary regression model to a dichotomized lung function measure loses a substantial amount of information compared to an analysis that treats lung function (e.g., FEV1) on its natural, continuous scale. However, with standard regression analyses of continuous scales, parametric assumptions on the distribution are required to capture important quantities, such as exceedance probabilities.

A compelling approach to tackle the challenges associated with nonstandard response distribution modeling is to treat continuous response variables as if they were ordinal using cumulative probability models (CPMs), also known as cumulative link models (Agresti, 2010; Liu et al., 2017). The CPM is a semiparametric linear transformation model (Zeng & Lin, 2006) that assumes a linear model following an unspecified response transformation. Rather than making an assumption about the appropriate transformation to apply, CPM fitting uses the data to estimate the transformation nonparametrically with a step function. The CPM is invariant to any monotonic transformation of the response variable because only order information is used for regression parameter estimation. Therefore, no pretransformation of the response variable is needed. A nice feature of CPMs is that regression parameters from CPMs are interpretable (based on the chosen link function), and because the cumulative distribution function (CDF) is modeled, conditional (on covariates) means, quantiles, and exceedance probabilities can be extracted from the CPM fit. The use of CPMs for cross-sectional continuous response variables, even with thousands of unique outcomes, is computationally feasible with applications of sparse matrix calculations, and it has been implemented in the 	exttt{rms} R package (Harrell, 2020).

Although CPMs have been shown to be an excellent tool for the analysis of cross-sectional continuous response data, they are not directly applicable to the CD4:CD8 study or the Lung Health Study because these studies include repeated measures on individuals. Clustered continuous data are common in practice and important for studying exposure–response associations. Generalized estimating equations (GEE; Liang & Zeger, 1986; Zeger & Liang, 1986) extend quasi-likelihood estimation (Wedderburn, 1974) for generalized linear models (McCullagh & Nelder, 1983), from independent to correlated data settings. However, even though valid inferences are possible with GEE when second and higher order moments are misspecified, GEE for correlated data is challenged by nonstandard distributions in the same way linear regression is for cross-sectional response data.

Extending Liu et al. (2017) that discussed CPMs for scalar response data, we present CPMs for cluster-correlated continuous response variables. Specifically, we demonstrate that (1) CPMs can be fit to quantitative correlated data using GEE methods for ordinal data and (2) GEE for ordinal data can be applied to nonstandard, quantitative response distributions. Our proposed approach estimates time- and covariate-dependent CDFs, from which estimates of the mean, quantiles, and exceedance probabilities can be derived. In addition, we present software and strategies for implementing GEE methods for ordinal data to settings with large numbers (i.e., hundreds or thousands) of distinct levels.

## 2 REVIEW OF EXISTING METHODS

The CPM is a class of models for scalar ordinal response data (Liu et al., 2017). Let $Y$ be a continuous response variable, and $Y^* = h(Y)$ be a transformation of $Y$ with $h(\cdot)$ an unspecified nondecreasing function. Let $X$ be a vector of covariates with $X = 0$ as a reference value. Let $\epsilon$ be an error term. We assume the relationship between the transformed variable and covariates is linear, $Y^* = \beta^T X + \epsilon$, where $\epsilon$ follows a known distribution $F_\epsilon$ and $\beta$ is a vector of regression parameters. It follows that

$$Y = h^{-1}(Y^*) = h^{-1}(\beta^T X + \epsilon). \quad (1)$$

Letting $G = F_{\epsilon}^{-1}$ be a link function, (1) can be expressed as a CPM with $F(y|X) = P(Y \leq y|X) = P(\epsilon \leq h(y) - \beta^T X | X) = F_\epsilon(h(y) - \beta^T X)$, which implies $G(F(y|X)) = h(y) - \beta^T X$. The intercept $h(y) = G(F(y|X = 0))$
represents the link-transformed CDF for $X = 0$ (i.e., the reference CDF), and $\beta^T X$ represents shifts in this CDF that depend on the values of $X$. The interpretation of $\beta$ depends on the choice of the link function. For example, $\beta$ is interpreted as a log odds ratio with the logit link (i.e., $F_\epsilon$ logistic distribution) and a log hazard ratio with the complementary log-log link (i.e., $F_\epsilon$ extreme value distribution).

Assume there are $N$ subjects and denote $y_{ij}$ to be the $j$th smallest observed response value ($j = 1, ..., J$). Rather than specifying a functional form for $h(\cdot)$, we can estimate it using a step function with $h_j = h(y_{ij})$. Since $h(\cdot)$ is estimated nonparametrically, CPMs belong to the class of semiparametric linear transformation models (Zeng & Lin, 2006). For $(y_i, X_i), i \in 1, ..., N$ with $y_i \equiv y_j$ to the $j$th-ordered value, $y_{ij}$, the CPM is given by

$$G(F(y_i | X_i)) = G(F(y_{ij} | X_i)) = y_j - \beta^T X_i. \quad (2)$$

Letting $\theta = (\gamma^T, \beta^T)^T$ and $\gamma = (\gamma_1, ..., \gamma_{J-1})^T$, the likelihood for $\theta$ is

$$L(\theta) = \prod_{j=1}^{J} \prod_{y_{ij} \in y_{j}} \{ F(y_i | X_i) - F(y_{ij}^{-} | X_i) \}, \quad (3)$$

where $F(y_{ij}^{-} | X_i) = \lim_{t \downarrow y_{ij}} F(t | X_i)$. A nonparametric likelihood can be obtained by substituting $F(y_{j-1}^{-} | X_i)$ for $F(y_{ij}^{-} | X_i)$ as follows:

$$L(\theta) = \prod_{j=1}^{J} \prod_{y_{ij} \in y_{j}} \{ G^{-1}(\gamma_j - \beta^T X_i) - G^{-1}(\gamma_{j-1} - \beta^T X_i) \}, \quad (4)$$

where $-\infty \equiv y_0 < \gamma_1 < \cdots < \gamma_{J-1} < \gamma_J \equiv \infty$. From this likelihood, nonparametric maximum likelihood estimates (NPMLEs) of $\theta$ can be obtained.

The CPM in (2) is identical to the cumulative link model used for ordinal data. For example, CPMs with the logit link are referred to as proportional odds models for ordinal data (McCullagh, 1980). It follows that a semiparametric linear transformation model can be fitted with an ordinal CPM where each distinct value of the continuous $Y$ is treated as its own ordinal category. There will be $N$ categories for truly continuous $Y$. CPMs model the continuous response variable CDF as a linear function of covariates after an unspecified, monotonic transformation has been applied. The transformation is estimated nonparametrically from the observed data with a step function.

CPMs have several attractive properties for fitting continuous response data (Liu et al., 2017; Tian et al., 2020). Since only ordinal information is used for estimating $\beta$, CPMs are invariant to any monotonic transformation of the response, which means no transformation is needed. They have been shown to work well with continuous response variables subject to detection limits (Tian et al., 2022), and, under some mild conditions, CPMs yield estimates that are consistent and asymptotically normal (Li, Tian, et al., 2022) with variances that can be estimated with the inverse of the observed information matrix. Other quantities, such as quantiles, exceedance probabilities, and expectations conditional on covariates can be derived from the CPM model fit. For example, the expectation can be estimated with $\hat{E}(Y | X) = \sum_{j=1}^{J} \sum_{y_{ij} \in y_{j}} y_{ij} \{ F(y_{ij} | X) - \hat{F}(y_{j-1}^{-} | X) \}$. Standard errors for CDFs and expectations can be calculated using the delta method, and quantiles can be estimated with linear interpolations of the inverse of the CDFs (Liu et al., 2017; Tian et al., 2022).

Until recently, use of CPMs for continuous responses was rare due, in part, to computational costs. The orx() function in the rms package in R is a computationally efficient implementation of CPMs that can be fit with tens of thousands of distinct responses. It takes advantage of the sparse structure of the Hessian matrix, which permits efficient inversion by Cholesky decomposition in a Newton–Raphson algorithm (Harrell, 2020; Liu et al., 2017).

### 3 METHODS

#### 3.1 CPMs for clustered continuous response variables

We extend CPMs to the cluster-correlated response setting for the same reason they were developed in the cross-sectional response setting; namely, we would like to avoid parametric and often ad hoc transformations of the response to satisfy modeling assumptions. Although our notation throughout suggests we are in a setting with longitudinal data, our methodology applies to more general clustered data.

Let $N$ be the number of subjects in a sample with $i \in \{1, ..., N\}$ indexing subjects (clusters), and $M_i$ is the number of observations for subject $i$. Denote the response for subject $i$ at time $t$ with $Y_{it}$, and $Y_i = (Y_{i1}, ..., Y_{iM_i})^T$. Across all subjects, $Y = (Y_1, ..., Y_N)^T$ has a total of $J$ distinct values; with truly continuous $Y$, $J = \sum_{i=1}^{N} M_i$. Let $Z_{it,j} = I(Y_{it} \leq y_{ij})$ and $\mu_{it,j} = E(Z_{it,j} | X_{it}) = F(Y_{it} \leq y_{ij} | X_{it})$, where $y_{ij}$ corresponds to the $j$th smallest value among the $J$ levels of the response.
variable, and $X_{it}$ is the design vector (without an intercept) for subject $i$ at time $t$. Let the vector of binary indicator variables for subject $i$ at time $t$ be $Z_{it} = (Z_{it1}, \ldots, Z_{itM})^T$, and $\mu_{it} = (\mu_{it1}, \ldots, \mu_{itM})^T$. Finally, for subject $i$, let $Z_i = (Z_{i1}^T, \ldots, Z_{iM}^T)^T$ and $\mu_i = (\mu_{i1}, \ldots, \mu_{iM})^T$.

Suppose $Y_{it}$ has a linear relationship with the covariates $X_{it}$ after an unspecified monotonic transformation $h(\cdot)$ has been applied. This leads to a linear transformation model

$$Y_{it} = h^{-1}(Y_{it}') = h^{-1}(\beta^T X_{it} + \epsilon_{it}),$$

where $\epsilon_{it}$ follows a specified distribution and $\epsilon_{it}$ is independent of $\epsilon_{i't'}$ for $i \neq i'$, but not if $i = i'$. By linear transformation models, we have $\mu_{it,j} = P(Y_{it} \leq y_{(j)} | X_{it}) = P(\epsilon_{it} \leq h(y_{(j)}) - \beta^T X_{it}) = F_{\epsilon}(h(y_{(j)}) - \beta^T X_{it})$, which implies $F_{\epsilon}^{-1}(\mu_{it,j}) = h(y_{(j)}) - \beta^T X_{it}$. Then the CPM for a clustered continuous response variable is

$$G(\mu_{it,j}) = \gamma_j - \beta^T X_{it},$$

where $G = F_{\epsilon}^{-1}$ is the specified link function, $\gamma_j = h(y_{(j)})$, and $\theta = (\gamma^T, \beta^T)^T$. The interpretation of $\beta$ depends on the link function. For example, if $G(\cdot)$ is the log odds link, $\beta$ is a log odds ratio; if $G(\cdot)$ is the log-log link, $\beta$ is a log hazard ratio. The intercepts $\gamma$ are the link function transformed CDFs when all covariates are set equal to 0. This also represents the transformation needed for the response variable to be modeled by a linear model.

With clustered data, we cannot directly fit NPMLEs for CPMs because observations are not independent. However, since the CPM is parameterized as an expectation $\mu_{it,j} = E(Z_{it,j} | X_{it})$, we can rely on GEE techniques to estimate parameters in (6) for valid inference, GEE requires correct specification of the marginal model for the response mean. GEE also permits specification of within cluster response dependence with a working correlation structure. The working correlation structure does not have to represent the true structure; however, to the extent that it differs from the true structure, efficiency losses incur (Liang & Zeger, 1986; Zeger & Liang, 1986). GEE methods for clustered ordinal responses have been developed (Heagerty & Zeger, 1996; Huang et al., 2002; Lipsitz et al., 1994; Parsons et al., 2006; Touloumis et al., 2013).

We estimate $\theta$ in (6) using ordinal GEE methods by solving the equation

$$A_{\theta}(\theta; \alpha) = \sum_{i=1}^{N} D_i^T W_i^{-1} (Z_i - \mu_i) = 0,$$

where $D_i = \frac{\partial}{\partial \theta} \mu_i$, $W_i = S_i^T R_i(\alpha) S_i^T$, and $\alpha$ is a vector of association parameters. $R_i(\alpha)$ is a working correlation matrix for $Z_i$, and $S_i$ is a $M_i(J-1) \times M_i(J-1)$ block matrix with elements based on the variance of $Z_{it,j}, \{\mu_{it,j}(1 - \mu_{it,j})\}$. $W_i^{-1}$ can be considered as a weight matrix for subject $i$. Efficiency is improved to the extent that the working correlation matrix $R_i(\alpha)$ is a better approximation to the true correlation structure of $Z_i$. The structure of $R_i(\alpha)$ is assumed by the analyst, and $\alpha$ can then be estimated with a second estimating function that will be described in more detail in Section 3.3. The covariance of $\hat{\theta}$ is given by

$$V_{\theta}(\alpha) = \left( \sum_{i=1}^{N} D_i^T W_i^{-1} D_i \right)^{-1} \left( \sum_{i=1}^{N} D_i^T W_i^{-1} \text{Cov}(Z_i) W_i^{-1} D_i \right) \left( \sum_{i=1}^{N} D_i^T W_i^{-1} D_i \right)^{-1},$$

which can estimated by replacing $\theta$ with $\hat{\theta}$ and $\text{Cov}(Z_i)$ with $(Z_i - \mu_i)(Z_i - \mu_i)^T$.

Since $\mu_{it,j} = F(Y_{it}|X_{it})$ is a CDF, other quantities can be readily obtained from a fitted CPM. The CDF can be calculated with $\hat{F}(y|X) = G^{-1}(\hat{\gamma}_j - \hat{\beta}^T X)$, where $j$ is the index such that $y_{(j)} = \max\{y_j: y_j \leq y\}$. We can derive its standard error with the delta method. Similar to the scalar response setting, cross-sectional summaries (e.g., quantiles and expectations) and confidence intervals can be calculated from $\hat{\theta}$ and $\hat{V}_{\theta}(\alpha)$.

Fitting ordinal GEE methods to clustered continuous response data is computationally challenging. Specifically, for each observation $Y_{it}$, we need $J - 1$ indicators $Z_{it,j} = I(Y_{it} \leq y_{(j)})$, and $J$ is usually a large number for continuous data, which implies that $W_i$ and $D_i$ in (7) and (8) can be high-dimensional. In the following subsections, we introduce two computationally efficient implementations, first with independent working correlation structures and second with more complex working correlation structures commonly used for GEE-based estimation.

### 3.2 CPMs with independent working correlation

It is well known that working covariance weighting can be more efficient than working independent weighting, particularly for parameters corresponding to time-varying covariates. However, the independent working correlation structure is simpler and therefore easier to implement than other structures because it does not require estimating $\alpha$, and the computational burden of matrix inversion is reduced with a diagonal structure. In addition, there are settings where using an independent working correlation structure is recommended for statistical reasons, the most common of which occurs when interest is
in the cross-sectional $E(Y_{ij} | X_{ij})$ but where $E(Y_{ij} | X_{ij}) \neq E(Y_{ij} | X_{11}, ..., X_{M})$. In such settings, one must use an independent working correlation to ensure consistent estimates of time-varying covariate parameters (Diggle et al., 2002; Pepe & Anderson, 1994; Schildcrout & Heagerty, 2005).

CPMs for scalar response data can be fitted when there are thousands of distinct values. With an independent working correlation structure, solving (7) for $\theta$ and plugging $\hat{\theta}$ into (8) to estimate the variance is equivalent to treating the response data as unclustered, computing the NPMLEs of CPMs as described in Section 2, and then correcting estimates of uncertainty by using a sandwich-variance estimate (see Web Appendix B). Specifically, we fit CPMs to clustered continuous response variables by maximizing the marginal likelihood

$$L(\theta) = \prod_{j=1}^{J} \prod_{i \in j} \left( G^{-1}(y_{ij} - \theta^T X_{ij}) - G^{-1}(y_{j-1} - \theta^T X_{il}) \right)$$

$$= \prod_{j=1}^{J} \prod_{i \in j} \left( \mu_{il,j} - \mu_{i,j-1} \right).$$

(9)

To correct for correlated responses within each cluster, we use the Huber–White sandwich estimator for covariance (Freedman, 2006). Since the clusters are independent but observations within clusters are dependent, we group within-cluster observations. Let

$$l(\theta) = \log(L(\theta)) = \sum_{j=1}^{J} \sum_{i \in j} \log(f_{il,j})$$

be the log-likelihood of (9) under the assumption of independent observations, where $f_{il,j} = \mu_{il,j} - \mu_{i,j-1}$. The first and second-order partial derivatives of $l(\theta)$ with respect to $\theta$ are

$$l'(\theta) = \frac{\partial l(\theta)}{\partial \theta} = \sum_{j=1}^{J} \sum_{i \in j \cap y_{ij} = y_{ij}} \frac{\partial \log(f_{il,j})}{\partial \theta} = \sum_{j=1}^{J} \sum_{i \in j \cap y_{ij} = y_{ij}} g_{il,j},$$

$$l''(\theta) = \frac{\partial^2 l(\theta)}{\partial \theta^2} = \sum_{j=1}^{J} \sum_{i \in j \cap y_{ij} = y_{ij}} \frac{\partial^2 \log(f_{il,j})}{\partial \theta^2},$$

and the Huber–White sandwich estimator for Cov($\hat{\theta}$) is given by

$$(l''(\hat{\theta}))^{-1} \left( \sum_{i=1}^{N} \left( \sum_{j=1}^{M_i} \hat{g}_{il,j} \right) \left( \sum_{j=1}^{M_i} \hat{g}_{il,j} \right)^T \right) (l''(\hat{\theta}))^{-1},$$

(10)

where $\sum_{i=1}^{M_i} \hat{g}_{il,j}$ is the sum of the plug-in estimators for the first partial derivative elements within a cluster. Consistency and asymptotic normality of estimates and the validity of the sandwich estimators are shown under the conditions provided in Web Appendix C (Ni et al., 2022). Point estimates and robust covariances for CPMs can be obtained with the or() and robcov() functions in the rms package in R (Harrell, 2020).

### 3.3 CPMs with exchangeable/AR1 working correlation

Though computationally efficient, CPMs with an independent working correlation structure can be statistically inefficient if the within cluster correlation is high and/or clusters are large. GEE methods for ordinal response variables allow for more complicated working correlation structures to improve efficiency, and approaches have been proposed. For example, Lipsitz et al. (1994) estimated association parameters with Pearson residuals; Heagerty and Zeger (1996) extended alternating logistic regression for binary longitudinal outcomes to ordinal longitudinal outcomes with pairwise log-odds ratio parameters as the association parameters (Carey et al., 1993; Lipsitz et al., 1991); and Touloumis et al. (2013) captured response association with local odds ratios based on Goodman’s row and column effects models.

To improve efficiency over the independent working correlation approach described above, we appeal to the framework proposed by Parsons et al. (2006, 2009) that specifies the association parameter $\alpha$ as a correlation and that estimates the parameter iteratively by minimizing the determinant of $V_{\alpha}(\alpha)$. This method, which Parsons et al. (2009) called “repolr” (repeated measures proportional odds logistic regression), estimates $\alpha$ based on the covariance matrix, whose dimension is relatively manageable. In contrast, other ordinal GEE methods require enumerating all pairs of observations within each cluster to estimate $\alpha$, which is computationally intensive for continuous response data.

In repolr, $R(\alpha)$ is constructed as $R(\alpha) = K(\alpha) \otimes C$, where $K(\alpha)$ is a $M_i \times M_i$ within-cluster working correlation matrix and $C$ is a $(J-1) \times (J-1)$ matrix of correlations among elements in $Z_{il}$. $C$ is set to be common for every pair of binary, ordinal-level indicators for each subject at each time point, so that

$$C = \begin{bmatrix} \rho_{11} & \ldots & \rho_{1(J-1)} \\ \vdots & \ddots & \vdots \\ \rho_{(J-1)1} & \ldots & \rho_{(J-1)(J-1)} \end{bmatrix},$$
where \( \rho_{pq} \) is expected correlation between \( Z_{ip} \) and \( Z_{iq} \) for \( i = 1, \ldots, N \). With the logit link, \( \rho_{pq} = \rho_{qp} = \{\exp(\gamma_p - \gamma_q)\}^{\frac{1}{2}} \) where \( p < q \) (Kenward et al., 1994). Two common structures for \( K(\alpha) \) are exchangeable (also called uniform or compound symmetric) and first-order autoregressive (AR1) structures (Diggle et al., 2002). For the exchangeable structure, \( K(\alpha) = 1 \) if \( p = q \) and \( K(\alpha) = \alpha \) otherwise; for the AR1 structure, \( K(\alpha) = 1 \) if \( p = q \) and \( K(\alpha) = \alpha^{1-q+1} \) otherwise. Note that the working correlation structure is specified for \( Y \), not \( Z \); this is illustrated in Web Appendix D with a simple example. The additional estimating equation for the association parameter \( \alpha \) in repolr is

\[
\frac{\partial \log |V_\theta(\alpha)|}{\partial \alpha} = 0, \tag{11}
\]

which is equivalent to estimating \( \alpha \) by minimizing \( \log |V_\theta(\alpha)| \). The equation solves for \( \alpha \) that minimizes the confidence region size of the \( \theta \). The algorithm iterates between solving (7) for \( \theta \) and solving (11) for \( \alpha \) until convergence. This approach can be applied with the \texttt{repolr()} function in the \texttt{repolr} package in R (Parsons, 2016) for complete data and for the logit link.

With continuous response variables, it may still be expensive to run a fully iterated repolr model; hence, we propose a one-step GEE estimator for repolr (Lipsitz et al., 2017). In our setting, instead of iterating between the two estimating equations (7) and (11) until convergence, we start with an estimate of \( \theta \) under an independent working correlation structure, \( \hat{\theta}_i \), which can be efficiently estimated with CPMs. We then obtain the association parameter \( \hat{\alpha} \) by solving (11) with \( V_{\theta_i}(\alpha) \). Finally, we solve (7) using \( \hat{\alpha} \) to get \( \hat{\theta}_i \), which is asymptotically equivalent to the fully iterated GEE estimator (Lipsitz et al., 2017).

We built an R package, \texttt{cpmgee} (Tian, 2022), that applies this one-step estimation procedure for exchangeable and AR1 working correlation structures. This package also fits CPMs with independent working correlation.

Although the one-step GEE estimator for repolr can substantially reduce the computational burden, computation with exchangeable and AR1 working correlation structures may still be intensive if the number of distinct values of a continuous response variable is large. One may seek to reduce the number of distinct values in the response by binning. Specifically, the \( N' = \sum_{i=1}^{N} M_i \) observations can be divided into \( M_b \) bins, where the value assigned to each observation in the bin is the median value for observations in that bin. The median is a natural choice to capture the center of the bin, although it should be noted that \( \beta \) coefficients are not impacted by the chosen value as long as the order of the bins is preserved. Approximately equal-quantile binning can be achieved by expressing \( N' \) as

\[
N' = M_b q + r = (M_b - r) q + r(q + 1),
\]

where \( q \) is the integer quotient of \( N'/M_b \) and \( r \) is the remainder. Thus, \( M_b - r \) bins have \( q \) observations, and \( r \) bins have \( q + 1 \) observations. Rounding to a decimal place is yet another way to reduce the number of distinct values. Strategies for binning and rounding cross-sectional CPMs with very large sample sizes are provided elsewhere (Li, Chen, et al., 2022).

## 4 | SIMULATIONS

We studied the performance of our estimators applying CPMs with independent, exchangeable, and AR1 working correlation to continuous clustered data under various simulation settings. Responses were generated in the following manner for subject \( i \) at time \( t \): 

\[
Y_{it} = \text{Inv}X^2(\Phi(Y^*_{it})/2, df = 5), \quad Y^*_{it} = X_i \beta_X + T_{it} \beta_T + \epsilon_{it},
\]

where Inv-\(X^2(\cdot, \ df = 5)\) is the inverse of the CDF for a chi-square distribution with 5 degrees of freedom and \( \Phi(\cdot) \) is the probability density function of the standard normal distribution. The transformation has been used in earlier work (Tian et al., 2020) and was chosen because it does not correspond to a commonly used closed-form transformation.

In the primary study setting, we set the sample size \( N = 1000 \), \( X \) was a time-invariant covariate with standard normal distribution, and \( T \) represented time, a time-varying covariate, with values \{0, 0.2, 0.4, 0.6, 0.8, 1\}. We imposed dropout completely at random uniformly across \( t \in \{3, 4, 5, 6\} \), such that the number of observations per subject varied from 2 to 6. A logistic residual distribution was used, and the correlation structure was exchangeable with \( \alpha = 0.7 \). We set \( \beta_X = 1 \) and \( \beta_T = 1 \). For CPMs with exchangeable and AR1 working correlation structures, we fit models using equal-quantile binning with \( M_b = 300 \).

We also explored scenarios with a smaller \( \alpha \), different values of \( M_b \) for binning, and rounding with different decimal places. Additional simulation settings including the identity transformation; complete data; differing sample sizes, cluster sizes, time effects, and correlation structures; and link function misspecification are shown in Web Appendix E.

We replicated each scenario 1000 times and evaluated for percent bias, root mean squared error (RMSE), empirical standard error, average estimated standard error, and coverage of 95\% confidence intervals. We compared our
methods with a gold standard for estimating $\beta$, in which GEE with an identity link and the correct working correlation structure was fit to the correctly transformed continuous response variable. We also investigated the performance of estimates of the conditional expectation, median, and CDF—specifically, $E(Y|X = 1, T = 0.2)$, $Q(0.5|X = 1, T = 0.2)$, and $F(5|X = 1, T = 0.2)$, respectively—that were estimated from the fitted CPMs. We do not show the average estimated standard error of $Q(0.5|X = 1, T = 0.2)$ because its confidence interval was obtained from linear interpolation of the inverse of the confidence interval for the conditional CDF (Liu et al., 2017).

Computation times for our CPM methods are shown in Web Appendix F. CPM fits with independent working correlation structure are computationally efficient and can handle thousands of distinct values in the response variable. CPM fits with exchangeable working correlation are much slower, although they can be sped up by binning the outcome.

### 4.1 The primary setting

Simulation results under the primary study setting with $N = 1000$ and $\alpha = 0.7$ as well as a secondary setting with $\alpha = 0.3$ are shown in Table 1. For the primary setting ($\alpha = 0.7$), CPMs performed quite well with low bias and generally good coverage for $\beta_X$, $\beta_T$, $E(Y|X = 1, T = 0.2)$, $Q(0.5|X = 1, T = 0.2)$, and $F(5|X = 1, T = 0.2)$. CPMs with an independent working correlation structure had minimal bias and coverage near 0.95. Estimates of $\beta_T$ from CPMs with a properly specified, exchangeable working correlation structure exhibited modest bias ($\sim 3\%$) that resulted in somewhat lower coverage (0.91). As expected, the exchangeable working correlation structure was much more efficient than the independent structure (e.g., empirical SE (Empirical standard error) equal to 0.069 vs. 0.091).

We observed efficiency losses when fitting CPMs with an exchangeable working correlation (empirical SE equal to 0.069) compared to the gold standard GEE estimator that applies the appropriate transformation and then uses an identity link with an exchangeable working correlation structure (empirical SE equal to 0.060). However, it is important to note that in practice, the appropriate transformation is not known to the analyst. Further, estimates of the conditional expectation cannot be extracted from the gold standard GEE estimator because $h[E(Y^*|X = 1, T = 0.2)] \neq E[h(Y^*)|X = 1, T = 0.2] = E(Y|X = 1, T = 0.2)$, nor can estimates of the conditional quantities and exceedance probabilities without making additional distributional assumptions. Working exchangeable and independent structures yielded approximately equal precision when estimating the conditional mean, median, and CDF. This is not surprising because the conditional quantities are estimated using linear combinations of the estimated $\gamma$, $\beta_X$, and $\beta_T$ values, and substantive efficiency gains associated with the exchangeable structure are only observed for $\beta_T$ estimates.

In the secondary setting, when the within cluster correlation was relatively low ($\alpha = 0.3$), CPMs were approximately valid and all relative efficiencies compared to the gold standard GEE analysis were much closer to 1. We note that for the CPM with an exchangeable working correlation structure, the bias in the parameter estimate was similar to the primary data setting ($\sim 3\%$). However, because there was more uncertainty (empirical SE equal to 0.086 vs. 0.069) the impact of the biased parameter estimate had less impact on coverage.

#### 4.2 Equal-quantile binning and rounding

In the primary simulation setting, when applying CPMs with exchangeable working correlation, we used equal-quantile binning with $M_b = 300$. We observed substantial efficiency gains for estimating $\beta_T$ compared CPMs with independent working correlation that used no binning, but we also observed a small bias in the parameter estimate. To investigate the sensitivity of the procedure to binning choice, we repeated simulations while imposing different binning/rounding strategies. Table 2 shows results.

As we decreased $M_b$ from 200 to 25, we observed that the bias in the $\beta_T$ estimate effectively vanished, and the coverage improved. Importantly, there was no noticeable change in estimation precision from coarsening the data. With that said, we observed slightly worse performance in estimates of the conditional quantities with fewer bins, particularly when $M_b$ was quite low (25 or 50). In contrast, rounding to 0 decimal places resulted in 169 categories in the response variable on average, resulting in poor performance for estimating $Q(0.5|X = 1, T = 0.2)$ and $F(5|X = 1, T = 0.2)$. Rounding is a suboptimal choice for such right-skewed responses because many distinct values at the lower end of the distribution are rounded to a single value. There were 498 ordinal levels on average if the response variable was rounded to one decimal place; consistent with binning results for $M_b = 300$, coverage for $\beta_T$ was 0.91, but the performance of the conditional estimators was quite good.

To summarize, as expected, working correlation structures can improve efficiency compared to working independent structures, particularly for estimating coefficients for time-varying covariates when correlation within clusters is high. This is the primary motivation for their use. However, CPMs with nonindependent working cor-
TABLE 1 Simulation results for CPMs for the primary setting and its modifications with lower within cluster correlation ($\alpha = 0.3$). For comparison, gold standard models were fit using GEE with the identity link and the correct exchangeable working correlation structure after correctly transforming the response variable.

| $\alpha$ | Method   | Metric | $\beta_X$ | $\beta_T$ | $E(Y|X=1,T=0.2)$ | $Q(0.5|X=1,T=0.2)$ | $F(5|X=1,T=0.2)$ |
|----------|----------|--------|-----------|-----------|-----------------|------------------|-----------------|
| 0.7      | Bias(%)  | $-0.010$ | $0.087$ | -         | -               | -                | -               |
|          | RMSE     | $0.050$  | $0.060$ | -         | -               | -                | -               |
|          | Gold Empirical SE | $0.050$  | $0.060$ | -         | -               | -                | -               |
|          | Standard Average SE | $0.051$  | -       | -         | -               | -                | -               |
|          | Coverage | $0.953$  | $0.944$ | -         | -               | -                | -               |
|          | RE       | reference | reference | -     | -               | -                | -               |
|          | Bias(%)  | $0.129$  | $0.270$ | $-0.009$ | $-0.074$        | $-0.169$          | -               |
|          | RMSE     | $0.054$  | $0.091$ | $1.232$  | $1.199$         | $0.171$           | -               |
|          | CPM Empirical SE | $0.054$  | $0.091$ | $0.139$  | $0.132$         | $0.016$           | -               |
|          | (ind) Average SE | $0.055$  | $0.088$ | $0.142$  | -               | -                | -               |
|          | Coverage | $0.957$  | $0.942$ | $0.956$  | $0.958$         | $0.956$           | -               |
|          | RE       | $1.129$  | $2.279$ | -         | -               | -                | -               |
|          | Bias(%)  | $0.234$  | $2.983$ | $-0.181$ | $-0.270$        | $-0.077$          | -               |
|          | RMSE     | $0.052$  | $0.075$ | $1.224$  | $1.191$         | $0.170$           | -               |
|          | CPM Empirical SE | $0.052$  | $0.069$ | $0.135$  | $0.130$         | $0.015$           | -               |
|          | (ex) Average SE | $0.053$  | $0.067$ | $0.137$  | -               | -                | -               |
|          | Coverage | $0.957$  | $0.910$ | $0.948$  | $0.956$         | $0.958$           | -               |
|          | RE       | $1.047$  | $1.310$ | -         | -               | -                | -               |
| 0.3      | Bias(%)  | $-0.061$ | $0.127$ | -         | -               | -                | -               |
|          | RMSE     | $0.040$  | $0.089$ | -         | -               | -                | -               |
|          | Gold Empirical SE | $0.040$  | $0.089$ | -         | -               | -                | -               |
|          | Standard Average SE | $0.040$  | $0.087$ | -         | -               | -                | -               |
|          | Coverage | $0.955$  | $0.943$ | -         | -               | -                | -               |
|          | RE       | reference | reference | -     | -               | -                | -               |
|          | Bias(%)  | $0.063$  | $0.254$ | $-0.015$ | $-0.054$        | $-0.069$          | -               |
|          | RMSE     | $0.041$  | $0.092$ | $1.236$  | $1.204$         | $0.171$           | -               |
|          | CPM Empirical SE | $0.041$  | $0.092$ | $0.107$  | $0.105$         | $0.013$           | -               |
|          | (ind) Average SE | $0.042$  | $0.091$ | $0.105$  | -               | -                | -               |
|          | Coverage | $0.959$  | $0.946$ | $0.957$  | $0.952$         | $0.959$           | -               |
|          | RE       | $1.063$  | $1.073$ | -         | -               | -                | -               |
|          | Bias(%)  | $0.160$  | $2.929$ | $-0.196$ | $-0.249$        | $0.023$           | -               |
|          | RMSE     | $0.041$  | $0.091$ | $1.227$  | $1.195$         | $0.170$           | -               |
|          | CPM Empirical SE | $0.041$  | $0.086$ | $0.105$  | $0.105$         | $0.013$           | -               |
|          | (ex) Average SE | $0.041$  | $0.085$ | $0.109$  | -               | -                | -               |
|          | Coverage | $0.961$  | $0.936$ | $0.953$  | $0.943$         | $0.959$           | -               |
|          | RE       | $1.041$  | $0.943$ | -         | -               | -                | -               |

Notes: The average computation time for the gold standard, CPM(ind) and CPM(ex) estimators under the primary setting ($\alpha = 0.7$) was 0.13, 42, and 2937 s. The time was based on 100 simulation replications performed on a 64-bit Linux server equipped with 24 Intel Xeon E5-2620 processors running at 2.40 GHz. Abbreviations: CPM, cumulative probability model; RMSE, root mean squared error; Empirical SE, empirical standard error; Average SE, average estimated standard error; RE, relative efficiency.

4.3 Other simulation results

Results for other simulation settings are in Web Appendix E. With complete data and the same time-varying covariate pattern across all subjects, CPMs with
TABLE 2  Simulation results for fitting CPMs with exchangeable working correlation with equal-quantile binned and rounded response data based on the primary setting. For equal-quantile binning, we show results of $M_b = 25, 50, 100$ and 200. Results of rounding to 0 and 1 decimal place are also shown.

| Scenario | Metric | $\beta_X$ | $\beta_T$ | $E(Y|X=1,T=0.2)$ | $Q(0.5|X=1,T=0.2)$ | $F(5|X=1,T=0.2)$ |
|----------|--------|-----------|-----------|------------------|---------------------|------------------|
| Binning $M_b = 25$ | Bias(%) | 0.171 | 0.548 | -1.064 | -2.258 | 0.142 |
| | RMSE | 0.052 | 0.068 | 1.205 | 1.107 | 0.170 |
| | Empirical SE | 0.052 | 0.067 | 0.134 | 0.132 | 0.018 |
| | Average SE | 0.053 | 0.067 | 0.129 | - | 0.016 |
| | Coverage | 0.956 | 0.945 | 0.889 | 0.881 | 0.901 |
| Binning $M_b = 50$ | Bias(%) | 0.174 | 0.757 | -0.039 | -0.053 | -0.233 |
| | RMSE | 0.052 | 0.068 | 1.205 | 1.166 | 0.171 |
| | Empirical SE | 0.052 | 0.068 | 0.135 | 0.131 | 0.016 |
| | Average SE | 0.053 | 0.067 | 0.133 | - | 0.016 |
| | Coverage | 0.958 | 0.942 | 0.929 | 0.923 | 0.935 |
| Binning $M_b = 100$ | Bias(%) | 0.187 | 1.193 | -0.316 | -0.493 | -0.174 |
| | RMSE | 0.052 | 0.069 | 1.217 | 1.181 | 0.171 |
| | Empirical SE | 0.052 | 0.068 | 0.135 | 0.130 | 0.016 |
| | Average SE | 0.053 | 0.067 | 0.135 | - | 0.016 |
| | Coverage | 0.957 | 0.936 | 0.945 | 0.948 | 0.953 |
| Binning $M_b = 200$ | Bias(%) | 0.208 | 2.069 | -0.197 | -0.311 | -0.112 |
| | RMSE | 0.052 | 0.071 | 1.223 | 1.189 | 0.170 |
| | Empirical SE | 0.052 | 0.068 | 0.135 | 0.131 | 0.015 |
| | Average SE | 0.053 | 0.067 | 0.136 | - | 0.016 |
| | Coverage | 0.957 | 0.924 | 0.946 | 0.952 | 0.958 |
| Rounding 0 decimal place | Bias(%) | 0.196 | 0.799 | -0.015 | -7.316 | -20.965 |
| | RMSE | 0.052 | 0.070 | 1.231 | 0.940 | 0.222 |
| | Empirical SE | 0.054 | 0.069 | 0.136 | 0.155 | 0.014 |
| | Average SE | 0.053 | 0.068 | 0.139 | - | 0.014 |
| | Coverage | 0.959 | 0.937 | 0.952 | 0.244 | 0.004 |
| Rounding 1 decimal place | Bias(%) | 0.210 | 3.180 | -0.123 | -0.693 | -2.147 |
| | RMSE | 0.052 | 0.076 | 1.229 | 1.175 | 0.176 |
| | Empirical SE | 0.052 | 0.070 | 0.136 | 0.130 | 0.015 |
| | Average SE | 0.053 | 0.067 | 0.138 | - | 0.016 |
| | Coverage | 0.957 | 0.907 | 0.953 | 0.942 | 0.943 |

Abbreviations: CPM, cumulative probability model; RMSE, root mean squared error; Empirical SE, empirical standard error; Average SE, average estimated standard error.

Independent working correlation were as efficient as those with exchangeable working correlation structure (Table S1 in Web Appendix). This is consistent with other studies of the performance of GEE with exchangeable versus independent structures with balanced and complete data (Mancl & Leroux, 1996). With $N = 100$, CPMs with independent working correlation showed good performance while CPMs with exchangeable correlation had substantial bias; this bias decreased as sample size increased. With $N = 500$, the relationship between the number of bins and bias for $\beta_T$ with exchangeable working correlation was similar to that seen with $N = 1000$ and was seemingly dependent on $N/M_b$ (Table S4 in Web Appendix). With data generated with AR1 correlation, CPMs with AR1 working correlation performed well and were almost as efficient as the gold standard GEE with the identity link under the correct transformation with AR1 working correlation structure (Table S7 in Web Appendix). CPMs had reasonable performance with moderate link function misspecification, that is, when data were generated with normal residuals but fit using the logit link function (Table S8 in Web Appendix).

5 APPLICATIONS

To illustrate the use of the proposed CPM methods, we applied them on the CD4:CD8 ratio and Lung Health Study datasets presented in the Introduction.


**Table 3** Odds ratio estimates of higher CD4:CD8 ratios with pointwise 95% confidence intervals from CPMs with independent working correlation and CPMs with exchangeable working correlation with binning (1000 equal-quantile bins) and rounding (two decimal places) are shown. Variance ratios (VRs) are calculated by the variances of the log-odds ratios from CPMs with exchangeable working correlation divided by the variances of the log-odds ratios from CPMs with independent working correlation. Notably, VRs are the same up to two decimals for binning and rounding.

| Predictor          | Independent | Exchangeable (binning) | Exchangeable (rounding) | VR   |
|--------------------|-------------|------------------------|-------------------------|------|
| Time (years)       | 1.22 (1.08, 1.37) | 1.23 (1.14, 1.33) | 1.23 (1.13, 1.32) | 0.43 |
| Enrollment year    | 1.01 (0.98, 1.04) | 1.01 (0.99, 1.04) | 1.014 (0.99, 1.04) | 0.81 |
| Race               |             |                        |                         |      |
| African American   | (Reference) |                        |                         |      |
| Caucasian          | 1.01 (0.83, 1.24) | 1.07 (0.89, 1.29) | 1.06 (0.88, 1.28) | 0.88 |
| Hispanic           | 0.68 (0.46, 0.99) | 0.73 (0.50, 1.06) | 0.72 (0.50, 1.05) | 0.98 |
| Other              | 0.72 (0.47, 1.12) | 0.74 (0.49, 1.12) | 0.73 (0.48, 1.11) | 0.89 |
| Baseline age (10 years) | 0.67 (0.61, 0.74) | 0.68 (0.62, 0.74) | 0.68 (0.62, 0.74) | 0.88 |
| Sex                |             |                        |                         |      |
| Male               | (Reference) |                        |                         |      |
| Female             | 1.72 (1.32, 2.25) | 1.80 (1.40, 2.32) | 1.80 (1.40, 2.32) | 0.90 |
| Route              |             |                        |                         |      |
| Heterosexual       | (Reference) |                        |                         |      |
| Injection drug use | 0.99 (0.68, 1.46) | 0.93 (0.64, 1.35) | 0.93 (0.64, 1.35) | 0.93 |
| MSM                | 0.90 (0.70, 1.17) | 0.90 (0.71, 1.15) | 0.90 (0.71, 1.14) | 0.89 |
| Other/unknown      | 0.79 (0.47, 1.35) | 0.86 (0.54, 1.38) | 0.85 (0.53, 1.37) | 0.79 |
| HCV                | 0.82 (0.60, 1.14) | 0.81 (0.60, 1.09) | 0.81 (0.60, 1.09) | 0.85 |
| HBV                | 0.99 (0.66, 1.49) | 0.92 (0.64, 1.31) | 0.92 (0.64, 1.32) | 0.77 |

Abbreviations: CPM, cumulative probability model; HBV, hepatitis B virus; HCV, hepatitis C virus; MSM, men who have sex with men.

### 5.1 CD4:CD8 ratio

Longitudinal measures of the CD4:CD8 ratio were collected on 1763 adults with HIV on stable ART (1 year on ART with undetectable viral load) at the Vanderbilt Comprehensive Care Clinic. We are interested in modeling the CD4:CD8 ratio during the second year on ART. A mean of 2.9 CD4:CD8 measurements per subject was available (median = 3; range = 1–7) with 3862 distinct outcome values. Time-invariant covariates considered were calendar year at baseline (1 year after ART initiation), race, baseline age, sex, probable route of infection, hepatitis C virus infection status, and hepatitis B virus infection status. Time (in years) after baseline was the only time-varying covariate.

CPMs with independent working correlation are able to handle 3862 ordinal levels efficiently, while CPMs with exchangeable/AR1 working correlation require binning or rounding due to computational complexities. Thus, we divided the outcome into 1000 bins and rounded to two decimal places. Equal-quantile binning resulted in 979 ordinal levels due to ties. The two decimal place rounding led to 234 levels. The logit link was used in all models.

Odds ratio estimates and 95% confidence intervals from the fitted CPMs are shown in Table 3. The results suggest that time, race, baseline age, and sex are associated with the CD4:CD8 ratio. For example, fixing other variables, a 10-year increase in baseline age is associated with a 33% decrease in the odds of having a higher CD4:CD8 ratio based on the CPM with an independent working correlation. Results are fairly similar across all three fitted CPMs.

There were some differences in efficiency of estimates across different CPM estimating procedures. The variance ratios in Table 3 correspond to the variances of the log-odds ratios from CPMs with exchangeable working correlation divided by the variances of the log-odds ratios from CPMs with independent working correlation. The variances for the estimated log-odds ratio for the time-varying covariate, time, for the two exchangeable working correlation models was 0.43 times that for the independent working correlation model. We saw variance ratios ranging from 0.77 to 0.98 for time-invariant covariate parameter estimates.

In addition to odds ratios, other quantities can be estimated from the fitted CPMs. Conditional means, medians, and probabilities of CD4:CD8 being greater than 1 are shown as a function of time in Figure 1 with other covariates fixed at their median (for continuous covariates) or mode (for categorical covariates) levels. The CD4:CD8...
ratio above 1 is considered normal for people without HIV (Petoumenos et al., 2017). Results from the three models were generally very close. We also included the conditional mean obtained by a standard GEE model with an identity link and independent working correlation without transforming the response data for purpose of comparison; results from this analysis are also fairly similar.

5.2 Lung function health study

The purpose of the Lung Health Study was to determine whether a smoking intervention program and the use of an inhaled bronchodilator could slow the rate of decline in lung function in participants with early COPD (Anthonisen et al., 1994). Our interest was in the association of a single nucleotide polymorphism, rs12194741, with lung function decline over 5 years as measured with FEV1 (Hansel et al., 2013). rs12194741 was represented with a binary indicator for the presence of at least 1 copy of the T allele. We evaluate the genetic contribution to lung function decline by the interaction of rs12194741 and visit with data collected from participants’ annual visits. We included participants who were continuous smokers (dropping observations after smoking stopped) with at least two observations. 2562 subjects were included, and 1694 (66%) completed five visits. Baseline adjustment covariates included age, study site, body mass index (BMI, weight (kg)/height (m²)), lifetime smoking status (in pack years), and average number of cigarettes smoked per day over the year prior to enrollment. BMI change from baseline and study visit was included as time-varying covariates. The distribution of the responses, FEV1, was fairly symmetric (Web Appendix G).

We applied CPMs with independent and AR1 working correlation structures and used the logit link function. Neither binning nor rounding was applied prior to fitting the models as we felt data from 2562 participants was more than sufficient to support the 361 distinct values of the outcome. Table 4 shows odds ratio estimates of higher FEV1 and 95% confidence intervals obtained from the two methods. The odds ratios from the two models were very close. The variance ratios shown in the last column indicate that, as expected, the log-odds ratio estimates obtained by CPMs with the AR1 working correlation structure were more precise than those from CPMs with the independent structure, particularly for time-varying covariates (visit and BMI change). The confidence interval for the interaction term did not cover 1, consistent with rs12194741 being associated with more rapid lung function decline.

Conditional quantities including means, medians, and $P(\text{FEV1} \leq 2 \mid X)$ as a function of study visit and genotype were derived from the models and are in Web Appendix G.

6 DISCUSSION

We extended CPMs, a class of ordinal regression models for cross-sectional responses, to analyze clustered continuous response data. In scalar-response settings, CPMs have been used to fit different types of continuous response variables (Liu et al., 2017). Only rank information is used in CPMs when estimating $\hat{\beta}$, and thus fitting such ordinal regression models can avoid transformations of response
TABLE 4  Odds ratios estimates for higher FEV1 with 95% confidence intervals from CPMs with independent and AR1 working correlation. The last column shows the variance ratios (VRs) calculated by the variances of the log-odds ratios from CPMs with AR1 working correlation divided by the variances of the log-odds ratios from CPMs with independent working correlation. Nine site indicators (LHS (Lung Health Study) was a 10-site study) were also included in the regression analyses but are not shown.

| Predictor                         | Independent        | AR1            | VR  |
|-----------------------------------|--------------------|----------------|-----|
| Visit                             | 0.86 (0.84, 0.88)  | 0.86 (0.85, 0.87) | 0.64 |
| rs12194741                        | 1.12 (0.97, 1.29)  | 1.12 (0.97, 1.29) | 0.97 |
| Visit × rs12194741 interaction     | 0.97 (0.94, 0.99)  | 0.97 (0.95, 0.99) | 0.61 |
| BMI Change (per 5 kg/m²)           | 0.65 (0.53, 0.80)  | 0.66 (0.56, 0.76) | 0.56 |
| Baseline age (per 10 year)         | 0.34 (0.30, 0.39)  | 0.34 (0.30, 0.39) | 0.90 |
| Baseline BMI (per 5 kg/m²)         | 1.48 (1.34, 1.63)  | 1.48 (1.35, 1.62) | 0.88 |
| Cigarettes/day (per 10 cigs/day)   | 0.98 (0.92, 1.03)  | 0.98 (0.92, 1.03) | 0.96 |
| Pack years (per 20 pack year)      | 1.19 (1.09, 1.30)  | 1.19 (1.09, 1.30) | 0.98 |

Abbreviations: AR1, first order autoregressive working response dependence structure; BMI, body mass index.

variables. To account for correlation between observations within each cluster, we estimated parameters in CPMs using GEE techniques. We proposed feasible and computationally efficient approaches for fitting CPMs with independent, exchangeable, and AR1 working correlation structures. With the estimated parameters, we can obtain CDFs, expectations, and quantiles conditional on covariates to help interpret regression results. Our approaches work well under a variety of simulation settings studied for this paper. We built an R package, cpmgee, for CPMs with independent, exchangeable and AR1 working correlation.

Fitting CPMs with an exchangeable/AR1 working correlation structure is more computationally burdensome than using an independent working correlation structure and often requires binning/rounding. In addition, exchangeable/AR1 working correlation structures can be problematic when the full covariate conditional model is not equal to the fitted cross-sectional model (Pepe & Anderson, 1994; see Section 5.7 in Web Appendix E for brief simulation and discussion). However, the goal of using a working covariance weighting over independent weighting is statistical efficiency. With time-varying covariates, unequal cluster sizes, and high within cluster correlation, our simulations are consistent with prior studies in that substantial efficiency gains can be realized by appropriately employing exchangeable/AR1 working correlation structures. This holds even after binning. Considerations for the extent of binning should depend on analytical goals. A smaller number of bins is more computationally efficient and provides highly efficient and approximately valid inferences for both time-invariant and time-varying covariate coefficients. In contrast, if one is focused on estimating conditional expectations, medians, or exceedance probabilities, we recommend using a larger number of bins or using an independent working correlation structure.

There are several avenues of interest that were not addressed fully in this paper and warrant further study. These include the following: (1) The semiparametric method proposed herein requires specifying a link function. Characterizing the extent to which the inference is impacted under misspecified link functions would be useful, though some small-scale simulations are presented in Table S7 of the Web Appendix. (2) A deeper study that provides additional guidance on binning strategies is warranted. (3) In simulation studies that are not shown, we noticed circumstances where the one-step repolr estimator appeared more efficient than a fully iterated repolr estimator. This observation was surprising to us and warrants further investigation. (4) In future research, we plan to extend CPMs to include sampling weights, which we believe will allow us to fit fully continuous clustered data with more complex working correlation structures. Implementation of weights would also allow us to extend these methods to address challenges associated with missing data and/or confounding.

ACKNOWLEDGMENTS
We thank Jessica Castilho for the data for the HIV study. Data for The Lung Health Study were downloaded from the National Center for Biotechnology Information’s Database of Genotypes and Phenotypes (accession no. phs000335.v2.p2). The project was supported by funding from the U.S. National Institutes of Health grants R01 AI093234, R01 HL094786, R01 HL072966, P30 AI110527, K23 AI120875, and N01-HR-46002.

DATA AVAILABILITY STATEMENT
The data for the CD4:CD8 ratio in this paper will not be made publicly available for ethical and privacy reasons. The dataset contains sensitive information. Furthermore, we do not own the data, and we made an agreement with those who collected the data that we will not share it. We are happy to provide others with the relevant information for how they might be able to obtain the data.
under reasonable requests (please contact Bryan Shepherd at bryan.shepherd@vumc.org). Data from the Lung Health Study are available from the National Institutes of Health’s database of Genotypes and Phenotypes (dbGaP, https://www.ncbi.nlm.nih.gov/gap/), and one may request access to these data by using the Study Accession number: phs000335.v3.p2.

ORCID
Yuqi Tian https://orcid.org/0000-0003-0601-3454
Bryan E. Shepherd https://orcid.org/0000-0002-3758-5992
Chun Li https://orcid.org/0000-0002-8819-2443
Donglin Zeng https://orcid.org/0000-0003-0843-9280
Jonathan S. Schildcrout https://orcid.org/0000-0003-2276-2551

REFERENCES
Agresti, A. (2010) Analysis of ordinal categorical data, Wiley Series in Probability and Statistics, volume 656. New York: John Wiley & Sons.

Anthonisen, N.R., Connett, J.E., Kiley, J.P., Altose, M.D., Bailey, W.C., Buist, A.S., et al. (1994) Effects of smoking intervention and the use of an inhaled anticholinergic bronchodilator on the rate of decline of FEV1: the Lung Health Study. Journal of the American Medical Association, 272, 1497–1505.

Carey, V., Zeger, S.L. & Diggle, P. (1993) Modelling multivariate binary data with alternating logistic regressions. Biometrika, 80, 517–526.

Castilho, J.L., Shepherd, B.E., Koethe, J., Turner, M., Behawy, S., Logan, J., et al. (2016) CD4/CD8 ratio, age, and risk of serious non-communicable diseases in HIV-infected adults on antiretroviral therapy. AIDS, 30, 899.

da Silva, C.M., de Peder, L.D., Silva, E.S., Previdelli, I., Pereira, O.C.N., Teixeira, J.J.V., et al. (2018) Impact of HBV and HCV coinfection on CD4 cells among HIV-infected patients: a longitudinal retrospective study. The Journal of Infection in Developing Countries, 12, 1009–1018.

Diggle, P., Diggle, P.J., Heagerty, P., Liang, K.-Y., Zeger, S., et al. (2002) Analysis of longitudinal data. New York: Oxford University Press.

Freeman, D.A. (2006) On the so-called “Huber sandwich estimator” and “robust standard errors”. The American Statistician, 60, 299–302.

Global Initiative for Chronic Obstructive Lung Disease (2021) Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease (2022) Global Initiative for Chronic Lung Disease, Inc. https://goldcopd.org/wp-content/uploads/2020/11/GOLD-REPORT-2021-v1.1-25Nov20_WMV.pdf

Gras, L., May, M., Ryder, L.P., Trickey, A., Helleberg, M., Obel, N., et al. (2019) Determinants of restoration of CD4 and CD8 cell counts and their ratio in HIV–1–positive individuals with sustained virological suppression on antiretroviral therapy. Journal of Acquired Immune Deficiency Syndromes, 80, 292.

Hansel, N.N., Ruczinski, I., Rafaels, N., Sin, D.D., Daley, D., Malinina, A., et al. (2013) Genome-wide study identifies two loci associated with lung function decline in mild to moderate COPD. Human Genetics, 132, 79–90.

Harrell, F. (2020) rms: regression modeling strategies. R package version 6.1.0. https://CRAN.R-project.org/package=rms

Heagerty, P.J. & Zeger, S.L. (1996) Marginal regression models for clustered ordinal measurements. Journal of the American Statistical Association, 91, 1024–1036.

Huang, G.-H., Bandeen-Roche, K. & Rubin, G.S. (2002) Building marginal models for multiple ordinal measurements. Journal of the Royal Statistical Society Series C: Applied Statistics, 51, 37–57.

Kenward, M.G., Lesaffre, E. & Molenberghs, G. (1994) An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. Biometrics, 50, 945–953.

Li, C., Chen, G. & Shepherd, B.E. (2022) Fitting semiparametric cumulative probability models for big data. arXiv [Preprint]. Available from: https://doi.org/10.48550/arXiv.2207.06562 (accessed Wed, 13 July 2022).

Li, C., Tian, Y., Zeng, D. & Shepherd, B.E. (2022) Asymptotic properties for cumulative probability models for continuous outcomes. arXiv [Preprint]. Available from: https://doi.org/10.48550/arXiv.2206.14426 (accessed 18 Feb 2023).

Liang, K.-Y. & Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. Biometrika, 73, 13–22.

Lipsitz, S., Fitzmaurice, G., Sinha, D., Hevelone, N., Hu, J. & Nguyen, L.L. (2017) One-step generalized estimating equations with large cluster sizes. Journal of Computational and Graphical Statistics, 26, 734–737.

Lipsitz, S.R., Kim, K. & Zhao, L. (1994) Analysis of repeated categorical data using generalized estimating equations. Statistics in Medicine, 13, 1149–1163.

Lipsitz, S.R., Laird, N.M. & Harrington, D.P. (1991) Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. Biometrika, 78, 153–160.

Liu, Q., Shepherd, B.E., Li, C. & Harrell Jr, F.E. (2017) Modeling continuous response variables using ordinal regression. Statistics in Medicine, 36, 4316–4335.

Mancl, L.A. & Leroux, B.G. (1996) Efficiency of regression estimates for clustered data. Biometrics, 52, 500–511.

McCullagh, P. (1980) Regression models for ordinal data. Journal of the Royal Statistical Society, Series B (Statistical Methodology), 42, 109–127.

McCullagh, P. & Nelder, J.A. (1983) Generalized linear models. New York: Routledge.

Parsons, N. (2016) repolr: Repeated Measures Proportional Odds Logistic Regression. R package version 3.4. https://CRAN.R-project.org/package=repolr

Parsons, N.R., Costa, M.L., Achten, J. & Stallard, N. (2009) Repeated measures proportional odds logistic regression analysis of ordinal score data in the statistical software package R. Computational Statistics & Data Analysis, 53, 632–641.

Parsons, N.R., Edmondson, R. & Gilmour, S. (2006) A generalized estimating equation method for fitting autocorrelated ordinal score data with an application in horticultural research. Journal of the Royal Statistical Society. Series C (Applied Statistics), 55, 507–524.

Pepe, M.S. & Anderson, G.L. (1994) A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. Communications in Statistics-simulation and Computation, 23, 939–951.
Petoumenos, K., Choi, J.Y., Hoy, J., Kiertiburanakul, S., Ng, O.T., Boyd, M., et al. (2017) CD4:CD8 ratio comparison between cohorts of HIV-positive Asians and Caucasians upon commencement of antiretroviral therapy. *Antiviral Therapy*, 22, 659–668.

Sauter, R., Huang, R., Ledergerber, B., Battegay, M., Bernasconi, E., Cavassini, M., et al. (2016) CD4/CD8 ratio and CD8 counts predict CD4 response in HIV-1-infected drug naive and in patients on cART. *Medicine*, 95, e5094.

Schildcrout, J.S. & Heagerty, P.J. (2005) Regression analysis of longitudinal binary data with time-dependent environmental covariates: bias and efficiency. *Biostatistics*, 6, 633–652.

Tian, Y. (2022) The cpmgee package. Available from: https://github.com/YuqiTian35/cpmgee (accessed June 2022).

Tian, Y., Hothorn, T., Li, C., Harrell Jr, F.E., & Shepherd, B.E. (2020) An empirical comparison of two novel transformation models. *Statistics in Medicine*, 39, 562–576.

Tian, Y., Li, C., Tu, S., James, N.J., Harrell, F.E. & Shepherd, B.E. (2022) Addressing detection limits by semiparametric cumulative probability models. arXiv [Preprint]. Available from: https://doi.org/10.48550/arXiv.2207.02815

Touloumis, A., Agresti, A. & Kateri, M. (2013) GEE for multinomial responses using a local odds ratios parameterization. *Biometrics*, 69, 633–640.

Wedderburn, R.W. (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61, 439–447.

Zeger, S.L. & Liang, K.-Y. (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42, 121–130.

Zeng, D. & Lin, D. (2006) Efficient estimation of semiparametric transformation models for counting processes. *Biometrika*, 93, 627–640.

**SUPPORTING INFORMATION**

Web Appendices, Tables, and Figures referenced in Section 3-5, and R package cpmgee and code for simulation studies and application analysis referenced in Section 3-5, are available with this paper at the Biometrics website on Wiley Online Library. The R package cpmgee and code are also on Github (package: github.com/YuqiTian35/cpmgee, code: github.com/YuqiTian35/CPMGEE_Code).

**How to cite this article:** Tian, Y., Shepherd, B.E., Li, C., Zeng, D. & Schildcrout, J.S. (2023) Analyzing clustered continuous response variables with ordinal regression models. *Biometrics*, 79, 3764–3777. https://doi.org/10.1111/biom.13904