Journal of Data Science **1**(2003), 149-165

A Classification Statistic for GEE Categorical Response Models

John M. Williamson[1], Hung-Mo Lin[2] and Huiman X. Barnhart[3]
[1]*Centers for Disease Control and Prevention,*
[2]*Penn State College of Medicine and* [3]*Emory University*

Abstract: A kappa-like classification statistic is proposed for assessing the fit of GEE regression models with a categorical response. The proposed statistic is a summary measure depicting how well categorical responses are predicted from the fitted GEE model. The statistic takes on a value of 1 if prediction is perfect and a value of 0 if the fitted model fares no better than random chance, i.e., fitting the repeated categorical responses with an intercept-only model. To demonstrate the usefulness of the classification statistic, we present simulation results as well as two examples from biomedical studies.

Key words: Cumulative logistic regression, kappa statistic, logistic regression, ordinal response.

1. Introduction

Data analysts are often confronted with the problem of modeling a response variable with various covariates. It is then of crucial importance to assess how well the final model fits the data. For normally distributed outcomes, numerous approaches are available for assessing model fit (Neter, Wasserman, and Kutner, 1985, chap. 4). However, in the biomedical and social sciences the response of interest is often a categorical variable, whether binary, nominal, or ordinal. Assessing the aptness of a categorical regression model is difficult because the residuals are not normally distributed, as is true of linear regression. In logistic regression, for example, the residuals

are of only two forms, $-\widehat{\pi}$ and $1-\widehat{\pi}$, in which $\widehat{\pi}$ is the fitted probability, further complicating model fit.

Often the categorical responses will be correlated in clusters, e.g., in repeated measure studies. The observations within a cluster will often be positively correlated, i.e., the observations will tend to be more like each other than like observations from other clusters. This correlation must be taken into account when analyzing clustered studies for proper inference and valid hypothesis testing.

Generalized estimating equations (GEE) are useful for analyzing such correlated data with categorical or continuous responses (Liang and Zeger, 1986; Zeger and Liang, 1986). Parameter estimation is conducted through estimating equations which converge to a sum of mean zero random variables if the mean structure is correctly specified. There is no need to specify a joint distribution for the responses. However, assessing model fit is further complicated with GEE than for models assuming independence because no likelihood is available and the residuals are correlated within a cluster.

Some methods are available for assessing the fit of GEE regression models with binary responses. Horton, Bebchuck, Jones, Lipsitz, Catalano, Zahner, and Fitzmaurice (1999) developed a goodness-of-fit test for assessing such model fit by extending Hosmer and Lemeshow's (1989) goodness-of-fit statistic for ordinary logistic regression. Their proposed test statistic has an approximate chi-squared distribution when the model is specified correctly. Barnhart and Williamson (1998) also propose a goodness-of-fit statistic for assessing the fit of GEE binary regression models. They extend Tsiatis' method (1980) for assessing the fit of ordinary logistic regression models. This approach involves partitioning the space of covariates into distinct regions and forming score statistics that are asymptotically distributed as chi-square random variables with the appropriate degrees of freedom. Barnhart and Williamson's approach is best employed in the situation when there are only discrete covariates available because then there is no need to partition the covariates. Pan (2002) has proposed goodness-of-fit tests for GEE with correlated binary data. Pan's two tests result in the Pearson chi-square and an unweighted sum of residual squares, both of which are based on the residuals. These two tests can only be used when there is at least one continuous covariate available.

If the possibility of influential observations is of concern to the data analyst, Preisser and Qaqish (1996) have proposed deletion diagnostics for generalized estimating equations. The diagnostics consider leverage and residuals to measure the influence of a subset of observations on the fitted regression parameters. Preisser and Qaqish (1999) also generalize the GEE procedure to produce parameter estimates and fitted values that are resistant to influential data.

Here, the concern is assessing the fit of GEE categorical response models by determining how well the covariates predict the subject's responses. In Section 2, we present the proposed kappa-like classification statistic, which indicates how well the proposed model predicts the categorical response. Simulation results are presented in Section 3 to demonstrate the usefulness of the proposed classification statistic. Analyses of two biomedical studies are used to illustrate the proposed statistic in Section 4, and we conclude with a short discussion.

2. Classification statistic

Lipsitz, Kim, and Zhao (1994) extend the GEE approach of Liang and Zeger (1986) to the analysis of correlated categorical response data. They propose models for the correlation between repeated nominal or ordinal categorical responses and their methods reduce to Liang and Zeger's method when the repeated responses are binary. The repeated categorical responses can arise from a longitudinal study or from other correlated data settings, such as familial or ophthalmologic studies.

We outline Lipsitz, Kim, and Zhao's (1994) method as follows. Assume the response of interest is a categorical outcome with K categories denoted $Z_{it} = k$ if the tth subunit from the ith cluster falls in the kth category, for $i = 1, \cdots, N$; $t = 1, \cdots, T_i$ ($T = \max(T_i) \, \forall \, i = 1, \cdots, N$); and $k = 1, \cdots, K$. For example, in an ophthalmologic study the subunit will be either the left or right eye and the cluster will be the subject under observation. For simplicity, we will assume that the data are balanced, i.e., $T_i = T$ for $i = 1, \ldots, N$. The following method will still be applicable in the case of unbalanced data. The $T(K-1) \times 1$ response vector $\mathbf{Y}_i$ for cluster i consists of the binary random variables Y_{itk}, where $Y_{itk} = 1$ if $Z_{it} = k$.

Typically one models the marginal cumulative probabilities of response, $\vartheta_{itk} = Pr(Z_{it} \leq k)$ for $k = 1, \cdots, K-1$. The marginal probabilities are denoted by $\pi_{itk} = Pr(Z_{it} = k) = Pr(Y_{itk} = 1) = E(Y_{itk}) = \vartheta_{itk} - \vartheta_{it,k-1}$ and will comprise the $T(K-1)\times 1$ vector $\boldsymbol{\pi}_{\mathbf{i}}$. The vectors $\mathbf{Y_i}$ and $\boldsymbol{\pi}_i$ require only $T(K-1)$ elements instead of TK elements because $\sum_{k=1}^{K} Y_{itk} = \sum_{k=1}^{K} \pi_{itk} = 1$, for $i = 1, \cdots, N$ and $t = 1, \cdots, T$. Let $\mathbf{X_{it}}$ be the $p \times 1$ covariate vector for the tth subunit of the ith cluster.

The cumulative marginal response probabilities will be related to the covariates via the link function g, the kth intercept λ_k, and the $p\times 1$ marginal parameter vector β,

$$g(\vartheta_{itk}) = \lambda_k + \mathbf{X}'_{it}\boldsymbol{\beta}.$$

The intercepts are in increasing order: $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{K-1}$. For an ordinal response, the function g may be any link function such as the logit function, probit function (Φ^{-1}), or the complementary log-log function. For a nominal response, one may model the marginal probabilities directly using the polytomous link function. Lipsitz, Kim, and Zhao (1994) suggest that one estimates $\boldsymbol{\beta}$ with the following set of generalized estimating equations:

$$\boldsymbol{\nu}_1(\boldsymbol{\beta}) = \sum_{i=1}^{N} \mathbf{D}'_i \mathbf{V}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\pi}_i) = \mathbf{0},$$

where $\mathbf{D}_i = \mathbf{D}_i(\boldsymbol{\beta}) = d\,\boldsymbol{\pi}_i(\boldsymbol{\beta})/d\boldsymbol{\beta}$, $\mathbf{V}_i = \mathbf{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) \approx \text{var}(\mathbf{Y}_i)$ is a "working" covariance matrix of $\mathbf{Y}_i$ (Liang and Zeger, 1986; Zeger and Liang, 1986) and $\boldsymbol{\alpha}$ is a $q \times 1$ vector of correlation parameters. The parameters $\boldsymbol{\alpha}$ are associated with the correlation between the elements of the vectors $\mathbf{Y}_{is}$ and $\mathbf{Y}_{it}$. For example, if one chooses an exchangeable correlation matrix as the "working" correlation matrix, then $\boldsymbol{\alpha}$ is comprised of just one parameter (α) which represents the correlation between any two responses in a cluster, which are all assumed equal. See Lipsitz, Kim, and Zhao (1994) for further details and for other choices of $\mathbf{V}_i$.

We propose a kappa-like statistic to assess model fit for GEE categorical response models. Historically, the kappa coefficient has been used to determine the agreement of binary (Cohen, 1960) and categorical (Fleiss, 1971) outcomes between raters. Kappa corrects the percentage of agreement between raters by taking into account the proportion of agreement

expected by chance. Kappa has been used as a measure of reproducibility in many epidemiologic settings, such as studies involving twin similarity (Klar, Lipsitz, and Ibrahim, 2000) and control-informant agreement collected from case-control studies (Korten, Jorm, Henderson, McCusker, and Creasey, 1992). The general expression for the kappa statistic is

$$\kappa = \frac{P_o - P_e}{1 - P_e},$$

where P_o is the observed proportion of agreement and P_e is the proportion of agreement expected by chance alone (Fleiss, 1971). A value of 0 for κ indicates no agreement beyond chance and a value of 1 indicates perfect agreement, among many of κ's desirable properties (Fleiss, 1981). Thus, larger values of κ indicate greater agreement between the outcomes.

Here we use κ as a measure of agreement between the predicted and observed categorical responses to assess the fit of the GEE model. We estimate κ in a second set of estimating equations, similar to Lipsitz, Laird, and Brennan (1994); Klar, Lipsitz, and Ibrahim (2000); and Williamson, Manatunga, and Lipsitz (2000). With Lipsitz, Kim, and Zhao's (1994) method, we estimate the probability of the response falling in each of the K categories. Denote this estimated probability for the kth category, tth subunit, of the ith cluster as $\widehat{\pi}_{itk}$. We do not have a straightforward predicted response as with linear regression. However, if we did have a predicted response for the tth subunit of the ith cluster, denoted $\widehat{Z}_{it}$, it is natural to assume that $\widehat{Z}_{it}$ would equal k ($k = 1, \cdots, K$) with probability $\widehat{\pi}_{itk}$. Let P_{oit} denote the probability that the predicted response from the model is equal to the observed response, i.e., $\widehat{Z}_{it} = Z_{it}$. A natural estimate of P_{oit} is obtained by using $\widehat{\pi}_{itZ_{it}}$, the estimated probability from the fitted model that the response falls into the observed category for the tth subunit of the ith cluster. We define κ_{it} as the agreement between the predicted and observed responses for the tth subunit of the ith cluster as follows:

$$\kappa_{it} = \frac{P_{oit} - P_e}{1 - P_e},$$

where P_{oit} is defined above and P_e is the probability of correct prediction expected by chance alone.

As an estimate of P_e, we suggest fitting an intercept-only model. Cox and Snell (1989, pp. 208-209) and Nagelkerke (1991) proposed using an intercept-only model as a baseline model when generalizing the coefficient of determination for assessing the fit of a logistic regression model. Thus, we will fit a model with just the intercepts, the λ_k parameters for $k = 1, \cdots, K-1$, and no covariates. This "baseline" model will provide a good starting point from which to compare the proposed model. The estimated category probabilities from the intercept-only model will be the same for all clusters and subunits, and will be denoted

$$\widehat{p}_{itk} = \widehat{p}_k = \sum_{i=1}^{N}\sum_{t=1}^{T} I(Z_{it} = k)/NT = \sum_{i=1}^{N}\sum_{t=1}^{T} Y_{itk}/NT = n_k/NT,$$

where n_k is the sum of observations in category k, i.e., $n_k = \sum_{i=1}^{N}\sum_{t=1}^{T} Y_{itk}$ for $k = 1, \ldots, K$. All n_k observations with response category k will each be correctly predicted with probability $\widehat{p}_k$; accordingly, the estimate of P_e is

$$\widehat{P}_e = \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{t=1}^{T} I(\widehat{Z}_{it} = Z_{it} = k)/NT = \sum_{k=1}^{K} n_k\widehat{p}_k/NT = \sum_{k=1}^{K} \widehat{p}_k^2.$$

The agreement between two raters for assessing a categorical outcome with K categories can be depicted in a $K \times K$ contingency table (Fleiss, 1981, p. 219). The row and column total probabilities for the kth outcome category are $p_{k.}$ and $p_{.k}$, the marginal probabilities that the two raters assess the outcome in the kth category. The estimate of the probability expected by chance is calculated assuming independence between the rows and columns in the contingency table and is $\widehat{P}_e = \sum_{k=1}^{K} p_{k.}p_{.k}$, which is similar to the estimate above.

We will estimate an overall κ to ascertain the fit of the model ($\kappa = \kappa_{it}$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$). By noting that $P_{oit} = P_e + \kappa(1.0 - P_e)$, we propose a second set of estimating equations as follows. Let $\mathbf{P}_{oi}$ and $\mathbf{U}_i$ denote the $T \times 1$ vectors $[P_{oi1}, \ldots, P_{oiT}]'$ and $[\widehat{\pi}_{i1Z_{i1}}, \ldots, \widehat{\pi}_{iTZ_{iT}}]'$. The second set of estimating equations are, thus,

$$\boldsymbol{\nu}_2(\kappa, \boldsymbol{\beta}) = \sum_{i=1}^{N} \mathbf{C}_i^{'}\, \mathbf{W}_i^{-1}\{\mathbf{U}_i(\boldsymbol{\beta}) - \mathbf{P}_{oi}(\kappa)\} = \mathbf{0},$$

where $\mathbf{C}_i = d\mathbf{P}_{oi}/d\kappa = [1 - \widehat{P}_e, \cdots, 1 - \widehat{P}_e]'$ and $\mathbf{W}_i \approx \text{var}(\mathbf{U}_i)$ is the $T \times T$ working covariance matrix of $\mathbf{U}_i$. To compute $(\widehat{\boldsymbol{\beta}}, \widehat{\kappa})$, one can use a Fisher-scoring-type algorithm such as

$$\widehat{\boldsymbol{\beta}}^{(m+1)} = \widehat{\boldsymbol{\beta}}^{(m)} - \Big[\sum_{i=1}^{N} \widehat{\mathbf{D}}_i' \, \widehat{\mathbf{V}}_i^{-1} \widehat{\mathbf{D}}_i\Big]^{-1} \Big[\sum_{i=1}^{N} \widehat{\mathbf{D}}_i' \widehat{\mathbf{V}}_i^{-1} \{\mathbf{Y}_i - \widehat{\boldsymbol{\pi}}_i(\widehat{\boldsymbol{\beta}}^{(m)})\}\Big]$$

and

$$\widehat{\kappa}^{(m+1)} = \widehat{\kappa}^{(m)} - \Big[\sum_{i=1}^{N} \widehat{\mathbf{C}}_i' \widehat{\mathbf{W}}_i^{-1} \widehat{\mathbf{C}}_i\Big]^{-1} \Big[\sum_{i=1}^{N} \widehat{\mathbf{C}}_i' \widehat{\mathbf{W}}_i^{-1} \{\mathbf{U}_i(\widehat{\boldsymbol{\beta}}^{(m+1)}) - \mathbf{P}_{oi}(\widehat{\kappa}^{(m)})\}\Big],$$

where m denotes the iteration. We use Liang and Zeger's (1986) empirically corrected variance estimate of $\widehat{\boldsymbol{\beta}}$ and Prentice's (1988) empirically corrected variance estimate of $\widehat{\kappa}$.

The second set of estimating equations can be solved non-iteratively if we choose the $T \times T$ identity matrix for $\mathbf{W}_i$:

$$\widehat{\kappa} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{\pi}_{itZ_{it}}/NT - \widehat{P}_e}{1 - \widehat{P}_e}.$$

The term $\sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{\pi}_{itZ_{it}}/NT$ is the average predicted probability corresponding to the observed responses. If the fitted model predicts the categorical response perfectly, i.e., $\widehat{\pi}_{itZ_i} = 1.0$, then $\widehat{\kappa} = 1.0$. If the fitted model predicts the responses no better than an intercept-only model, then $\widehat{\kappa} = 0.0$. This kappa-like classification statistic should be interpreted as the average probability of predicting the observed responses above and beyond the prediction by the intercept-only model.

3. Simulations

To assess the performance of the proposed classification statistic, we conducted analyses by using simulated data with known distributions. The goal of a first set of simulations was to use the proposed statistic to evaluate GEE regression models with correlated binary responses when significant and nonsignificant predictors were added to the model used to analyze the

data. We generated 500 data sets, each containing 100 clusters of three binary responses, $Y_{it} = 0, 1$ for $i = 1, \cdots, 100$ and $t = 1, 2, 3$. We generated multivariate normally distributed data with an exchangeable correlation structure (correlation equal to 0.3) and collapsed each of the three normally distributed random variables into a binary variable. The marginal model generating the data was $probit\{Pr(Y_{it} = 1)\} = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3it}$ for $i = 1, \cdots, 100$ and $t = 1, 2, 3$, with $x_{1i} = -0.5$ for $i \leq 50$ and 0.5 otherwise, x_{2i} distributed as a uniform$(-0.5, 0.5)$ random variable, and $x_{3it} = 1.0$ for $t = 3$ and 0.0 for $t = 1, 2$. Thus, we have two cluster-specific covariates and a time-varying covariate indicating the third occasion. The parameter values were $\beta_1 = 2.0$, $\beta_2 = 1.0$, $\beta_3 = 0.0$, and $\alpha = 0.0$. In essence, the responses were generated with only the covariates x_1 and x_2 because $\beta_3 = 0.0$. The data were analyzed with four GEE models using a marginal probit link function: (1) with just x_1 as a covariate, (2) with x_1 and x_3, (3) with x_1 and x_2, and (4) with all three variables as predictors. The kappa-like statistic ($\widehat{\kappa}$) and its standard error ($\widehat{se}(\widehat{\kappa})$) were calculated for each data set.

Table 1: Simulation results of GEE model with binary response

Model	$\widehat{\kappa}$	$\widehat{se}(\hat{\kappa})$
$probit\{Pr(Y_{it} = 1)\} = \alpha + \beta_1 x_1$	0.376	0.052
$probit\{Pr(Y_{it} = 1)\} = \alpha + \beta_1 x_{1it} + \beta_3 x_{3it}$	0.378	0.052
$probit\{Pr(Y_{it} = 1)\} = \alpha + \beta_1 x_{1it} + \beta_2 x_{2it}$	0.459	0.055
$probit\{Pr(Y_{it} = 1)\} = \alpha + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it}$	0.461	0.055

Values shown are the means of 500 simulations. The third model generated the data: $\beta_1 = 2.1, \beta_2 = 1.0$ and $\beta_3 = 0.0$.

Table 1 is a summary of these simulations. The average value over the 500 data sets of $\widehat{\kappa}$ is 0.376 for the model with only x_1 as a predictor. The average value of $\widehat{\kappa}$ increased only minimally to 0.378 when x_3 was added to the model, which was expected because the data were generated with $\beta_3 = 0.0$. When a covariate (x_2) with a moderate parameter value ($\beta_2 = 1.0$) was added to the model with x_1 and x_3, the average values of $\widehat{\kappa}$ increased

substantially to 0.461. For the model with only the covariates x_1 and x_2, the average estimate κ is 0.459, slightly smaller than the corresponding value from the full model. It is obvious from these simulation results that the estimated value of κ will increase when extra significant covariates are entered into the model and the inclusion of nonsignificant covariates will have little or no impact on κ.

A second set of simulations were conducted with an ordered categorical response ($K = 4$ categories) generated for three occasions. We again generated multivariate normally distributed data with an exchangeable correlation structure (correlation equal to 0.3) and collapsed the three normally distributed random variables into 4 categories. Two sets of 500 data sets were randomly generated with a person-specific binary covariate and a second time-varying covariate distributed as a Uniform$(-1, 1)$ random variable. In each data set, half of the observations were generated for each binary covariate value ($x_{1i} = -0.5$ or 0.5). The first set of simulations contained data sets of sample size of 100, and the second set of simulations contained data sets of sample size of 1000 to demonstrate the effect of sample size on the classification statistic. Thus, the model that generated the data is $probit\{Pr(Y_{it} \leq k)\} = \alpha_k + \beta_1 x_{1i} + \beta_2 x_{2it}$, for $k = 1, 2, 3$; $i = 1, \cdots, 100$ and 1000; and $t = 1, 2, 3$. The parameter values were $\alpha_1 = -1.250$, $\alpha_2 = 0.0$, $\alpha_3 = 1.250$, and $\beta_2 = 2.0$, corresponding to the uniformly distributed time-varying continuous covariate. Simulations were conducted on data generated randomly as β_1 (the binary covariate) ranged from 0.0 to 5.0 in increments of 1.0. The goal of this was to evaluate the performance of the proposed classification statistic when the model increases its predictive strength (as β_1 increases). The data sets were analyzed with GEE using the same model that generated the data.

Table 2 is a summary of these simulations. The average value over 500 data sets of $\widehat{\kappa}$ increased from 0.222 to 0.608 as β_1 increased from 0.0 to 5.0 for the data sets of a sample of size 100. The corresponding value of $\widehat{\kappa}$ for the data sets of a sample of size 1000 are very similar to the respective values for the data sets of a sample of size 100. The simulation results indicate that κ is not overly dependent on sample size; larger data sets do not result in assessments of poorer model fit.

Table 2: Simulation results of GEE model with ordinal response

	β_1 [a]	0.0	1.0	2.0	3.0	4.0	5.0
$n = 100$	$\widehat{\kappa}$	.222	.255	.331	.426	.517	.608
		(.018)	(.018)	(0.20)	(.025)	(.031)	(.034)
$n = 1000$	$\widehat{\kappa}$	.218	.248	.326	.420	.511	.062
		(.005)	(.005)	(.006)	(.008)	(.010)	(.011)

Values shown are the means of 500 simulations. Standard errors are shown below in parentheses.

[a] The data were generated by $probit\{Pr(Y_{it} \leq k)\} = \alpha_k + \beta_1 x_{1i} + \beta_2 x_{2it}$, where β_1 is the parameter associated with the binary covariate and $\beta_2 = 2.0$.

4. Examples

4.1 Wisconsin epidemiologic study of diabetic retinopathy

A total of 996 insulin-taking younger-onset diabetics in southern Wisconsin were examined for severity of diabetic retinopathy. The goal of the study was to determine the risk factors for diabetic retinopathy. The severity of diabetic retinopathy was originally graded on a 10-point scale. For the purpose of illustration, the 10-point scale was combined to form two categories: absence and presence of diabetic retinopathy. The cross-classification of diabetic retinopathy in left and right eyes for the 720 individuals with complete information is presented in Table 3. In this example, the cluster is the individual with correlated binary responses and the subunits are the left and right eyes. For more details concerning the study, see Klein, Klein, Moss, Davis, and Demets (1984).

First we fit a GEE logistic regression model with main effects terms only. The resulting kappa-like classification index had a value of 0.397 ($se = 0.046$). This model fit the data poorly according to the goodness-of-fit test statistic ($p < 0.001$) proposed by Barnhart and Williamson (1998).

Table 3: Frequency of diabetic retinopathy

Right eye \ Left eye	Absence	Presence	Total
Absence	237	38	275
Presence	31	414	445
Total	268	452	720

We then examined various interaction and quadratic variables for entry into the regression model at the significance level of 0.05. The quadratic terms for duration of diabetes and body mass index were significant and entered into the final model. The kappa-like classification index for this final model increased to 0.453 ($se = 0.039$) indicating a better fit. With the inclusion of the two quadratic terms, the model fit the data quite well ($p = 0.37$) according to Barnhart and Williamson's (1998) test. Results of these GEE regression analyses are presented in Table 4.

4.2 Arthritis clinical trial

We now illustrate the proposed classification statistic with a GEE ordinal logistic regression analysis of an arthritis clinical trial (Bombardier, Ware, and Russell; 1986). The trial compared the drug auranofin and placebo therapy for the treatment of rheumatoid arthritis. The response of interest is the self-assessment of arthritis, classified as 1 =poor, 2 =fair, 3 =good. Patients were randomized into either the auranofin treatment group or the placebo group after baseline self-assessment of arthritis, and then observed at one-, three- and five-month examinations. In this example, the cluster is the individual and the subunits are the three examinations. Table 5 presents the occasion-specific distribution of self-assessment of arthritis by treatment group. For more details of this GEE analysis, see Lipsitz, Kim, and Zhao (1994).

We first fit a GEE ordinal logistic regression model with the covariates Time_2, Time_3 (indicator variables for the second and third examinations), Sex (1 =male, 0 =female), Age (in years), and Auranofin (1 =yes,

Table 4: GEE logistic regression models for diabetic retinopathy study

Covariates	(a) Estimate	SE	(b) Estimate	SE
Intercept	-6.46	0.88	-14.58	2.02
Duration of diabetes (years)	0.23	0.03	0.44	0.03
Glycosylated hemoglobin level	0.13	0.03	0.15	0.03
Diastolic blood pressure	0.03	0.01	0.03	0.01
Body mass index	0.16	0.17	4.19	1.03
(Duration of diabetes)^2			-0.007	0.0007
(Body mass index)^2			-0.55	0.14
κ	0.397	0.046	0.453	0.039

(a)= model without quadratic terms, (b)= final model. All p-values are less then 0.001 except for the case of body mass index without the quadratic term, which is 0.35.

0 =placebo). The resulting estimate of κ was 0.012. Next we added Baseline (baseline self-assessment of arthritis, treated as continuous) to the model. The estimate of κ increased to 0.084 indicating a better, although still poor, fit. Results of these GEE analyses are presented in Table 6.

5. Discussion

The proposed kappa-like classification statistic is a more appropriate indicator of how well the model predicts the observed responses at the cluster level (e.g., an individual) as opposed to how well the model fits the data at the group level (e.g., treatment category). Often a model can fit the data well in terms of predicting the proportion of positive responses for a group of individuals, but is not necessarily useful for predicting a particular individual's response. The kappa-like classification index is an intuitive measure for assessing model fit in that it estimates the probability of an observation being correctly predicted by the fitted model. Then, this prob-

Table 5: Marginal distributions of self-assessment of arthritis

Treatment	Response	Occation Baseline	1 month	3 months	5 months
Auranofin	Poor	50	18	30	22
	Fair	69	77	52	51
	Good	34	56	66	73
	Total	153	151	148	146
Placebo	Poor	46	44	41	37
	Fair	70	50	63	52
	Good	33	54	44	58
	Total	149	148	148	147

ability is corrected for chance by comparing it to the probability that an intercept-only model would have correctly predicted the observation. The kappa-like statistic takes on a value of 0.0 for the intercept-only model and a value of 1.0 for the saturated model (NT parameters for N observations and T subunits per cluster, i.e., $\widehat{\pi}_{itZ_{it}} = 1.0$ for $i = 1, \cdots, N$ and $t = 1, \ldots, T$). An advantage of the proposed statistic is that no decisions need be made when calculating it, unlike methods based on covariate partitioning (where to partition, how many partitioned categories), Hosmer and Lemeshow's approach (how many groups, what to do when predicted probabilities are tied), rank correlation methods and classification tables (what the cutoff probability should be).

However, interpretation of the kappa statistic is not always straightforward; see Fleiss (1971) and Landis and Koch (1977) for details. In this context, we use kappa to compare the agreement between predicted probabilities and a categorical response. Therefore, P_{oit} for any observation will not obtain a maximum of 1.0 and, accordingly, high values of kappa from a fitted model will be unlikely. Also, Maclure and Willett (1987) noted that kappa is dependent upon the number of categories and that kappa will tend to be smaller for categorical responses with a greater number of categories.

Table 6: GEE ordinal logistic regression models for arthritis clinical trial

	Baseline Excluded			Baseline Included		
Covariates	Estimate	SE	p-value	Estimate	SE	p-value
Intercept_1	-1.489	0.465	0.001	0.667	0.537	0.214
Intercept_2	0.260	0.464	0.575	2.617	0.555	< 0.001
Time2	0.040	0.114	0.728	0.046	0.125	0.715
Time3	-0.246	0.105	0.019	-0.266	0.118	0.024
Sex (1 =male)	-0.108	0.197	0.584	-0.133	0.184	0.470
Age (years)	0.011	0.008	0.175	0.006	0.008	0.477
Auranofin	-0.479	0.177	0.007	-0.548	0.176	0.002
Baseline arthritis (continuous)				-1.035	0.137	< 0.001
κ	0.012	0.022		0.084	0.019	

However, there is no formula relating how kappa decreases as the number of categories of the outcome increases. Therefore, we suggest that data analysts interpret lower kappa values more favorably than would be the case in a traditional agreement study, especially for a categorical response with more than 2 categories. Similar to Landis and Koch's labeling of kappa values, we suggest interpreting the values of kappa for this classification index as follows:

Kappa Statistic	Fit of Model
0.00−0.20	Poor
0.21−0.40	Fair
0.41−0.60	Good
0.61−1.00	Excellent

For the diabetic retinopathy study, we would recommend that the kappa value (0.453) for the final model indicated good prediction, but the kappa

value (0.084) for the final model of the arthritis study indicated poor prediction. If some disagreement is acceptable when predicting a categorical response, then a weighted kappa approach may be formulated (Cohen, 1968; Gonin, Lipsitz, Fitzmaurice, and Molenberghs, 2000).

Acknowledgements

The authors thank Kevin Ward of Emory University and Dr. John Karon of the Centers for Disease Control and Prevention for helpful discussions and manuscript review.

References

Barnhart, H. X. and Williamson, J. M. (1998). Goodness-of-fit tests for GEE modeling with binary responses. *Biometrics* **54,** 720-729.

Bombardier, C., Ware, J. H. and Russell, I. J. (1986). Auranofin therapy and quality of life in patients with rheumatoid arthritis. *American Journal of Medicine* **81,** 565-578.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20,** 37-46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* **70,** 213-220.

Cox, D. R. and Snell, E. J. (1989). *The Analysis of Binary Data*, 2nd ed. Chapman and Hall, London.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin* **76,** 378-382.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions.* Wiley, New York.

Gonin, R., Lipsitz, S. R., Fitzmaurice, G. and Molenberghs, G. (2000). Regression modelling of weighted kappa by using generalized estimating equations. *Applied Statistics* **49,** 1-18.

Horton, N. J., Bebchuck, J. D., Jones, C. L., Lipsitz, S. R., Catalano, P. J., Zahner, G. E. P. and Fitzmaurice, G. M. (1999). Goodness-of-fit for GEE:

An example with mental health service utilization. *Statistics in Medicine* **18,** 213-222.

Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logistic Regression.* Wiley, New York.

Klar, N., Lipsitz, S. R. and Ibrahim, J. (2000). An estimating equations approach for modelling kappa. *Biometrical Journal* **42,** 45-58.

Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D., and Demets, D. L. (1984). The Wisconsin Epidemiologic Study of Diabetic Retinopathy: II. Prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years. *Archives of Ophthalmology* **102,** 520-526.

Korten, A. E., Jorm, A. F., Henderson, A. S., McCusker, E. and Creasey, H. Control-informant agreement on exposure history in case-control studies of Alzheimer's disease. *International Journal of Epidemiology* **21,** 1121-1131.

Landis, R. J. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33,** 159-174.

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

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

Lipsitz, S. R., Laird, N. M. and Brennan, T. A. (1994). Simple moment estimates of the κ-coefficient and its variance. *Applied Statistics* **43,** 309-323.

Maclure, M. and Willett, W. C. (1987). Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology* **126,** 161-169.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika* **78,** 691-692.

Neter, J., Wasserman, W. and Kutner, M. H. (1985). *Applied Linear Statistical Models.* Richard D. Irwin, Homewood, IL.

Pan, W. (2002). Goodness-of-fit tests for GEE with correlated binary data. *Scandinavian Journal of Statistics* **29,** 101-110.

Preisser, J. S. and Qaqish, B. F. (1996). Deletion diagnostics for generalised estimating equations. *Biometrika* **83,** 551-562.

Preisser, J. S. and Qaqish, B. F. (1999). Robust regression for clustered data with application to binary responses. *Biometrics* **55,** 574-579.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44,** 1033-1048.

Tsiatis, A. A. (1980). A note on a goodness-of-fit test for the logistic regression model. *Biometrika* **67,** 250-251.

Williamson, J. M., Manatunga, A. K. and Lipsitz, S. R. (2000). Modeling kappa for measuring dependent categorical agreement data. *Biostatistics* **1,** 191-202.

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

Received September 26, 2001; accepted June 15, 2002

John M. Williamson
Division of HIV/AIDS Prevention (MS E-37)
National Center for HIV, STD, and TB Prevention
Centers for Disease Control and Prevention, 1600 Clifton Rd., NE
Atlanta, Georgia 30333, USA.
jow5@cdc.gov

Hung-Mo Lin
Department of Health Evaluation Sciences
Penn State College of Medicine
P. O. Box 855, A210, 600 Centerview Drive
Hershey, Pennsylvania 17033, USA.
hlin@psu.edu

Huiman X. Barnhart
Department of Biostatistics
The Rollins School of Public Health of Emory University
1518 Clifton Rd., NE
Atlanta, Georgia 30322, USA.
hbarbha@sph.emory.edu