American Journal of Epidemiology

Vol. 165, No. 4
DOI: 10.1093/aje/kwk020
Advance Access publication November 22, 2006

Original Contribution

Marginal Modeling of Nonnested Multilevel Data using Standard Software

Diana L. Miglioretti[1] and Patrick J. Heagerty[2]

[1] Group Health Center for Health Studies, Seattle, WA.
[2] Department of Biostatistics, University of Washington, Seattle, WA.

Received for publication November 1, 2005; accepted for publication July 11, 2006.

Epidemiologic data are often clustered within multiple levels that may not be nested within each other. Generalized estimating equations are commonly used to adjust for correlation among observations within clusters when fitting regression models; however, standard software does not currently accommodate nonnested clusters. This paper introduces a simple generalized estimating equation strategy that uses available commercial or public software for the regression analysis of nonnested multilevel data. The authors describe how to obtain empirical standard error estimates for constructing valid confidence intervals and conducting statistical hypothesis tests. The method is evaluated using simulations and illustrated with an analysis of data from the Breast Cancer Surveillance Consortium that estimates the influence of woman, radiologist, and facility characteristics on the positive predictive value of screening mammography. Performance with a small number of clusters is discussed. Both the simulations and the example demonstrate the importance of accounting for the correlation within all levels of clustering for proper inference.

clustered data; generalized estimating equation; generalized linear model

Abbreviations: C1ID, observations belonging to cluster 1 (C2ID defined analogously); GEE, generalized estimating equation; ID, cluster-identifying variable; n-ID, ID for neighborhood; p-ID, ID for provider; PPV, positive predictive value.

Statistical analyses of epidemiologic studies often need to adjust for correlation among observations that arise in ''clusters.'' For example, outcomes for individual subjects may be clustered within providers and clinics. In other settings, clusters may arise from the repeated measurement of individuals over time in longitudinal studies or through social or geographic organization of individuals within communities or neighborhoods. Often, the multiple levels of clustering may not be perfectly nested within each other. For example, a population of patients may be treated at clinics served by a common collection of physicians, and physicians may practice at multiple clinics. Alternatively, correlation among healthcare utilization outcomes may be induced via shared providers or clinics and unmeasured geographic factors. When evaluating the relation between individual outcomes and variables measured at any cluster level, one must consider in a proper statistical analysis the potential correlation induced by unmeasured heterogeneity at each level of clustering.

One approach to analyzing multilevel data explicitly represents sources of unmeasured heterogeneity that induce clustering or correlation. Such methodology is referred to as multilevel, hierarchical, random effects, or mixed modeling and is implemented through computationally intensive maximum likelihood or Bayesian methods. In the case of one or two levels of clustering, nonlinear regression models (e.g., logistic regression) can be fit using standard software, such as SAS (SAS Institute, Inc., Cary, North Carolina) and STATA (StataCorp LP, College Station, Texas), even when the clusters are nonnested. However, when interest is in marginal (population-averaged) effects, most conditionally

Correspondence to Dr. Diana Miglioretti, Group Health Center for Health Studies, 1730 Minor Avenue, Suite 1600, Seattle, WA 98101 (e-mail: miglioretti.d@ghc.org).

specified multilevel nonlinear models do not directly address the scientific question and must either be marginalized to obtain model summaries or reparameterized to allow direct inference on marginal contrasts (1, 2). Ritz and Spiegelman (3) and Wang and Louis (4) discuss situations when coefficients from conditional and marginal regression models are equivalent. Several authors provide guidance on how to determine whether marginal or conditional models are of scientific interest (2, 5, 6).

When interest is in marginal effects, generalized estimating equations (GEEs) are commonly used (7). Regression analyses based on GEEs directly model the marginal mean and may be computationally feasible even with large numbers of observations and clusters. Although popular statistical packages, such as SAS and STATA, easily fit generalized linear models using GEEs, they do not currently directly accommodate nonnested clusters. For nonnested, multilevel, binary data with two levels of clustering, Miglioretti and Heagerty (8) compare and contrast a marginalized multilevel model fit using a likelihood (Bayesian) approach and a simple three-step, moment-based GEE method that uses standard software. In this paper, we describe this extension of GEEs, which is straightforward to implement and has the potential to deal with several levels of clustering. We evaluate the approach using simulated data and illustrate the method using data from the Breast Cancer Surveillance Consortium, with which we estimate the influence of woman, radiologist, and facility characteristics on the positive predictive value of screening mammography.

GENERALIZED ESTIMATING EQUATIONS APPROACH

Let Y_i represent the outcome of interest for the ith observation, $i = 1, \ldots, N$. We first consider the case when Y_i is clustered within two nonnested levels, C_1 and C_2 (e.g., a single outcome observed on each patient clustered within physicians and clinics where physicians may practice at more than one clinic). We model the influence of a $p \times 1$ vector of covariates $\mathbf{X}_i$ on the marginal mean $\mu_i = E(Y_i|\mathbf{X}_i)$ using a generalized linear model (9):

$$g(E(Y_i|\mathbf{X}_i)) = \mathbf{X}_i\boldsymbol{\beta}.$$

The covariates $\mathbf{X}_i$ may be measured at the level of the observation and/or the level of one or both levels of clustering.

GEEs were originally introduced for analysis of clustered data that arise through repeated measurements of individuals in a prospective longitudinal study. Standard software implementations of GEEs allow the user to select a ''working correlation'' matrix that is used to form a weighted version of the standard regression estimator. The specific choice of the correlation model is not crucial for valid inference (7). A ''sandwich variance'' estimator is empirically calculated from the data, and this accounts for arbitrary correlation among observations within a cluster. Thus, it is possible to simply use an independence working correlation matrix, which does not modify the estimated regression coefficients but provides inference on coefficient estimates that properly accounts for correlation within the specified clusters. In addition, if multiple clusters are perfectly nested, GEE clustering on the top level cluster accounts for the multilevel correlation structure through the sandwich variance estimator (10). For nonnested clusters, we detail use of GEEs with ''working independence'' to obtain the necessary sandwich variance calculations to provide valid standard errors for regression coefficients.

Implementation of GEEs requires a cluster-identifying variable (ID) that signifies potentially correlated outcomes. The sandwich variance is estimated by the sum of weighted residual cross-product terms $w_j(Y_j - \mu_j) \times w_k(Y_k - \mu_k)$ for all pairs of observations that are from the same cluster (with weights w_j and w_k specific to the regression model). One way to view this correction is through a simple logical operation on the ID variable. If ID(j) = ID(k), the observation pair (j, k) contributes a term to the sandwich variance estimator. For nonnested multilevel data, empirical covariance contributions from pairs of observations that are clustered on any one of the clustering variables need to be included. For example, if we have an ID for provider (p-ID) and an ID for neighborhood (n-ID), we need to include correlation contributions whenever (p-ID(j) = p-ID(k)) or (n-ID(j) = n-ID(k)). Viewed logically, we can evaluate the ''or'' operation using (share p-ID) or (share n-ID) = (share p-ID) + (share n-ID) – (share p-ID and n-ID). This connection reveals how we can use standard software to obtain a standard error correction for clustering within both providers and neighborhoods. Using GEE clustering on p-ID will include all cross-products that share a provider, and GEE clustering on n-ID will account for all intraneighborhood correlations. Finally, subtraction of cross-product terms for observations that have been counted twice, because they share both provider and neighborhood, is accomplished by using GEE clustering on the unique combinations of provider and neighborhood. This reasoning also demonstrates why analysis of perfectly nested clusters requires clustering only on the top-level cluster. If providers are nested within neighborhoods, observations that share p-ID also share n-ID; therefore, the third term (share p-ID and n-ID) is identical to the first term (share p-ID).

More generally, let C1ID identify observations belonging to the same cluster C_1, C2ID identify observations belonging to C_2, and C1C2ID identify observations belonging to both C_1 and C_2. The covariance matrix for the estimated regression coefficients $\hat{\boldsymbol{\beta}}$ is simply a linear combination of three covariance matrices estimated via GEE:

$$\mathbf{V}(\hat{\boldsymbol{\beta}}) = \mathbf{V}^1 + \mathbf{V}^2 - \mathbf{V}^3$$

where $\mathbf{V}^1$, $\mathbf{V}^2$, and $\mathbf{V}^3$ are the covariance matrices estimated from working independence GEE models clustering on C1ID, C2ID, and C1C2ID, respectively. The mathematical derivation of this result is provided in Miglioretti and Heagerty (8). Note that $\hat{\boldsymbol{\beta}}$ for all three models will be identical when using working independence. Given $\hat{\boldsymbol{\beta}}$ and $\mathbf{V}\left(\hat{\boldsymbol{\beta}}\right)$, standard Wald-based confidence intervals and test statistics may be calculated.

Sample SAS and STATA code for fitting a logistic regression model using the GEE approach with two levels of clustering is included in Appendix A.

Extension to more than two levels of clustering

The GEE approach can be extended to more than two levels of clustering; however, the number of models that must be fit grows quickly with increasing numbers of cluster levels ($2^K - 1$ fits are required for K levels of clustering). For example, three levels of clustering, C_1, C_2, and C_3, require fitting seven models: three clustering separately on each of the clustering factors; three clustering on each of the pairwise combinations of the cluster levels; and one clustering on the combination of all three cluster levels. The corrected covariance matrix is given by the addition of the first three covariance matrices minus each of the second three matrices plus the final matrix. This calculation follows the logical representation of (share C_1) or (share C_2) or (share C_3) using (share C_1) + (share C_2) + (share C_3) − (share C_1 and C_2) − (share C_1 and C_3) − (share C_2 and C_{13}) + (share C_1 and C_2 and C_3). Mathematical details are provided in Appendix B.

Performance with a small number of clusters

Sandwich variance estimates were originally proposed for use with independent data having nonconstant variance (heteroskedasticity) and for correlated longitudinal data. In each of these scenarios, typical applications involve hundreds of observations or clusters, and the empirical standard error estimator has been shown to perform well. However, when the number of independent clusters is small, the use of GEEs with a sandwich variance estimator has been shown to be anticonservative, resulting from biased standard error estimates (11, 12).

Mancl and DeRouen (13) overview several corrections to improve small-sample performance including the following: resampling methods such as the bootstrap or jackknife; alternative tailored sandwich estimators that explicitly account for the estimation of the regression coefficient; and a simple scaling by $K/(K-p)$, where K is the number of clusters and p is the number of regression parameters. Use of the jackknife correction involves multiple reestimation of the regression coefficients after removal of an independent block of data (usually one cluster). This approach could be adopted for estimation of the sandwich variance components ($\mathbf{V}^1, \mathbf{V}^2, \mathbf{V}^3$) that form the basis for estimation with nonnested data; however, direct use of jackknife methods is not straightforward when data have a crossed correlation structure, because independent blocks are not easily identified. Alternatively, the simple scaling correction, $K/(K-p)$, could be easily applied to each covariance matrix before they are combined. When the number of clusters within a level is less than 50, some sensitivity analysis is warranted to evaluate the potential for small-sample bias.

ILLUSTRATIONS USING SIMULATIONS

In this section, we use simulations to demonstrate the potential magnitude of standard error adjustment that can result from proper correction for clustering within nonnested levels and to evaluate our proposed method. We show the inaccuracy that may result from adjusting for only one level of clustering and demonstrate that the proposed GEE correction for both nonnested levels of clustering provides standard error estimates that approximate the true sampling variability and thus provide for valid inference. Previous research has shown that sandwich variance estimates can perform poorly when the number of clusters is small and correlation within clusters is large (13). Therefore, we present a series of simulations where we vary the number of clusters and the strength of dependence (or equivalently the magnitude of heterogeneity among the clusters).

We considered two levels of clustering in the generation of the data, and we generically refer to one level as *provider* with identifying variable p-ID and the second level as *neighborhood*, identified by the variable n-ID. Examples of such data include health-care utilization outcomes where observations are grouped into geographic units such as ZIP codes or census tracts (neighborhoods), but observations are also clustered into medical clinics (providers) or reliability studies where each medical image (forms neighborhood) is read and classified by multiple radiologists (providers).

To simulate data, we initially (scenario 1) created $J = 30$ provider IDs (p-ID) and $K = 50$ neighborhood IDs (n-ID). For each combination of p-ID and n-ID, we generated between $i = 1$ and 5 binary observations, Y_{jki}, representing a dichotomous outcome measured on individuals from provider j and neighborhood k. In contrast to other sections where we represent an outcome as Y_i, in this section we subscript using individual (i), provider (j), and neighborhood (k) to make explicit the level at which the covariates are measured. We generated Y_{jki} with a marginal regression model structure using a provider-level covariate $X_{1,j}$, a neighborhood-level covariate $X_{2,k}$, and an individual-level covariate $X_{3,jki}$,

$$\text{logit}P(Y_{jki} = 1|\mathbf{X}) = \beta_0 + \beta_1 X_{1,j} + \beta_2 X_{2,k} + \beta_3 X_{3,jki},$$

where $\beta_0 = -1.5$, $\beta_1 = \beta_2 = 0.5$, and $\beta_3 = 0.25$. Each binary covariate was generated from a Bernoulli distribution with a prevalence of 0.5. Correlation was induced using normally distributed random provider and neighborhood effects with standard deviations of 0.5 and 1.0 for scenario 1. Additional scenarios differed from the first scenario as follows: $J = 20$ and $K = 80$ to illustrate performance with a very small number of clusters for one dimension; correlation present in only one level of clustering (either within-provider only or within-neighborhood only); and smaller within-cluster dependence using random effects standard deviations half the size used for scenario 1.

We used marginalized random effects models to generate the data (14), which permit fixing the marginal logistic regression parameters while allowing us to vary the magnitude of unexplained variation across both providers and neighborhoods. We generated 5,000 simulated data sets and estimated marginal logistic regression parameters $\boldsymbol{\beta}$ using GEEs with working independence. For standard error estimation, we compared naïve estimates assuming independent data, GEE estimates clustering on p-ID only, GEE estimates clustering on n-ID only, and the proposed three-step correction for clustering on both neighborhood and provider. Because

TABLE 1. Results of simulations examining estimated regression coefficients, standard errors, and 95% coverage probabilities from the naïve (unadjusted) model and generalized estimating equation estimates that adjust for correlation within provider only, neighborhood only, and both provider and neighborhood*

Parameter†	Regression coefficients		SE‡ and 95% CI‡ coverage								
			Sampling SE	Naïve		Generalized estimating equation					
						Provider only		Neighborhood only		Both levels	
	Truth	Estimated		SE	95% CI coverage	SE	95% CI coverage	SE	95% CI coverage	SE	95% CI coverage
Scenario 1: $J = 30$, $K = 50$, $\sigma(j) = 0.5$, $\sigma(k) = 1.0$											
Intercept β_0	−1.50	−1.51	0.049	0.054	0.96	0.019	0.77	0.038	0.91	0.050	0.94
Provider β_1	0.50	0.50	0.028	0.047	0.99	0.030	0.95	0.0055	0.61	0.028	0.94
Neighborhood β_2	0.50	0.50	0.061	0.047	0.91	0.0056	0.44	0.065	0.95	0.063	0.94
Individual β_3	0.25	0.25	0.0055	0.046	1.00	0.0060	0.95	0.0058	0.95	0.0063	0.96
Scenario 2: $J = 20$, $K = 80$, $\sigma(j) = 0.5$, $\sigma(k) = 1.0$											
Intercept β_0	−1.50	−1.52	0.044	0.052	0.97	0.026	0.84	0.025	0.85	0.044	0.94
Provider β_1	0.50	0.51	0.040	0.044	0.96	0.043	0.94	0.0049	0.50	0.041	0.94
Neighborhood β_2	0.50	0.51	0.040	0.044	0.96	0.0054	0.51	0.041	0.95	0.040	0.95
Individual β_3	0.25	0.25	0.0049	0.044	1.00	0.0059	0.95	0.0052	0.95	0.0061	0.95
Scenario 3: $J = 30$, $K = 50$, $\sigma(j) = 0$, $\sigma(k) = 0.5$											
Intercept β_0	−1.50	−1.50	0.015	0.054	1.00	0.007	0.79	0.016	0.95	0.016	0.95
Provider β_1	0.50	0.50	0.005	0.047	1.00	0.006	0.95	0.0056	0.94	0.005	0.93
Neighborhood β_2	0.50	0.50	0.024	0.047	0.99	0.0059	0.65	0.025	0.95	0.025	0.95
Individual β_3	0.25	0.25	0.0052	0.046	1.00	0.0059	0.96	0.0056	0.95	0.0062	0.96
Scenario 4: $J = 30$, $K = 50$, $\sigma(j) = 0.5$, $\sigma(k) = 0$											
Intercept β_0	−1.50	−1.51	0.022	0.054	1.00	0.024	0.95	0.006	0.71	0.024	0.94
Provider β_1	0.50	0.50	0.036	0.047	0.97	0.039	0.95	0.0056	0.55	0.038	0.95
Neighborhood β_2	0.50	0.50	0.005	0.047	1.00	0.0059	0.95	0.005	0.94	0.005	0.93
Individual β_3	0.25	0.25	0.0054	0.047	1.00	0.0059	0.95	0.0057	0.95	0.0063	0.95
Scenario 5: $J = 30$, $K = 50$, $\sigma(j) = 0.25$, $\sigma(k) = 0.5$											
Intercept β_0	−1.50	−1.51	0.019	0.055	1.00	0.011	0.84	0.016	0.91	0.020	0.94
Provider β_1	0.50	0.50	0.012	0.048	1.00	0.014	0.95	0.0055	0.80	0.013	0.95
Neighborhood β_2	0.50	0.50	0.023	0.047	1.00	0.0058	0.66	0.024	0.95	0.024	0.95
Individual β_3	0.25	0.25	0.0053	0.047	1.00	0.0059	0.96	0.0056	0.95	0.0062	0.95

* Binary outcomes were generated from a logistic regression model including three binary covariates: a provider-level covariate, a neighborhood-level covariate, and an individual-level covariate.

† J is the number of providers, K is the number of neighborhoods, $\sigma(j)$ is the standard deviation of the provider-specific random effects, and $\sigma(k)$ is the standard deviation of the neighborhood-specific random effects.

‡ SE, standard error; CI, confidence interval.

we explore scenarios with relatively small numbers of clusters, we applied the simple scalar correction discussed in the previous section, multiplying the empirical variance estimator by $J/(J-p)$ and $K/(K-p)$ when using GEEs to cluster on p-ID and n-ID, respectively. For consistency, we also applied the correction when clustering on the combination of both p-ID and n-ID, although the size of this correction is quite small (e.g., $JK/(JK-p) = 1,500/1,496$ for scenario 1).

The simulation results show that inference for regression parameters can be quite inaccurate unless properly accounting for clustering (table 1). For example, in the initial scenario (scenario 1), the naïve standard error is estimated on average as 0.046 for β_3 (corresponding to the individual-level covariate), while any of the clustering corrections yield an estimated standard error approximately 10-fold smaller. The naïve estimate also yields standard errors for the coefficient of the neighborhood covariate, β_2, that are approximately 25 percent too small (0.047/0.061 = 77 percent) and therefore contain the true parameter value only 91 percent of the time rather than the nominal 95 percent coverage. Table 1 shows that inference regarding the provider-level covariate is approximately correct when clustering on provider, but

clustering only on provider makes inference on the neighborhood covariate grossly incorrect. The parallel is observed when clustering only on neighborhood, where the average variance estimated for the coefficient of the provider covariate, β_1, is 0.0056, while the true sampling variance is approximately fivefold greater (i.e., 0.0291). Scenario 1 shows that, for valid inference on person, provider, and neighborhood covariates, a method that accounts for both provider and neighborhood clustering is necessary.

The results of the simulations in table 1 suggest that the proposed standard error estimator provides approximately correct inference for a range of plausible scenarios. For example, confidence interval coverage of the proposed estimator is between 93 percent and 95 percent for the situations considered, while naïve methods that do not adjust standard errors or correct for only one level of clustering yield confidence interval coverage ranging from 42 percent (for the neighborhood covariate in scenario 3 where we cluster on provider only, but in reality correlation is purely within-neighborhood) to 100 percent for naïve estimates of individual-level effects.

APPLICATION

In this section, we demonstrate use of our proposed approach to adjust for correlation within multiple levels of clustering when estimating the influence of woman, radiologist, and facility factors on the positive predictive value (PPV) of screening mammography. Mammography outcomes are clustered within women, radiologists, and facilities, and these clusters are not necessarily nested, because women may go to different facilities and because mammograms taken on the same woman may be read by different radiologists. In addition, radiologists often interpret mammograms at more than one facility. Data were collected by the Breast Cancer Surveillance Consortium (http://breastscreening.cancer.gov), a National Cancer Institute-sponsored collaboration among seven population-based mammography registries in the United States (15). Each registry prospectively collects demographic, risk-factor, and clinical information each time a woman receives a mammogram at a participating facility. Mammography registries link to regional or state cancer registries and pathology databases to determine cancer status. This study includes mammograms from four of the seven Breast Cancer Surveillance Consortium mammography registries: Group Health Cooperative in western Washington, New Hampshire Mammography Network, San Francisco Mammography Registry, and the Vermont Breast Cancer Surveillance System. Each registry has approval from its institutional review board to collect these data for research purposes.

Our analytical goal is to estimate the influence of woman-, radiologist-, and facility-level characteristics on the PPV of screening mammography. We include three woman-level factors previously shown to influence mammography performance: age, mammographic breast density, and time since previous mammography (16, 17). In addition, we include the radiologist average annual volume of mammography interpretation and the rural/urban makeup of the facility's population of patients. Previous studies have found conflicting results for the effect of radiologist volume on interpretive performance (18–20), and determining the effects of reader volume on interpretive performance was listed as a priority in the recently released Institute of Medicine report on improving breast-imaging quality standards (21). No previous studies have examined differences in mammography performance by urban/rural location of the facility.

We offer the following statistical model. PPV is estimated from mammograms with a positive assessment (i.e., those recalled for additional workup). Thus, we included all mammograms given a BI-RADS (22) assessment of 0, 4, 5, or 3 with a recommendation for immediate follow-up. Let D_i represent breast cancer status for the ith mammogram, such that $D_i = 1$ if the mammogram was associated with a diagnosis of invasive carcinoma or ductal carcinoma in situ within 1 year; $i = 1, \ldots, N$. We model the marginal probability of breast cancer given a positive assessment as a function of a $p \times 1$ vector of covariates $\mathbf{X}_i$ using logistic regression (23):

$$\text{logit}P(D_i = 1|\mathbf{X}_i) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}.$$

We fit this logistic regression model assuming independent observations (naïve model) and using the proposed GEE strategy to adjust for correlation within women, radiologists, and facilities. All models are additionally adjusted for mammography registry; however, results by registry are not shown for confidentiality reasons.

RESULTS

Of the 584,784 screening mammograms performed from 1998 to 2002, 62,226 mammograms (10.6 percent) on 57,015 women were recalled for additional workup by 169 radiologists at 56 facilities (table 2). Among these recalled mammograms, 2,531 were associated with breast cancer within the 1 year, resulting in a crude PPV of 4.1 percent. The majority of women (92 percent) were recalled on only one mammogram, 8 percent were recalled on two separate examinations, and less than 1 percent were recalled on three or more examinations. Radiologists recalled 12–2,233 screening mammograms (median = 266) with 6–8,798 mammograms recalled from each facility (median = 695).

Based on the raw data (table 2), the PPV of screening mammography increases with age. PPV is lower for women without a previous mammogram and for women with denser breasts. PPV is highest for radiologists who interpreted an average of more than 4,000 mammograms per year. There are no clear trends between PPV and urban/rural makeup of the facility's population of patients.

Table 3 displays the influence of covariates on the odds of breast cancer given a positive assessment, adjusted for other covariates in the table and mammography registry, along with 95 percent confidence intervals estimated from the naïve (unadjusted) model and models adjusting for correlation within women only, radiologists only, facilities only, and all three levels of clustering. As expected, the odds of having breast cancer following a positive mammogram are significantly higher among older women and among women screened 3 or more years prior relative to those screened more recently. The PPV is higher for women with denser

TABLE 2. Characteristics of women, radiologists, and facilities and raw (unadjusted) estimates of positive predictive value*

	Total no.	%	No. recalled	Recall rate†	No. of cancers among recalled	Positive predictive value
Total no.	587,784		62,226	10.6	2,531	4.1
Woman's age (years)						
40–44	92,872	15.8	12,014	12.9	181	1.5
45–49	102,334	17.4	12,399	12.1	307	2.5
50–54	110,654	18.8	12,162	11.0	394	3.2
55–59	82,426	14.0	8,380	10.2	391	4.7
60–69	116,503	19.8	10,596	9.1	684	6.5
70–79	82,995	14.1	6,675	8.0	574	8.6
Woman's mammographic breast density						
Almost entirely fat	44,118	7.5	2,102	4.8	99	4.7
Scattered fibroglandular densities	264,715	45.0	23,955	9.0	1,084	4.5
Heterogeneously dense	235,611	40.1	30,689	13.0	1,174	3.8
Dense	43,340	7.4	5,480	12.6	174	3.2
Woman's time since previous mammography						
No previous mammography	33,180	5.6	5,792	17.5	174	3.0
1–2 years	490,347	83.4	47,863	9.8	1,981	4.1
≥3 years	64,257	10.9	8,571	13.3	376	4.4
Radiologist's average annual volume of mammogram interpretations						
481–750	15,350	2.6	1,378	9.0	57	4.1
751–1,000	34,282	5.8	3,961	11.6	151	3.8
1,001–1,500	165,457	28.2	19,614	11.9	702	3.6
1,501–2,500	208,343	35.5	19,989	9.6	854	4.3
2,501–4,000	76,026	12.9	9,468	12.5	317	3.3
>4,000	88,326	15.0	7,816	8.8	450	5.8
% of facility's patients that live in rural area						
0–24	213,787	36.4	22,862	10.7	1,048	4.6
25–49	152,294	25.9	17,913	11.8	648	3.6
50–74	169,816	28.9	17,332	10.2	644	3.7
75–100	51,887	8.8	4,119	7.9	191	4.6

* The data used to illustrate the authors' proposed analytical approach are from the Breast Cancer Surveillance Consortium begun in 1994 (http://breastscreening.cancer.gov).

† Number recalled per 100 mammograms.

breasts, but this is not statistically significant. Although women with denser breasts are at higher risk of breast cancer (24), mammography is also less accurate in these women (16, 25), which would lower the PPV. Radiologists who interpret more than 4,000 mammograms a year have significantly higher PPV relative to most other volume groups; however, in general, the relation between interpretive volume and PPV is not monotonic. PPV is lower for facilities with less than 50 percent of their patients from rural areas, but this is not statistically significant after adjusting for correlation within clusters.

In general, if covariates vary only between clusters, variance estimates will be larger after adjusting for correlation within that cluster. For example, when adjusting for correlation within radiologists, we found that the variance estimates for radiologist volume increase compared with the naïve model. A similar pattern can be seen for the urban/rural makeup of the facility when adjusting for correlation within facilities. In contrast, if a covariate varies only within clusters (all clusters have outcomes measured for all levels of the covariate such as in cross-over trials), the corresponding variance estimate will decrease after adjustment. There are no covariates that vary only within clusters in this study. When covariates vary both within and between clusters, as is the case here when considering all three levels of clustering, it is difficult to determine a priori if adjustment for

TABLE 3. Influence of covariates on the odds of disease given a positive mammogram and 95% confidence intervals from the naïve (unadjusted) model and generalized estimating equation models that adjust for correlation within women only, radiologists only, facilities only, and all clusters (women, radiologists, and facilities)*

	Odds ratio†	95% confidence intervals				
		Naïve	Generalized estimating equation			
			Woman only	Radiologist only	Facility only	Woman, radiologist, and facility
Woman's age (years)						
40–44	0.16	0.16	0.13, 0.19	0.13, 0.19	0.13, 0.19	0.13, 0.18
45–49	0.27	0.27	0.23, 0.31	0.23, 0.31	0.23, 0.31	0.23, 0.31
50–54	0.36	0.36	0.31, 0.41	0.31, 0.41	0.30, 0.42	0.30, 0.43
55–59	0.53	0.53	0.46, 0.60	0.46, 0.60	0.45, 0.61	0.46, 0.60
60–69	0.74	0.74	0.66, 0.83	0.66, 0.83	0.65, 0.83	0.63, 0.86
70–79				Referent		
Woman's mammographic breast density						
Almost entirely fat	0.82	0.63, 1.06	0.64, 1.06	0.63, 1.07	0.66, 1.02	0.65, 1.03
Scattered densities	0.89	0.75, 1.05	0.75, 1.05	0.73, 1.08	0.74, 1.07	0.75, 1.06
Heterogeneously dense	0.91	0.77, 1.08	0.77, 1.08	0.74, 1.12	0.77, 1.08	0.77, 1.08
Dense				Referent		
Woman's time since previous mammography						
No previous mammography	0.85	0.70, 1.02	0.70, 1.03	0.68, 1.07	0.67, 1.08	0.67, 1.08
1–2 years	0.77	0.68, 0.86	0.68, 0.86	0.68, 0.86	0.67, 0.88	0.67, 0.88
≥3 years				Referent		
Radiologist's average annual volume of mammogram interpretations						
481–750	0.64	0.48, 0.86	0.48, 0.86	0.42, 0.97	0.40, 1.02	0.41, 1.00
751–1,000	0.60	0.49, 0.74	0.49, 0.74	0.42, 0.86	0.40, 0.91	0.39, 0.92
1,001–1,500	0.60	0.51, 0.70	0.51, 0.70	0.45, 0.80	0.42, 0.86	0.42, 0.87
1,501–2,500	0.73	0.63, 0.85	0.63, 0.85	0.56, 0.97	0.53, 1.02	0.53, 1.02
2,501–4,000	0.61	0.51, 0.72	0.51, 0.72	0.46, 0.81	0.41, 0.89	0.42, 0.89
>4,000				Referent		
% of facility's patients that live in rural area						
0–24	0.79	0.64, 0.97	0.64, 0.97	0.57, 1.08	0.54, 1.13	0.54, 1.15
25–49	0.71	0.59, 0.86	0.59, 0.86	0.53, 0.97	0.51, 1.00	0.50, 1.01
50–74	0.92	0.78, 1.10	0.78, 1.10	0.71, 1.21	0.65, 1.32	0.65, 1.32
75–100				Referent		

* The data used to illustrate the authors' proposed analytical approach are from the Breast Cancer Surveillance Consortium begun in 1994 (http://breastscreening.cancer.gov).

† Odds ratios are adjusted for the other covariates in the table plus mammography registry.

correlation will result in smaller or larger variance estimates. However, comparing the confidence interval widths in table 3 reveals some clear patterns. Adjusting for correlation within women had little to no effect on the confidence intervals. Perhaps this is because there are few observations per woman, with most women contributing only a single observation. Adjusting for correlation within radiologists and facilities had a small and inconsistent effect on the woman-level covariates but considerably widened the confidence intervals for radiologist and facility factors, which tend to vary more between than within radiologists and facilities. For example, the odds ratio comparing radiologists with an annual interpretive volume of 2,501–4,000 with those with a volume of greater than 4,000 has a confidence interval of 0.51, 0.72 using naïve methods. Clustering only on radiologist yields a 1.7-fold wider confidence interval of

0.46, 0.81, while clustering on woman, radiologist, and facility yields a 2.2-fold wider confidence interval of 0.42, 0.89. Similarly, for the urban/rural makeup of the facility, the odds ratio comparing facilities with less than 25 percent rural patients compared with those with 75 percent or more rural patients has a confidence interval of 0.64, 0.97 using naïve methods. Clustering on all three levels yields a 1.8-fold wider confidence interval of 0.54, 1.15, which includes the value of 1.0.

DISCUSSION

In this paper, we describe a simple GEE method that accounts for correlation within multiple nonnested clusters when fitting marginal generalized linear models. The approach relies on a working independence assumption coupled with a multistep method for obtaining empirical standard errors using standard software. These corrected standard errors may be used to calculate confidence intervals and to conduct Wald hypothesis tests. Sample SAS and STATA code for implementing the GEE approach with two levels of clustering is included in Appendix A. Extensions to three clusters are straightforward.

The proposed GEE method relies on empirical covariance estimates. When the number of clusters is small (<50 clusters), these estimates may be biased. In our simulations, we explored the use of a simple approach that scales the variance estimates to correct for this small-sample bias. This correction has been found to perform well (13) and is easy to apply when using our proposed approach for nonnested clusters.

Both the simulations and the application demonstrate the importance of accounting for the correlation within all levels of clustering for proper inference. Accounting for clustering within only one level could lead to biased variance estimates for factors measured at other levels and possibly even factors measured at the level of clustering accounted for, depending on how the factor varies within and between the other unacknowledged clusters. When covariates vary both within and between clusters, as is typically the case when there are multiple levels of clustering, it is difficult to determine a priori how adjustment for correlation within clusters will influence variance estimates. In some cases, such as for the woman-level factors and the accuracy parameters estimated in the application, adjustment may not lead to important differences, while in other cases, factors found to be highly significant without adjustment may lose statistical significance after proper adjustment for clustering. This was found in the application when testing the influence of the facility's urban/rural makeup on PPV. Given this potential for biased inference, it is important to report variance estimates and hypothesis tests that adjust for all levels of clustering or to conduct a sensitivity analysis to ensure that results are not biased by unacknowledged clustering within the data.

ACKNOWLEDGMENTS

This work was supported by National Cancer Institute-funded cooperative agreement U01CA86076 and by National Heart, Lung, and Blood Institute grant HL72966. Data collection for the application was supported by National Cancer Institute-funded Breast Cancer Surveillance Consortium (http://breastscreening.cancer.gov) cooperative agreement (U01CA63740, U01CA86076, U01CA86082, U01CA63736, U01CA70013, U01CA69976, U01CA63731, U01CA70040).

Conflict of interest: none declared.

REFERENCES

1. Heagerty PJ, Zeger SL. Marginalized multilevel models and likelihood inference. Stat Sci 2000;15:1–26.
2. Diggle PJ, Heagerty PJ, Liang KY, et al. The analysis of longitudinal data. 2nd ed. Oxford, United Kingdom: Oxford University Press, 2002.
3. Ritz J, Spiegelman D. Equivalence of conditional and marginal regression models for clustered and longitudinal data. Stat Methods Med Res 2004;13:309–23.
4. Wang Z, Louis TA. Marginalized binary mixed-effects models with covariate-dependent random effects and likelihood inference. Biometrics 2004;60:884–91.
5. Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. Biometrics 1988; 44:1049–60.
6. Neuhaus JM, Kalbfleisch JD, Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. Int Stat Rev 1991;59: 25–35.
7. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–22.
8. Miglioretti DL, Heagerty PJ. Marginal modeling of multilevel binary data with time varying covariates. Biostatistics 2004;5: 381–98.
9. McCullagh P, Nelder JA. Generalized linear models. 2nd ed. London, United Kingdom: Chapman and Hall, 1989.
10. Betensky RA, Talcott JA, Weeks JC. Binary data with two, non-nested sources of clustering: an analysis of physician recommendations for early prostate cancer treatment. Biostatistics 2000;1:219–30.
11. Sharples K, Breslow N. Regression analysis of correlated binary data: some small sample results for the estimating equation approach. J Stat Comput Simul 1992;42:1–20.
12. Emrich LJ, Piedmonte MR. On some small sample properties of generalized estimating equation estimates for multivariate dichotomous outcomes. J Stat Comput Simul 1992;41: 19–29.
13. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. Biometrics 2001;57: 126–34.
14. Heagerty PJ. Marginally specified logistic-normal models for longitudinal binary data. Biometrics 1999;55:688–98.
15. Ballard-Barbash R, Taplin SH, Yankaskas BC, et al. Breast Cancer Surveillance Consortium: a national mammography screening and outcomes database. AJR Am J Roentgenol 1997;169:1001–8.
16. Carney PA, Miglioretti DL, Yankaskas BC, et al. Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. Ann Intern Med 2003;138:168–75.
17. Yankaskas BC, Taplin SH, Ichikawa L, et al. Association between mammography timing and measures of screening

performance in the United States. Radiology 2005;234:363–73.
18. Barlow WE, Chi C, Carney PA, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. J Natl Cancer Inst 2004;96:1840–50.
19. Smith-Bindman R, Chu P, Miglioretti DL, et al. Physician predictors of mammographic accuracy. J Natl Cancer Inst 2005;97:358–67.
20. Beam CA, Conant EF, Sickles EA. Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. J Natl Cancer Inst 2003;95:282–90.
21. Institute of Medicine. Improving breast imaging quality standards. Washington, DC: The National Academies Press, 2005.
22. American College of Radiology. American College of Radiology (ACR) breast imaging reporting and data system atlas (BI-RADS atlas). Reston, VA: American College of Radiology, 2003.
23. Pepe MS. The statistical evaluation of medical tests for classification and prediction. New York, NY: Oxford University Press, 2003.
24. Boyd NF, Rommens JM, Vogt K, et al. Mammographic breast density as an intermediate phenotype for breast cancer. Lancet Oncol 2005;6:798–808.
25. Buist DS, Porter PL, Lehman C, et al. Factors contributing to mammography failure in women aged 40–49 years. J Natl Cancer Inst 2004;96:1432–40.
26. Mayer-Hamblett N, Self S. A regression modeling approach for describing patterns of HIV genetic variation. Biometrics 2001;57:449–60.
27. Lumley T, Mayer Hamblett N. Asymptotics for marginal generalized linear models with sparse correlations. Seattle, WA: University of Washington, 2003. University of Washington Biostatistics Working Paper Series. (Working paper 207). (http://www.bepress.com/uwbiostat/paper207).

APPENDIX A

Below is sample SAS code to fit a logistic regression model to a binary outcome *Y*, adjusting for the correlation within two nonnested clusters using GEEs. Note that the clustering variable C1C2ID may be created by concatenating C1ID and C2ID in a data step using the command "C1C2ID = C1ID || '|' || C2ID;".

```
%macro gee(n=1,cluster=C1);
  proc genmod data=a descending;
    class &cluster;
    model y = x/dist=binomial;
    repeated subject=&cluster/type=indep ecovb;
    ods output GEEEmpPEst=beta GEERCov=V&n;
  quit;
%mend;
%gee(n=1,cluster=C1ID);
%gee(n=2,cluster=C2ID);
%gee(n=3,cluster=C1C2ID);
```

The covariance matrices may be read into PROC IML to combine and to calculate the corrected standard errors for the regression coefficients:

```
proc iml;
  use V1; read all var{rowname}; read all var(rowname) into V1; close V1;
  use V2; read all var(rowname) into V2; close V2;
  use V3; read all var(rowname) into V3; close V3;
  V=V1+V2−V3; SE=sqrt(vecdiag(V)); print SE;
```

Below is the corresponding STATA code to fit the same model. Note that the standard errors from STATA will be slightly larger than those from SAS, because STATA scales the robust covariance matrix to improve coverage probabilities (http://www.stata.com/support/faqs/stat/gee.html). For logistic regression models, STATA multiples the covariance matrix by $K/(K-1)$, where K is the number of clusters.

```
xtgee y x, family(bin) link(logit) corr(ind) robust i(C1ID)
mat V1=e(V)
xtgee y x, family(bin) link(logit) corr(ind) robust i(C2ID)
mat V2=e(V)
xtgee y x, family(bin) link(logit) corr(ind) robust i(C1C2ID)
mat V3=e(V)
mat V=V1+V2−V3
```

Other matrix functions may be performed on the covariance matrix **V** using STATA. For example, the vector of estimated variances can be obtained by selecting the diagonal elements of the covariance matrix using the command "mat var=vecdiag(V)" and printed using the command "matlist var."

APPENDIX B

The GEE approach to more than two levels of clustering is described below. GEE using working independence solves the estimating equation

$$\sum_{i=1}^{N} D_i^T V_i^{-1} (Y_i - \mu_i) = 0,$$

where $D_i = \partial\mu_i/\partial\beta$, and $V_i = \text{variance}(Y_i|\mathbf{X}_i)$. Based on results of Mayer Hamblett and Self (26) and Lumley and Mayer Hamblett (27), the solution to the estimating equations, $\hat{\beta}$, has an asymptotic variance given as

$$\text{variance}(\hat{\beta}) = A^{-1}BA^{-1}$$

$$A = \sum_{i=1}^{N} D_i^T V_i^{-1} D_i$$

$$B = \text{variance}\left(\sum_{i=1}^{N} U_i\right),$$

where $U_i = D_i^T V_i^{-1}(Y_i - \mu_i)$. A consistent estimate of B can be obtained using

$$\hat{B} = \sum_{i=1}^{N}\sum_{j=1}^{N} \delta(i,j) \cdot U_i U_j^T,$$

where $\delta(i, j) = 1$ if observations Y_i and Y_j share C_1, C_2, or C_3. The indicator $\delta(i, j)$ can be viewed as a logical "or" operator that captures the *same cluster 1* or the *same cluster 2* or the *same cluster 3*, and as such can be represented as:

$$\delta(i,j) = \delta_1(i,j) + \delta_2(i,j) + \delta_3(i,j) - \delta_1(i,j)\delta_2(i,j) - \delta_1(i,j)\delta_3(i,j) - \delta_2(i,j)\delta_3(i,j) + \delta_1(i,j)\delta_2(i,j)\delta_3(i,j),$$

where $\delta_k(i, j) = 1$ if both Y_i and Y_j come from the same cluster C_k and 0 otherwise.

This representation shows that the estimate $\hat{B}$ can be formed from seven contributions:

$$\hat{B} = \hat{B}_1 + \hat{B}_2 + \hat{B}_3 - \hat{B}_{12} - \hat{B}_{13} - \hat{B}_{23} + \hat{B}_{123}$$

$$\hat{B}_k = \sum_{i=1}^{N}\sum_{j=1}^{N} \delta_k(i,j) U_i U_j^T; \quad k = 1,2,3$$

$$\hat{B}_{12} = \sum_{i=1}^{N}\sum_{j=1}^{N} \delta_1(i,j)\delta_2(i,j) U_i U_j^T$$

$$\hat{B}_{13} = \sum_{i=1}^{N}\sum_{j=1}^{N} \delta_1(i,j)\delta_3(i,j) U_i U_j^T$$

$$\hat{B}_{23} = \sum_{i=1}^{N}\sum_{j=1}^{N} \delta_2(i,j)\delta_3(i,j) U_i U_j^T$$

$$\hat{B}_{123} = \sum_{i=1}^{N}\sum_{j=1}^{N} \delta_1(i,j)\delta_2(i,j)\delta_3(i,j) U_i U_j^T.$$

By use of working independence, the final estimated variance for $\hat{\beta}$ is simply a linear combination of variance estimates produced by GEEs:

$$
\begin{aligned}
\mathrm{var}(\hat{\beta}) &= A^{-1}BA^{-1} \\
&= (A^{-1}B_1A^{-1}) + (A^{-1}B_2A^{-1}) + (A^{-1}B_3A^{-1}) \\
&\quad - (A^{-1}B_{12}A^{-1}) - (A^{-1}B_{13}A^{-1}) - (A^{-1}B_{23}A^{-1}) \\
&\quad + (A^{-1}B_{123}A^{-1}) \\
&= V^1 + V^2 + V^3 - V^{12} - V^{13} - V^{23} + V^{123},
\end{aligned}
$$

where V^k is the estimated variance from a working independence GEE clustering on C_k. V^{kl} clusters on unique combinations of C_k and C_l, and V^{123} clusters on unique combinations of C_1, C_2, and C_3.