Journal of Data Science **5**(2007), 269-288

Missing Information as a Diagnostic Tool for Latent Class Analysis

Ofer Harel[1] and Diana Miglioretti[2]
[1] *University of Connecticut and* [2] *Group Health Cooperative*

Abstract: Latent class analysis (LCA) is a popular method for analyzing multiple categorical outcomes. Given the potential for LCA model assumptions to influence inference, model diagnostics are a particulary important part of LCA. We suggest using the rate of missing information as an additional diagnostic tool. The rate of missing information gives an indication of the amount of information missing as a result of observing multiple surrogates in place of the underlying latent variable of interest and provides a measure of how confident one can be in the model results. Simulation studies and real data examples are presented to explore the usefulness of the proposed measure.

Key words: Latent class, missing data, missing information, multiple imputation.

1. Introduction

Latent class analysis (LCA) has a long history in the social and behavioral sciences (Lazarsfeld, 1950; Lazarsfeld and Henry, 1968; McCutcheon, 1987; Clogg, 1995) and has gained considerable attention in biostatistics over the past two decades (Garrett, Eaton and Zeger, 2002; Garett and Zeger, 2000; Bandeen-Roche *et al.*, 1997; Formann, 1996). In general, LCA is used to explain relationships among multiple categorical variables. Specifically, LCA may be used to describe the prevalence and symptomatology of a mental disorder or health status that is measured via multiple indicators or to explore subgroups of the disorder or disease (Storr *et al.*, 2004; Moran *et el.*, 2004; Nestadt *et al.*, 2003; Fergusson *et al.*, 1995; Eaton and Bohrnstedt, 1989). In medical diagnostics, LCA may be used to measure the sensitivity and specificity of diagnostic tests in the absence of a gold standard (Garrett *et al.*, 2002, Formann, 1996; Butler *et al.*, 2003) or to develop or evaluate diagnostic criterion (Fossati *et al.*, 2001; Young *et al.*, 1983; Young, 1982). More recently, latent class models have been extended to regression settings. Latent class and latent transition regression have been proposed for quantifying the association between risk factors and latent health status when multiple surrogates are collected in lieu of a single adequate measure of health status (Miglioretti, 2003; Humphreys and Janson, 2000; Bandeen-Roche *et al.*, 1997; Dayton and Macready, 1988). Growth mixture modeling has been proposed for identifying and describing subgroups of individuals with different longitudinal trajectories (Muthen *et al.*, 2002; Muthen, 2004). Latent class survival models have been proposed for modeling time-to-event data (Rosen and Tanner, 1999; Lin *et al.*, 2002).

Given the potential for LCA model assumptions to influence inference, numerous model diagnostic methods exist. Numerical model checking statistics have been proposed for evaluating goodness-of-fit (Reiser and Lin, 1999; Formann, 1996; Collins *et al.*, 1993; Dayton and Macready, 1988; Hagenaars, 1988; Goodman, 1974), lack-of-fit (Rudas, Clogg and Linday, 1994), and identifiability (Bandeen-Roche *et al.*, 1997; Goodman, 1974; McHugh, 1956, 1958). Graphical displays have been proposed for evaluating model assumptions and goodness-of-fit (Miglioretti, 2003; Garrett and Zeger, 2000; Bandeen-Roche *et al.*, 1997) and checking for weak identifiability of model parameters (Garrett and Zeger, 2000).

In this paper, we propose a complementary model diagnostic measure, the "rate of missing information," which provides insight into the value of surrogates in measuring the latent variable of interest and the usefulness of the fitted latent class model. This measure may also be used to guide the design of future studies. The concept of information was introduced to statistics by Fisher in the 1920s. In the statistical sense, information refers to the amount of information in the sample about the population parameters of interest. For incomplete data sets, the amount of missing information can be estimated by the difference between the hypothetical information given complete data and the observed information in the incomplete data. The rate of missing information (Rubin, 1987) is the proportion of the missing information over the complete data information and provides a measure of how not observing the missing data contributes to uncertainty about the population parameters of interest. For LCA, this corresponds to the rate of information missing due to measurement error associated with the observed outcomes being surrogates of the underlying latent variable.

In LCA, the latent class memberships can be considered missing data, and the rate of missing information can be easily obtained when multiple imputation (Rubin, 1987; Schafer, 1997) with data augmentation (Tanner and Wong, 1987) is used for model estimation. In the LCA setting, the rate of missing information provides a measure of how observing surrogates in place of the latent variable of interest contributes to uncertainty about the parameters. We use simulations to explore how this measure depends on sample size, number of observed items, class size, class-specific item prevalences, and number of classes. In addition, we examine how it relates to bias and variability of the parameter estimates and the ability to accurately estimate true class membership.

We begin our article with a LCA review and introduction of the rate of missing information in this context. Next, we provide simulation results. We then apply these methods to real data examples. We end with a discussion of the proposed measure.

2. Methods

2.1 Latent class anlaysis

Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})'$ denote a vector of binary indicators for the ith individual; $i = 1, \ldots, N$; and J observed measures (e.g., diagnostic tests, health indicators, or some other surrogates of the underlying latent variable of interest). For simplicity, we consider the case of binary observed measures. Extension to categorical outcomes is straightforward. The basic idea of LCA is that association among $\mathbf{Y}_i$ arises because the study population is comprised of a mixture of subpopulations or classes (e.g., diseased and not diseased individuals in the medical diagnostics context). Let $S_i \in \{1, \ldots, K\}$ indicate latent class membership for ith individual and $\gamma_k = P(S_i = k)$ represent the prevalence of class k. There are two basic assumptions in LCA. First, individuals have common response probabilities within a class $k : \rho_{jk} = P(Y_{ij} = 1 | S_i = k)$. Second, observed responses $\mathbf{y}_i$ are independent given class membership S_i : $P(Y_{i1} = y_{i1}, \ldots, Y_{iJ} = y_{iJ} | S_i = k) = \prod_{j=1}^{J} P(Y_{ij} = y_{ij} | S_i = k)$. Given these two assumptions, the observed data likelihood may be expressed as

$$\prod_{i=1}^{N} \sum_{k=1}^{K} \gamma_k \prod_{j=1}^{J} \rho_{jk}^{y_{ij}} (1 - \rho_{jk})^{1-y_{ij}} .$$

Latent class regression extends the traditional latent class model to allow the probability of class membership γ_k to depend on a $1 \times P$ vector of covariates $\mathbf{x}_i$ (Dayton and Macready, 1988; Bandeen-Roche *et al.*, 1997) via polytomous regression:

$$\prod_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik} (\mathbf{x}_i, \beta) \prod_{j=1}^{J} \rho_{jk}^{y_{ij}} (1 - \rho_{jk})^{1-y_{ij}} .$$

Because the latent classes are not necessarily ordered, this relationship is typically modeled using a generalized logit link:

$$\log \left(\frac{\gamma_{ik} (\mathbf{x}_i, \boldsymbol{\beta}_k)}{\gamma_{iK} (\mathbf{x}_i, \boldsymbol{\beta}_K)} \right) = \mathbf{x}_i' \boldsymbol{\beta}_k$$

where $\boldsymbol{\beta}_K = 0$ for identifiability. Each latent class has a unique set of regression parameters, with the parameters for the reference class (here, the last class) set equal to zero for identifiability.

2.2 Rate of Missing Information

Latent classes can be viewed as variables that are missing with probability one; therefore, missing data methodology may be used to fit LCA models. For example, the EM algorithm has been long used to estimate LCA parameters (Goodman, 1978). When data are missing with probability one, it implies the data are missing completely at random (MCAR) (Rubin, 1987) i.e., the missingness (the missing data indicators or the process that causes the missing values) does not depend on any variable in the study. The common missing at random (MAR) assumption (Rubin, 1987), for which the missingness may depend on observed data but not on missing data, is a more general assumption that is implied by the MCAR assumption. In this case one can assume ignorability, and refrain from modelling the missingness (Schafer and Graham, 2002; Harel and Zhou, 2007).

Multiple imputation (MI) (Rubin, 1987, 1996; Schafer, 1997; Schafer and Graham, 2002; Harel and Zhou, 2007) is a simulation-based technique to deal with missing values. Generally speaking, each missing value is replaced with a set of $m > 1$ plausible values, resulting in m sets of complete data which differ only in the imputed values. Analyzing each of the complete data sets, using a complete-data methodology, and saving the estimates and standard errors of each set results in m sets of estimates and standard errors. Combining the results using Rubin's simple arithmetic rules produces a final result that takes into account the uncertainty in the data and the uncertainty due to the missing values. When using MI, the rate of missing information due to missing values (unobserved class memberships) may be easily estimated (Rubin, 1987).

We assume a joint model for the complete data $\mathbf{Y}_{com} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$ and the missingness $\mathbf{M}$, where $\mathbf{Y}_{obs}$ are the observed binary indicators, $\mathbf{Y}_{mis}$ are the missing latent class memberships, and $\mathbf{M}$ is the set of missingness indicators that separate the complete data into the observed and missing parts. To apply MI in the LCA context, m independent versions of the latent class memberships, $\mathbf{Y}_{mis}^{(1)}, \ldots, \mathbf{Y}_{mis}^{(m)}$, are imputed from $P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \mathbf{M})$. Because the latent class memberships are MCAR, we can ignore the missingness model $\mathbf{M}$ and impute from $P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs})$. Next, the m sets of original data with imputed class assignment are separately analyzed. Finally, the resulting m sets of point estimates and standard errors are combined using Rubin's (1987) rules, described below.

Let $\mathbf{Q}$ represent the set of LCA parameters where $\mathbf{Q} = (\gamma, \rho)$ is a $JK + K - 1$ dimensional vector for standard LCA and $\mathbf{Q} = (\beta, \rho)$ is a $JK + (K-1)P$ dimensional vector in the latent class regression setting. Let $\hat{\mathbf{Q}} = \hat{\mathbf{Q}}\left(\mathbf{Y}_{obs}, \mathbf{Y}_{mis}\right)$ denote the estimate for $\mathbf{Q}$ if the complete data were available and $\mathbf{U} = \mathbf{U}\left(\mathbf{Y}_{obs}, \mathbf{Y}_{mis}\right)$ denote its variance estimate. We assume that with complete data, each parameter estimate $\hat{Q}_q$ follows a normal distribution

$$(\hat{Q}_q - Q_q)/\sqrt{U_q} \sim Normal(0,1). \tag{2.1}$$

In the absence of $\mathbf{Y}_{mis}$, $\mathbf{Y}_{mis}^{(1)}, \ldots, \mathbf{Y}_{mis}^{(m)}$ are random versions from which the imputed-data estimates $\hat{\mathbf{Q}}^{(l)} = \hat{\mathbf{Q}}(\mathbf{Y}_{obs}, \mathbf{Y}_{mis}^{(l)})$ and their estimated variances $\mathbf{U}^{(l)} = \mathbf{U}(\mathbf{Y}_{obs}, \mathbf{Y}_{mis}^{(l)})$ are calculated, $l = 1, \ldots, m$. The overall estimate of Q_q is $\bar{Q}_q = m^{-1} \sum \hat{Q}_q^{(l)}$. The estimated total variance for $\bar{Q}_q$ is $T\,_q = (1+m^{-1})B_q + \bar{U}_q$ where $\bar{U}_q = m^{-1} \sum U_q^{(l)}$ is the estimated complete-data variance and $B_q = (m-1)^{-1} \sum (\hat{Q}_q^{(l)} - \bar{Q}_q)^2$ is the between-imputation variance. Tests and confidence intervals may be based on a Student's t approximation

$$(\bar{Q}_q - Q_q)/\sqrt{T_q} \sim t_\nu \tag{2.2}$$

with $\nu^{-1} = \frac{1}{(m-1)} \left[\frac{(1+m^{-1})B_q}{T_q}\right]^2$ degrees of freedom.

If $\mathbf{Y}_{mis}$ carries no information about Q_q, the imputed-data estimates $\hat{Q}_q^{(l)}$ would be identical and T_q would reduce to $\bar{U}_q$. Therefore, an estimate of the rate of missing information due to not observing Y_{mis}, *i.e.*, the rate of missing information, is

$$\hat{\lambda}_q = \frac{\bar{U}_q^{-1} - (T_q)^{-1}}{\bar{U}_q^{-1}} = \frac{(1+1/m)B_q}{\bar{U}_q + (1+1/m)B_q} = \frac{\hat{r}_q}{1+\hat{r}_q} \tag{2.3}$$

where $\hat{r}_q = (1+1/m)B_q/\bar{U}_q$. In LCA, this measure represents the rate of information missing due to the lack of knowledge about the individual class memberships, which can be translated to the model measurement error due to using the observed surrogates in lieu of the latent class memberships. If class memberships were observed, there would be no missing information, and hence the measurement error would be eliminated.

For LCA, the unobserved class memberships Y_{mis} can easily be imputed based on the posterior probabilities of class membership given the observed data and parameter estimates. To fully incorporate the variability of the estimated parameters, the posterior probabilities of class memberships may be calculated for each imputation by drawing parameter values from their posterior distribution as in data augmentation (Lanza *et al.*, 2005) or by drawing values from a multivariate normal distribution with mean and covariance equal to the maximum likelihood estimates. Given the imputed class memberships $\mathbf{Y}_{mis}^{(1)}, \ldots, \mathbf{Y}_{mis}^{(m)}$, the Binomial distribution may be used to estimate the model parameters given the complete data $\hat{\mathbf{Q}}^{(1)}, \ldots, \hat{\mathbf{Q}}^{(m)}$ and the corresponding complete data variances $\hat{\mathbf{U}}^{(1)}, \ldots, \hat{\mathbf{U}}^{(m)}$. For example, $\gamma_k = P(S_i = k)$ may be estimated as the proportion of subjects in each latent class n_k/N with variance given by $\hat{\gamma}_k(1 - \hat{\gamma}_k)/N$. The class specific item prevalences ρ_{jk} may be estimated as the number of subjects in class k with the item j equal to 1 with variance $\hat{\rho}_{jk}(1 - \hat{\rho}_{jk})/n_k$. For latent class regression (LCR), the regression coefficients and their variances given the imputed class memberships $\mathbf{Y}_{mis}^{(1)}, \ldots, \mathbf{Y}_{mis}^{(m)}$, may be estimated using standard complete-data polytomous regression.

3. Simulations

To explore the traits of the rate of missing information in the latent class setting we conducted a simulation study. We focus on two class models, because it is easier to manipulate the parameters in a systematic way to study the resulting behavior. We first explore the traditional LCA case, and then look at the LCR settings.

3.1 Latent class analysis

For our first simulation in the LCA settings, we generated 100 simulated data sets from 32 models with two latent classes as follows: The prevalence of class 1, γ_1, was set to 0.6 or 0.8 and the response probabilities given class membership, ρ, were set to 0.10, 0.15, 0.20, or 0.25 for class 1 and 0.90, 0.85, 0.80, or 0.75 for class 2. Data were generated from models with four and five items with sample sizes of 100 and 1000. We imputed 100 sets of class memberships from the posterior probabilities of class membership given the observed data after sampling 100 sets of parameter values from a multivariate normal distribution with mean and variance estimated from the LCA model and calculated the rates of missing information as described in the methods section. The mean rates of missing information across the 100 simulated data sets are summarized in Table 1. There was very little variation in the rates of missing information across ρ values within the same class, so the mean value across the four or five ρ values are presented for simplicity.

The most notable changes in the rate of missing information occurs with changes in the response probabilities. Response probabilities near 0.5 indicate a higher degree of measurement error, which is reflected in the dramatic increase in the rate of missing information as ρ moves from 0.10 and 0.90 towards 0.25 and 0.75 for classes 1 and 2, respectively. The rate of missing information is also a function of the number of items, with lower values for models with 5 items compared to those with only 4 items. The rate of missing information is lower

Table 1: Mean rates of missing information from 100 simulated LCA data sets under 32 conditions.

					Rate of missing information					
Number	Parameter Value				$N = 100$			$N = 1000$		
of items	γ_1	γ_2	$\boldsymbol{\rho_1}$	$\boldsymbol{\rho_2}$	γ	$\boldsymbol{\rho_1}$	$\boldsymbol{\rho_2}$	γ	$\boldsymbol{\rho_1}$	$\boldsymbol{\rho_2}$
4	0.6	0.4	0.25	0.75	0.76	0.50	0.58	0.83	0.55	0.63
4	0.6	0.4	0.20	0.80	0.57	0.39	0.49	0.60	0.40	0.48
4	0.6	0.4	0.15	0.85	0.33	0.29	0.36	0.32	0.27	0.34
4	0.6	0.4	0.10	0.90	0.12	0.17	0.22	0.12	0.16	0.21
4	0.8	0.2	0.25	0.75	0.68	0.35	0.58	0.87	0.47	0.74
4	0.8	0.2	0.20	0.80	0.54	0.29	0.52	0.66	0.33	0.59
4	0.8	0.2	0.15	0.85	0.32	0.19	0.41	0.37	0.20	0.44
4	0.8	0.2	0.10	0.90	0.13	0.11	0.26	0.14	0.11	0.29
5	0.6	0.4	0.25	0.75	0.62	0.37	0.45	0.70	0.40	0.48
5	0.6	0.4	0.20	0.80	0.41	0.28	0.34	0.42	0.26	0.34
5	0.6	0.4	0.15	0.85	0.19	0.17	0.23	0.17	0.15	0.20
5	0.6	0.4	0.10	0.90	0.06	0.09	0.12	0.05	0.07	0.10
5	0.8	0.2	0.25	0.75	0.63	0.29	0.53	0.77	0.33	0.61
5	0.8	0.2	0.20	0.80	0.39	0.18	0.40	0.48	0.21	0.44
5	0.8	0.2	0.15	0.85	0.18	0.10	0.26	0.22	0.12	0.29
5	0.8	0.2	0.10	0.90	0.06	0.05	0.15	0.06	0.05	0.15

Table 2: Mean rates of missing information from 100 simulated data sets under 12 conditions where $\gamma_1 = 0.5$ and ρ's are changing $\rho_{2j} = 1 - \rho_{1j}$.

Sample	Parameter Value						Rate of missing information					
size	γ_1	ρ_{11}	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{15}	γ_1	ρ_{11}	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{15}
100	0.5	0.25	0.25	0.25	0.25	0.25	0.66	0.46	0.43	0.44	0.42	0.42
100	0.5	0.25	0.25	0.25	0.25	0.10	0.43	0.27	0.26	0.25	0.26	0.54
100	0.5	0.25	0.25	0.25	0.10	0.10	0.26	0.14	0.14	0.15	0.38	0.39
100	0.5	0.25	0.25	0.10	0.10	0.10	0.16	0.09	0.09	0.25	0.24	0.25
100	0.5	0.25	0.10	0.10	0.10	0.10	0.10	0.06	0.17	0.17	0.17	0.17
100	0.5	0.10	0.10	0.10	0.10	0.10	0.06	0.11	0.13	0.11	0.10	0.11
1000	0.5	0.25	0.25	0.25	0.25	0.25	0.70	0.44	0.45	0.45	0.45	0.44
1000	0.5	0.25	0.25	0.25	0.25	0.10	0.49	0.26	0.26	0.26	0.26	0.62
1000	0.5	0.25	0.25	0.25	0.10	0.10	0.28	0.14	0.15	0.14	0.38	0.40
1000	0.5	0.25	0.25	0.10	0.10	0.10	0.15	0.08	0.08	0.23	0.25	0.24
1000	0.5	0.25	0.10	0.10	0.10	0.10	0.09	0.05	0.15	0.15	0.15	0.15
1000	0.5	0.10	0.10	0.10	0.10	0.10	0.04	0.09	0.08	0.08	0.09	0.08

Table 3: Percent bias and variance of parameter estimates and percent agreement of predicted and true class membership from 100 simulated LCA data sets under 16 conditions.

Number	Sample	Parameter Value			Percent Bias			Variance ($\times 10^2$)			Percent
of items	size	γ_1	ρ_1	ρ_2	γ	ρ_1	ρ_2	γ	ρ_1	ρ_2	agreement
4	100	0.6	0.25	0.75	-2.2	-5.0	-0.4	2.17	0.95	1.63	0.79
4	100	0.6	0.20	0.80	1.4	1.1	0.8	0.70	0.42	0.88	0.88
4	100	0.6	0.15	0.85	-0.8	-2.3	-0.2	0.39	0.31	0.53	0.93
4	100	0.6	0.10	0.90	-0.3	0.1	-0.2	0.30	0.17	0.33	0.97
4	1000	0.6	0.25	0.75	-0.5	-0.6	0.1	0.17	0.08	0.13	0.81
4	1000	0.6	0.20	0.80	0.3	-0.8	0.1	0.05	0.04	0.08	0.88
4	1000	0.6	0.15	0.85	0.3	0.3	0.0	0.04	0.03	0.05	0.93
4	1000	0.6	0.10	0.90	-0.0	-0.1	-0.1	0.03	0.02	0.03	0.97
5	100	0.6	0.25	0.75	-0.6	-3.1	0.7	1.30	0.66	1.36	0.83
5	100	0.6	0.20	0.80	-0.6	-3.5	0.2	0.49	0.43	0.66	0.91
5	100	0.6	0.15	0.85	0.3	-0.8	0.4	0.31	0.27	0.45	0.95
5	100	0.6	0.10	0.90	-0.9	1.7	0.1	0.28	0.17	0.27	0.98
5	1000	0.6	0.25	0.75	0.2	0.2	0.4	0.08	0.06	0.09	0.84
5	1000	0.6	0.20	0.80	0.9	0.4	0.1	0.04	0.04	0.05	0.91
5	1000	0.6	0.15	0.85	0.3	0.3	-0.0	0.04	0.03	0.04	0.96
5	1000	0.6	0.10	0.90	0.1	-1.2	-0.0	0.03	0.02	0.03	0.98

for classes with larger prevalences; however, there are no consistent patterns when comparing models fit to 100 versus 1000 observations.

In a second set of simulations, we varied the number of items with low measurement error (response probabilities of 0.10 and 0.90) and moderate measurement error (response probabilities of 0.25 and 0.75). The prevalence of class 1 was set to 0.50 for sample sizes of 100 and 1000 and five items. Equal class sizes were chosen to simplify reporting of results, because this would result in about equal rates of missing information for class 1 and class 2 parameters. Response probabilities for class 1 were set as follows: For model 1, we set one ρ to 0.10 the remaining four ρ's to 0.25. For model 2, we set two ρ's to 0.10 and the remaining three ρ's to 0.25. Similarly, for models 3 and 4, we set 3 and 4 ρ's to 0.10 and the remaining ρ's were set to 0.25. The ρ values for class 2 were set equal to 1 minus the values for class 1. Models with 0 and 5 ρ's equal to 0.10 and 0.90 were included for comparison.

Table 2 displays the results from the second simulation. As expected given the equal class sizes, the rate of missing information was similar for class 1 and class 2 parameters; therefore, we only present results for class 1. Increasing the number of items with low measurement error reduces the rates of missing information for the class prevalence γ and the response probabilities ρ with the same value; however, somewhat surprisingly, within a model, the rate of missing information is larger for items with lower measurement error, i.e., the values closer to zero or one compared to values closer to 0.5. As in the first simulation, there are no clear patterns with increasing sample size.

To better understand how the rate of missing information may provide insight into the value of surrogates in measuring the latent variable of interest and the usefulness of the fitted latent class model, we examined the finite sample bias and variability of the estimated parameter values for the simulated data sets (Table 3). Percent bias was defined as the difference between the mean of the estimated parameter values across the 100 simulated data sets and the true parameter value divided by the absolute value of the true parameter value. The variance across the 100 imputed data sets was also calculated. For all cases, the bias is very small. In general, both the bias and variance decrease as the rate of missing information decreases in addition to increasing sample size. This might suggest a connection between the rates of missing information and the required sample size for asymptotic results to hold. In other words, as the rates of missing information increase, a larger sample size is needed to get unbiased estimates.

We also estimated the percent agreement between the predicted and the true latent class memberships (Table 3). Class memberships were imputed from the posterior probabilities of class membership given the observed data and the maximum likelihood parameter estimates. The percent agreement follows the same pattern as the rates of missing information; the percent agreement is higher for models with lower rates of missing information. Roughly, for rates of missing information above 50%, the percent agreement is less than 90%. Thus, the rate of missing information sheds light on the usefulness of the surrogates for classifying individuals.

3.2 Latent class regression

We also conducted a simulation study for LCR to understand the behavior of the rate of missing information in the regression setting. We modeled the probabilities of class membership as a function of two covariates $\mathbf{x}$ where x_1 is a binary variable with prevalence 0.6 and x_2 is a continuous variable sampled from a normal $(0, 1)$ distribution. We set $\beta_0 = 0.4$, $\beta_1 = 1$, and $\beta_2 = -1$. The response probabilities, sample size, and number of items were varied as in the first simulation for LCA. We imputed 100 sets of class memberships from the posterior probabilities and estimated the rates of missing information for the class-specific item prevalences using the binomial distribution, as described above. Regression coefficients were estimated from the 100 sets of complete-data using logistic regression, and results were combined to estimate the rates of missing information using PROC MIANALYZE available in SAS version 8.2 or higher (SAS Institute, Inc., Cary, NC).

The rates of missing information for the LCR models are summarized in Table 4. As before, the most notable change in the rates of missing information occur with the change of response probabilities. Increasing the number of items decreases the rates of missing information, while the sample size does not have much of an effect.

Table 4: Mean rates of missing information from 100 simulated LCR data sets under 16 conditions where $\beta_0 = 0.4$, $\beta_1 = 1$, and $\beta_2 = -1$.

Item	Sample	ρ Value		Rate of missing information				
number	size	$\boldsymbol{\rho_1}$	$\boldsymbol{\rho_2}$	β_0	β_1	β_2	$\boldsymbol{\rho_1}$	$\boldsymbol{\rho_2}$
4	100	0.25	0.75	0.53	0.38	0.44	0.45	0.51
4	100	0.20	0.80	0.40	0.30	0.33	0.42	0.44
4	100	0.15	0.85	0.22	0.17	0.19	0.31	0.34
4	100	0.10	0.90	0.08	0.07	0.08	0.18	0.21
4	1000	0.25	0.75	0.66	0.45	0.50	0.53	0.55
4	1000	0.20	0.80	0.43	0.30	0.34	0.43	0.44
4	1000	0.15	0.85	0.22	0.16	0.18	0.31	0.31
4	1000	0.10	0.90	0.09	0.07	0.08	0.20	0.21
5	100	0.25	0.75	0.49	0.35	0.41	0.41	0.43
5	100	0.20	0.80	0.29	0.23	0.25	0.30	0.33
5	100	0.15	0.85	0.14	0.11	0.13	0.20	0.22
5	100	0.10	0.90	0.05	0.04	0.05	0.11	0.12
5	1000	0.25	0.75	0.52	0.37	0.41	0.43	0.44
5	1000	0.20	0.80	0.29	0.22	0.25	0.31	0.31
5	1000	0.15	0.85	0.14	0.11	0.13	0.20	0.20
5	1000	0.10	0.90	0.04	0.04	0.04	0.10	0.10

The bias and variance of the LCR parameter estimates from the simulated data are shown in Table 5. We only report the values for the β vector, because the information for the other parameters was discussed in the previous section. The percent bias, while generally small, is larger for the regression coefficients compared to the class and item-specific prevalences described in the previous section. As in the previous simulation, the variance decreases with decreasing rates of missing information and increasing sample size; however, the pattern is less consistent for the percent bias. A noteworthy result is that the bias for β_1 and β_2 are away from the null, i.e., the absolute size of the regression coefficients tend to be overestimated. The bias may be large enough in the models with 100 subjects to be of potential concern – A 16% bias on the log-odds scale results in an odds ratio of 2.7 being overestimated as 3.2. This is in contrast with the influence of nondifferential misclassification which is known to bias risk estimates towards the null (Copeland *et al.*, 1977; Flegal *et al.*, 1986). In the latent class regression setting, which attempts to correct for misclassification, it seems plausible that any potential finite-sample bias may be away from the null, because covariates are also being used to help identify an individual's true class membership. This may tend to maximize the estimated relationship between the covariates and class membership in small samples.

Table 5: Percent bias and variance from 100 simulated LCR data sets under 16 conditions where $\beta_0 = 0.4$, $\beta_1 = 1$, and $\beta_2 = -1$.

Number	Sample	ρ Value		Percent Bias			Variance		
of items	size	ρ_1	ρ_2	β_0	β_1	β_2	β_0	β_1	β_2
4	100	0.25	0.75	0.7	0.5	-9.0	0.538	0.468	0.259
4	100	0.20	0.80	-6.0	8.4	-11.8	0.383	0.454	0.194
4	100	0.15	0.85	10.8	3.7	-10.1	0.159	0.285	0.163
4	100	0.10	0.90	7.7	-2.7	-2.1	0.160	0.283	0.162
4	1000	0.25	0.75	-1.6	0.3	-1.9	0.043	0.049	0.020
4	1000	0.20	0.80	-1.3	5.8	-0.8	0.021	0.037	0.010
4	1000	0.15	0.85	2.4	0.9	-0.2	0.017	0.028	0.010
4	1000	0.10	0.90	0.7	2.0	0.7	0.012	0.025	0.008
5	100	0.25	0.75	10.5	15.9	-14.5	0.358	0.693	0.258
5	100	0.20	0.80	-7.3	9.5	-4.6	0.254	0.482	0.215
5	100	0.15	0.85	15.2	0.2	-10.7	0.158	0.215	0.110
5	100	0.10	0.90	17.3	5.5	-7.9	0.155	0.311	0.114
5	1000	0.25	0.75	5.3	-0.4	-1.2	0.027	0.041	0.015
5	1000	0.20	0.80	1.1	1.7	-0.6	0.020	0.037	0.010
5	1000	0.15	0.85	4.3	0.8	-1.3	0.016	0.034	0.009
5	1000	0.10	0.90	-2.1	1.8	-1.1	0.016	0.032	0.007

4. Examples

To illustrate the proposed measure, we reanalyze data from two previously published studies. The first is a latent class analysis presented by Garrett and Zeger (2000) on depression from the Epidemiologic Catchment Area Program. The second is a latent class regression analysis presented by (Badeen-Roche *et al.*, 1997) on mobility disability from the Woman's Health and Aging Study.

4.1 Latent class analysis

The National Institute of Mental Health (NIMH) Epidemiologic Catchment Area Program (ECA) is a five-site epidemiologic study focusing on mental health (Eaton, Reiger and Locke, 1981). Garrett and Zeger (2000) analyzed data from 2,938 individuals interviewed at the Baltimore site in 1981. The goal was to use 17 questions from the NIMH Diagnostic Interview Schedule to measure 6-month prevalence of depression. These questions were grouped into 9 items (see Table 6) and analyzed using latent class analysis fitted using a Bayesian approach. Garrett and Zeger (2000) concluded that three class model is ”statistically the most appropriate.” The four class model was not well identified and the three-class model was judged to fit better than the two class model.

We reanalyzed the data using the freeware WinLTA[1] (Collins *et al.*, 1999; Collins *et al.*, 2001). WinLTA uses the EM algorithm to find the maximum likelihood estimate and data augmentation (DA) for Bayesian estimation, variance estimation, and multiple imputation. When using the DA tab in winLTA for multiple imputation, the rates of missing information are given as a default. The maximum likelihood estimates for two and three class models are summarized in Table 6 and are similar to the results of Garrett and Zeger (2000), though some small differences exist due to the different fitting approaches.

Table 6: Two and three class model estimates and rates of missing information for the ECA depression data

	Two Classes				Three Classes					
	Estimate		Rate of missing info		Estimate			Rate of missing info		
Depression Status:	No	Yes	No	Yes	No	Minor	Major	No	Minor	Major
Symptom Prevalence										
Dysphoria	0.02	0.41	0.34	0.25	0.01	0.23	0.77	0.61	0.70	0.70
Loss of appetite	0.06	0.45	0.17	0.23	0.05	0.33	0.68	0.48	0.59	0.44
Insomnia, Hypersomnia	0.06	0.63	0.33	0.33	0.04	0.49	0.81	0.62	0.69	0.47
Slow movement, Restless	0.02	0.42	0.37	0.27	0.01	0.24	0.77	0.54	0.74	0.63
Disinterest in sex	0.01	0.20	0.22	0.15	0.01	0.12	0.34	0.44	0.57	0.32
Reduced energy, Fatigue	0.03	0.49	0.26	0.24	0.02	0.32	0.77	0.45	0.74	0.45
Guilty, Sinful	0.00	0.27	0.36	0.23	0.00	0.12	0.62	0.56	0.65	0.67
Reduced concentration	0.02	0.40	0.25	0.27	0.01	0.21	0.77	0.50	0.71	0.72
Thoughts of suicide	0.06	0.52	0.21	0.30	0.05	0.38	0.77	0.45	0.62	0.62
Class prevalence	0.87	0.13	0.50	0.50	0.81	0.16	0.03	0.85	0.76	0.85

When fitting a two class LCA model, the classes may be defined as a not depressed group and a depressed group. The depressed group, comprising 13% of the population, has moderate to high probabilities of all symptoms (20-63%)

[1] Available at http://methodology.psu.edu/downloads/winlta.html

with an average of about four total reported symptoms per person. The majority of the population (87%) are not depressed and thus have very low probability of reporting any symptoms ($< 6\%$). The three class model has a not depressed group and two depression groups which we labeled minor and major depression. The minor depression group has a prevalence of 16% and has low to moderate probabilities for all symptoms (12-49%), with an average of about two symptoms per person. The major depression group, comprising only a small fraction of the population (3%), has high probability of all symptoms (34-81%) with an average of six symptoms per person.

The rates of missing information are reasonable for the two class model ($20\% - 30\%$), but moderate to high for the three class model ($40\% - 70\%$). This suggests there is sufficient information in these nine symptoms to reliably classify individuals into depressed versus not depressed groups; however, there may not be enough information to reliably distinguish people with minor depression. This is consistent with the findings of (Gartett, Eaton and Zeger, 2002), who used a latent class approach to evaluate diagnostic criteria for depression. Based on the positive and negative predicted values estimated from the model, they concluded these nine symptoms provide essentially no information about minor depression. This is supported by the psychiatric literature, which has not yet developed consistently used criteria for diagnosing minor depression (Pincus, Davis and McQueen, 1999). The Diagnostic and Statistical Manual of Mental Disorders 4th edition (DSM-IV) defines widely accepted criteria for diagnosing major depressive disorder; however, criteria for minor depression is described in Appendix B with mental disorders that are considered to have “insufficient information to include as official categories” (APA, 1994).

4.2 Latent class regression

For the LCR analysis example we used the data previously analyzed by (Bandeen-Roche *et al.*, 1997) from the Women’s Health and Aging Study (WHAS), a study of the course of disability among moderately and severely disabled elderly women in Baltimore, Maryland. The WHAS study followed 1,002 disabled women aged 65 and older from November 1992 to February 1995. For this study we use population-based data from 3,543 women that were interviewed as part of the baseline screener.

The WHAS instrument included self-reported measures of disability, disease, and demographics. Following the analysis in Bandeen-Roche *et al.* (1997), we analyzed data from the following items that characterize mobility disability: “Without help, do you have any difficulty [doing a specific task]?” walking $\frac{1}{4}$ mile, climbing 10 steps, lifting up to 10 pounds, and getting in and out of bed or a chair. We regressed latent mobility disability status on age and arthritis status. We fit the

LCR model using a SAS (SAS Institute, Inc., Cary, NC) macro written by the second author[2]. This macro uses the EM algorithm with a Newton-Raphson step to find the maximum likelihood estimates for the model parameters (Bandeen-Roche *et al.*, 1997). The results for the two and three class models are summarized in Table 7. Bandeen -Roche *et al.* (1997) concluded the three class model provided a reasonable fit to the data.

To estimate the rates of missing information, we imputed 100 sets of class memberships from the posterior probabilities of class membership after sampling 100 sets of parameter values from a multivariate normal distribution with mean and variance estimated from the LCR model. Regression coefficients were estimated from the imputed data sets using polytomous regression and results were combined to estimate the rates of missing information using SAS's PROC MI-ANALYZE (SAS Institute, Inc., Cary, NC).

Table 7: LCR two and three class parameter estimates and rates of missing information for the WHAS mobility disability example.

	Two Classes				Three Classes					
	Estimate		Rate of missing info		Estimate			Rate of missing info		
Disability status:	No	Yes	No	Yes	No	Mild	Severe	No	Mild	Severe
Item prevalences										
Heavy housework	0.12	0.89	0.38	0.33	0.05	0.66	0.95	0.86	0.86	0.72
Walk 1/4 mile	0.12	0.83	0.33	0.29	0.06	0.57	0.94	0.78	0.84	0.78
Climb 10 steps	0.03	0.62	0.34	0.24	0.02	0.24	0.88	0.65	0.83	0.91
Lift 10 pounds	0.04	0.67	0.39	0.25	0.02	0.37	0.82	0.76	0.83	0.70
Getting in/out chair	0.02	0.39	0.22	0.14	0.01	0.14	0.57	0.50	0.76	0.68
Class Prevalence	0.64	0.36	0.34	0.34	0.52	0.29	0.19	0.85	0.76	0.86
Intercept	1.65	ref	0.30	ref	2.47	1.19	ref	0.68	0.50	ref
Age	-0.08	ref	0.17	ref	-0.10	-0.04	ref	0.58	0.27	ref
Arthritis	-1.47	ref	0.19	ref	-1.92	-0.78	ref	0.26	0.47	ref

The two class model shows 64% of women with no disability, having a low probability of reporting difficulty with any tasks (2-12%). The remaining 36% of women may be considered to have mobility disability, with high prevalence of task difficulties (39-89%) and on average, reporting difficulty with 3.4 of the 5 tasks. The odds of being in the disabled group is 4.3 times greater for women with arthritis (95% CI = 3.6 to 5.2) and 2.2 times greater for every 10 year increase in age (95% CI = 2.0 to 2.5).

The three class model shows 52% of women in the no disability group with very low probability ($< 6\%$) of difficulty with any task. Twenty-nine percent of women may be considered to have mild disability, with low to moderate probability of difficulty with each task (14-66%) and an average of 2 reported task difficulties.

[2] Available at http://www.centerforhealthstudies.org/perpages/migliore/software.html

The remaining 19% of women fall in the severe disability group with a high probability of task difficulties (57-95%) and reported difficulty with an average of 4 out of 5 tasks. The odds of being in the severe versus the no disability group is 6.8 times higher for women arthritis (95% CI = 5.2 to 8.9) and 2.9 times higher for every 10 year increase in age (95% CI = 2.5 to 3.3). The odds of being in severe versus the mild disability group is 2.2 times higher for women with arthritis (95% CI = 1.6 to 3.0) and 1.3 times higher for every 10 year increase in age (95% CI = 1.3 to 1.8).

The rates of missing information are high for three class model ($50\% - 80\%$) but reasonable for two class model ($20\% - 40\%$). This suggests that these items can be reliably used to classify patients into a healthy group of women who rarely report difficulty with any task and a disabled group with high probability of difficulty with three or more tasks; however, there may not be enough information in these items to reliably distinguish between women with mild versus severe disability. Despite this, given the large sample size, we may still be able to correctly quantify the influence of arthritis and age on the probability of being mild versus severely disabled based on the latent class regression model; however, uncertainty will be larger than for the two class model. Because Bandeen-Roche *et al.* (1997) found that the three class model fit the data better than the two class model, the three class model is preferable for making inference about the association between risk factors and disability.

5. Discussion

In this paper we introduce the rate of missing information in the context of LCA and explore the use of this measure as a diagnostic tool for latent class analysis and regression. The rate of missing information gives an indication of the amount of information missing as a result of observing multiple surrogates in place of the underlying latent variable of interest, and provides a measure of how confident one can be in the model results. If inference is based on high levels of missing information, one might be skeptical about the accuracy and usefulness of the LCA results, especially in small samples.

As demonstrated in the simulation studies and examples, the rates of missing information can be used to assess the potential of symptoms or other surrogates to be used as diagnostic criteria in the absence of a gold standard. Models with high rates of missing information (rates above approximately 50%) do not predict true class membership well and indicate the need for additional symptoms or surrogates with less measurement error for accurate classification. The rate of missing information may also be valuable for the design of future studies. By knowing if items have a strong effect on the rate of missing information, one can plan to add items, change items, or put emphasis on item quality in future studies.

In addition, high rates of missing information indicate that larger samples sizes are needed to obtain precise and unbiased estimates of the latent class model parameters.

The rate of missing information may also be useful when estimating diagnostic accuracy in absence of a gold standard. The rate of missing information provides a measure of whether it is appropriate to use LCA for measuring sensitivity, specificity, and prevalence. Based on our simulation results, if all tests have low sensitivity and specificity, there is likely to be a large amount of information missing by not directly observing a gold standard, and therefore, there is less faith in the latent class results. However, the addition of one or two tests with high sensitivity and specificity, say around 90%, is likely to increase the amount of information about all model parameters including the sensitivities and specificities of the other tests plus disease prevalence. If tests with high sensitivity and specificity are unavailable, larger sample sizes will be required to obtain precise and unbiased estimates.

When the observed data is incomplete, the missing data can be separated into two types, the missing latent class memberships and the missing surrogate values. Using two-stage MI (Harel, 2003), it would be interesting to separate the effect of the missing class memberships and the effect of the missing values on the overall uncertainty of the model. This is a topic for future study.

An unintended result of our simulation study was to find that latent class regression may bias risk estimates away from the null in small samples. In the case of 100 subjects, an odds ratio of 2.7 was overestimated to be as large as 3.2 on average. Future work should examine this issue in more detail. Until this potential bias is be better understood, care should be taken when fitting latent class regression models to small samples.

Acknowledgments

The work for this paper was partially supported by grant R01 HL62567 from the National Institute of Health and by grant MH069864 from the National Institute of Mental Health. The authors thank Elizabeth Garrett-Mayer, PhD for the use of the ECA data (National Institute of Mental Health grant MH56639-03) and Karen Bandeen-Roche, PhD for the use of the WHAS data (National Institute of Aging grant AG19905).

References

APA (1994). *Diagnostic and Statistical Manual of Mental Disorders*, 4-th ed. American Psychiatric Association.

Bandeen-Roche, K.,Miglioretti, D. L. and Zeger, S. L. and Rathouz, P.aul J. (1997). Latent variable regression for multiple discrete outcomes *Journal of the American Statistical Association* **92**, 1375-1386.

Butler, J. C. Bosshardt, S. C., Phelan, M., Moroney, S. M. Tondella, M. L. Farley, M. M., Schuchat, A. and Fields, B. S. (2003). Classical and latent class analysis evaluation of sputum polymerase chain reaction and urine antigen testing for diagnosis of pneumococcal pneumonia in adults. *J. Infect. Disease* **187**, 1416-1423.

Clogg, C. C. and Goodman, L. A. (1984). Latent structure analysis of a set of multidimentional contingency tables, *Journal of the American Statistical Association* **79**, 762-771.

Clogg, C. C. (1995). Latent class models. In *Handbook of Statistical Modeling for the Social and Behavioral Sciences* (Edited by Arminger, G., Clogg, Clifford C., and Sobel, M. E.), 311-360, Plenum Publishing Corporation.

Collins, L. M., Flaherty, B. P., Hyatt, S. L. and Schafer, J. L. (1999). *WinLTA User's Guide Part 1.* The Methodology Center, Penn State University.

Collins, L. M., Lanza, S. L. and Schafer, J. L. (2001). *WinLTA User's Guide for Data Augmentation.* The Methodology Center, Penn State University.

Collins, L. M., Fidler, P. L. Wugalter, S. E. and Long, J, D. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavior Research* **28**, 375-389.

Copeland, K. T., Checkoway, H., McMichael, A. J. and Holbrook, R. H. (1997). Bias due to misclassification in the estimation of relative risk. *American Journal of Epidemiology* **5**, 488-495.

Dayton, C. M. and Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association.* **83**, 173-178.

Eaton, W. W., Reiger, D. A. and Locke, B. Z. (1981). The NIMH epidemiologic catchment area progrom. *Public Health Reports* **1981**, 319-325.

Eaton, W. W. and Bohrnstedt, G. (1989). Introduction to latent variable models for dichotomous outcomes: Analysis of data from the epidemiologic catchment area program. *Sociological Methods and Research* **18**, 4-18.

Fergusson, D. M., Horwood, L. J. and Lynskey, M. T. (1995). The prevalence and risk factors associated with abusive or hazardous alcohol consumption in 16-year-olds. *Addiction* **90**, 935-946.

Flegal, K. M., Brownie, C. and Haas, J. D. (1986). The effects of exposure misclassification on estimates of relative risk. *American Journal of Epidemiology* **123**, 736-751.

Formann, A. K. (1996). Latent class analysis in medical research. *Statistical Methods in Medical Research* **5**, 179-211.

Fossati, A., Maffei, C., Battaglia, M., Bagnato, M., Donati, D., Donini, M., Fiorilli, M. and Novella, L. (2001). Latent class analysis of DSM-IV schizotypal personality disorder criteria in psychiatric patients. *Schizophrenia Bulletin* **27**, 59-71.

Garrett, E. S., Eaton, W. W. and Zeger, S. (2002). Methods for evaluating the performance of diagnostic tests in the absence of a gold standard: A latent class model approach. *Statistics in Medicine* **21**, 1289-1307.

Garrett, E. S. and Zeger, S. (2000). *Latent class model diagnosis. Biometrics* **56**, 1055-1081.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **1974**, 215-231.

Goodman, L. A. (1978). *Analyzing Qualitative/Categorical Data.* Cambridge.

Hagenaars, J. A. (1988). Latent structure models with direct effects between indicators: Local dependence models. *Sociological Methods and Research* **16**, 379-405.

Harel, O. (2003). Strategies for data analysis with two types of missing values. Pd.D. Thesis, The Pennsylvania State University.

Harel, O. and Zhou, X. H. (2007). Multiple imputation review of theory of implementation and softwaare. *Statistics in Medicine* (in press).

Humphreys, K. and Janson, H. (2000). Latent transition analysis with covariates, nonresponse, summary statistics and diagnostics: Modeling children's drawing development, *Multivariate Behavioral Research* **35**, 89-119.

Lanza, S. T., Collins, L. M., Schafer, J. L. and Flaherty, B. P. (2005). Using data augmentation to obtain standard errors and conduct hypothesis tests in latent class and latent transition analysis. *Psychological Methods* **10**, 84-100.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis In *Stadies in social psychology in world war II: Measurament and Prediction* **4** (Edited by Stouffer, S. A., Guttman, L., Suchman, E. A., Lazarsfeld, P. F., Star, S. A. and Clausen, J. A.), 362-412. Princton University Press,

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis.* Houghton-Mifflin.

Lin, H., and Turnbull, B. W., McCulloch, C. E. and Slate, E. H. (2003). Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association* **97**, 53-65.

McCutcheon, A. L. (1987). *Latent class analysis*, Sage Publications Inc.

McHugh, R. (1956). Efficient estimation and local identification in latent classes analysis. *Psychometrika* **21**, 331-347.

McHugh, R. (1958). Note on: Efficient estimation and local identification in latent classes analysis. *Psychometrika* **23**, 273-274.

Miglioretti, D. L. (2003). Latent transition regression for mixed outcomes. *Biometrics* **59**, 710-720.

Moran, M., Walsh, C., Lynch, A., Coen, R. F., Coakley, D. and Lawlor, B. A. (2004). Syndromes of behavioural and psychological symptoms in mild Alzheimer's disease. *Int. J. Geriatr Psychiatry* **19**, 359-364.

Muthen, B. O. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data In *Handbook of Quantitative Methodology for the Social Sciences* (Edited by Kaplan, D.), 345-368. Sage Publications.

Muthen, B. O., Brown, C. H., Masyn, K., Jo, B., Khoo, S. T., Yang, C. C., Wang, C. P., Kellam, S. G., Carlin, J. B. and Liao, J. (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics* **3**, 459-475.

Nestadt, G., Addington, A., Samuels, J., Liang, K. Y., Bienvenu, O. J., Riddle, M., Grados, M., Hoehn-Saric, R. and Cullen, B. (2003). The identification of OCD-related subgroups based on comorbidity. *Biol. Psychiatry* **53**, 914-920.

Pincus, H. A, Davis, W. W. and McQueen, L. E. (1999). Subthreshold' mental disorders. A review and synthesis of studies on minor depression and other 'brand names'. *Br. J. Psychiatry* **147**, 288-296.

Reiser, M. and Lin, Y. A. (1999). Goodness-of-fit test for the latent class model when expected frequencies are small. In *Sociological Methodology* (Edited by M. E. Sobel and M. P. Becker), 81-111. Blackwell.

Rosen, O. and Tanner, M. (1999). Mixtures of proportional hazards regression models. *Statistics in Medicine* **18**, 1119-1131.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* J. Wiley and Sons,

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of American Statistical Association*, **91**, 473-489.

Rudas, T., Clogg, C. C. and Lindsay, B. G. (1994). A new index of fit based on mixture methods for the analysis of contingency tables. *Journal of the Royal Statistical Society, Series B* **56**, 623-639.

Schafer, J. L. (1997) *Analysis of Incomplete Multivariate Data*, Chapman and Hall.

Schafer, J. L. and Graham, J. W. (2002). Missing data: a review. *Psychological Methods* **7**, 147-177.

Storr, C. L., Reboussin, B. A. and Anthony, J. C. (2004). Early childhood misbehavior and the estimated risk of becoming tobacco-dependent. *American Journal of Epidemiology* **160**, 126-130.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation". *J. American Statistical Association* **82**, 528-540.

Young, M. A. (1982). Evaluating diagnostic criteria: a latent class paradigm. *Journal of Psychiatric Research* **17**, 285-296.

Young, M. A., Abrams, R., Taylor, M. A. and Meltzer, H. Y. (1983). Establishing diagnostics criteria for mania. *Journal of Nervous and Mental Disease* **171**, 676-682.

Received August 6, 2005; accepted December 6, 2005.

Ofer Harel
Department of Statistics
University of Connecticut,
215 Glenbrook Road, Unit 4120, Storrs, CT, USA
oharel@stat.uconn.edu

Diana Miglioretti
Group Health Cooperative
1730 Minor Ave., Suite 1600
Seattle, WA 98101, USA
miglioretti.d@ghc.org